WWU MÜNSTER

› Stable and efficient Petrov-Galerkin methods for certain (kinetic) transport equations

Dissertation

Julia Brunken
2021

living.knowledge

Fach: Mathematik

Stable and efficient Petrov-Galerkin methods for certain (kinetic) transport equations

Inauguraldissertation
zur Erlangung des Doktorgrades der Naturwissenschaften
im Fachbereich Mathematik und Informatik
der Mathematisch-Naturwissenschaftlichen Fakultät
der Westfälischen Wilhelms-Universität Münster

eingereicht von
Julia Brunken
aus Varel

2021

Dekan: Prof. Dr. Xiaoyi Jiang
Erstgutachter: Prof. Dr. Mario Ohlberger
Zweitgutachterin: Dr. Kathrin Smetana
Drittgutachterin: Dr. Virginie Ehrlacher

Tag der mündlichen Prüfung: 08.03.2021
Tag der Promotion: 08.03.2021

Abstract

In this thesis, we develop stable and efficient Petrov-Galerkin discretizations for two different transport-dominated problems: first order linear transport equations and a kinetic Fokker-Planck equation. Based on well-posed weak formulations on the continuous level, the core of our numerical schemes is the choice of the discrete spaces for the Petrov-Galerkin projection. By first defining a discrete test space and then computing a problem-dependent discrete trial space such that the spaces consist of matching stable pairs of trial and test functions, we obtain efficiently computable uniformly stable discrete schemes.

For first order linear transport equations, we use an optimally conditioned ultraweak variational formulation. Then, the optimally stable discrete trial space results from the chosen discrete test space by an easy-to-compute application of the adjoint (differential) operator. For the kinetic Fokker-Planck equation, we derive a favorable lower bound for the inf-sup constant on the continuous level with methods inspired by well-posedness results for parabolic equations. Here, the stable discrete trial space is constructed from the test space by the application of the kinetic transport operator and the inverse velocity Laplace-Beltrami operator, so that the specific basis functions can be efficiently computed by low-dimensional elliptic problems. In both cases we thereby guarantee the discrete inf-sup stability with the same inf-sup constant as on the infinite-dimensional level independently of the chosen test spaces.

This guaranteed stability is especially beneficial when considering model reduction by the reduced basis method for parametrized first-order transport equations. Using our discretization strategy, we build a reduced model consisting of a fixed reduced test space generated by a greedy algorithm and parameter-dependent reduced trial spaces depending on the test space. Since the stability is inherently built into the method, we can avoid additional stabilization loops within the greedy algorithm, so that we obtain efficient reduced models by an easily implemented procedure.

Zusammenfassung

In dieser Arbeit entwickeln wir stabile und effiziente Petrov-Galerkin-Diskretisierungen für zwei verschiedene transportdominierte Probleme: lineare Transportgleichungen erster Ordnung und kinetische Fokker-Planck-Gleichungen. Aufbauend auf wohlgestellten schwachen Formulierungen liegt der Kern unserer numerischen Methoden in der Wahl der diskreten Räume für die Petrov-Galerkin-Projektion. Indem wir zunächst den diskreten Testraum definieren und dann einen problemangepassten diskreten Ansatzraum so berechnen, dass die Räume aus passenden stabilen Paaren von Ansatz- und

Testfunktionen bestehen, erhalten wir effizient berechenbare und uniform stabile diskrete Methoden.

Für lineare Transportgleichungen erster Ordnung benutzen wir optimal konditionierte “ultraschwache” Variationsformulierungen. Der optimal stabile diskrete Ansatzraum entsteht dann aus dem gewählten diskreten Testraum durch eine einfach zu berechnende Anwendung des adjungierten (Differential-)Operators. Für die kinetische Fokker-Planck-Gleichung leiten wir mit Methoden inspiriert von Wohlgestelltheitsresultaten für parabolische Gleichungen eine vorteilhafte untere Schranke für die Inf-Sup-Konstante auf der stetigen Ebene her. Hier wird der stabile diskrete Ansatzraum dann aus dem Testraum durch die Anwendung des kinetischen Transportoperators und des inversen Geschwindigkeits-Laplace-Beltrami-Operators konstruiert, sodass die speziellen Basisfunktionen durch niedrigdimensionale elliptische Probleme effizient berechnet werden können. In beiden Fällen garantieren wir dadurch diskrete Inf-Sup-Stabilität mit derselben Inf-Sup-Konstante wie auf der unendlichdimensionalen Ebene unabhängig von den gewählten Testräumen.

Diese garantierte Stabilität ist insbesondere auch für Modellreduktion durch die Reduzierte-Basis-Methode für parametrisierte Transportgleichungen erster Ordnung von Vorteil. Indem wir unsere Diskretisierungsstrategie nutzen, konstruieren wir ein reduziertes Modell bestehend aus einem festen, durch einen Greedy-Algorithmus generierten reduzierten Testraum und parameterabhängigen reduzierten Ansatzräumen, die aus dem Testraum hervorgehen. Da die Stabilität grundsätzlich in die Methode eingebaut ist, können wir zusätzliche Stabilisierungen im Greedy-Algorithmus vermeiden, sodass wir effiziente und leicht zu implementierende reduzierte Modelle erhalten.

Acknowledgements

First and foremost, I would like to thank my supervisors Prof. Dr. Mario Ohlberger and Dr. Kathrin Smetana. I am very grateful to Mario for giving me the possibility to work in his group, for offering his experience and advise, for always being patient and supportive of new research ideas and changes of plans, and for being a mentor for quite some years now.

I am deeply indepted to Kathrin for her tremendous support during my whole PhD years, for her valuable advice and ideas on so many questions and for still making sure that we stayed on my own path onward. I am extremely grateful for her encouragement and optimism especially in the more bumpy times, for her extensive and detailed feedback on all manuscripts that greatly improved my writing and presentation skills, for so many hours of Skype sessions thinking about our latest mathematical problems and often also discussing the recent problems and joys of the world in general. Thanks for all the good times at and especially besides conferences, and not least for the tips on whale-watching and the best ice cream in Boston.

I would like to thank Prof. Dr. Karsten Urban for our fruitful and enjoyable collaboration, it has been a great pleasure working with him. I also thank Dr. Virginie Ehrlacher for agreeing to co-examine this thesis.

I gratefully acknowledge financial support by the German Federal Ministry of Education and Research through the project *GlioMaTh: Gliomen, Mathematische Modelle und Therapieansätze* under grant BMBF 05M2016 and thank all collaborators of the GlioMaTh project for their valuable input. Furthermore, this work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2044 –390685587, *Mathematics Münster: Dynamics–Geometry–Structure.*

I would like to thank my current and former colleagues in the AG Ohlberger and the Institute for Analysis and Numerics for the friendly working atmosphere and their help. Special thanks go to Barbara Verfürth, Tobias Leibner, Julia Schleuß, Marie Tacke, and Tim Keil for all the fun lunch breaks, conferences, or – in the recent times – Zoom "tea breaks".

Thanks also to Barbara Verfürth and Martin Maiwald for proofreading parts of this thesis.

Finally, I sincerely thank my family and my friends, particularly my parents Marlies and Frank, my sister Marika, and Leonie, for all their support throughout the years. Special thanks for their encouragement, motivation, and the cookies that brought me through the final weeks of finishing this thesis.

And, most importantly: Martin, thank you for always being there for me, sharing the joyful moments, and cheering me up when I need it. I am so happy to have you in my life.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Transport phenomena arise in all major areas of science and technology and their numerical simulation is often crucial for the understanding of natural processes. One particular example of an application where complex transport equations arise lies in mathematical models describing the spreading of gliomas in the human brain. Gliomas are rarely curable brain tumors that are highly infiltrative and possess specific "finger-like" invasion patterns, which are believed to result from anisotropic cell movement aligned to white matter tracts of the brain, see [61, 62]. As a full resection of the tumor in surgery is in general not possible, chemotherapy and radiotherapy play an important part in the treatment of patients. The planning of radiotherapy treatment involves the assessment of the tumor margins, which is exacerbated by the fact that parts of the tumor may not be visible in medical imaging and the anisotropic invasion pattern is highly patient-specific.

Therefore, mathematical models of the tumor invasion that allow for a patient-specific simulation of the anisotropic cell movement might be useful to enhance the treatment planning. Multiscale models that aim at describing the different complex processes involved in glioma invasion have been developed, for instance, in [61, 62, 90, 95]. In these models, mesoscopic descriptions in form of kinetic equations for the density of the tumor cells are derived. The models include biological processes on the microscopic cell level as well as macroscopic data of the patient-specific brain structure leading to anisotropy in the cell movements. Simulations of the tumor invasion are then mostly based on macroscopic partial differential equations (PDEs) that result from a limiting procedure of the mesoscopic model [41, 61, 64]. However, additional insights might also be gained from simulations of the kinetic mesoscopic equations.

These kinetic equations are high-dimensional, contain several purely advective variables, and may also depend on additional biological parameters. These characteristics all pose significant challenges for the numerical solution. Since transport-dominated PDEs may often involve discontinuous solutions, standard numerical methods may lead to accuracy and stability problems when not taking the specific structure of the equation into account. While, for instance, projection-based methods like the finite element (FE) method are automatically well-posed for elliptic problems, simple Galerkin projections fail for transport-dominated problems, and additional stabilization is necessary. When we encounter complex settings including high-dimensional or parametrized equations, e.g., by biological parameters in the model, this stabilization problem also occurs when we employ model reduction methods to reduce the complexity of numerical simulations: Many standard approaches like the reduced basis (RB) method are also based on linear projections of weak solutions or detailed approximations onto reduced subspaces. Therefore, in the transport-dominated setting, a stabilization is also necessary in the generation of reduced models.

1.2 Goal and contribution of this work

In this thesis, we aim at building a framework for the efficient and stable simulation of (possibly kinetic) transport equations that may also involve parameters. To that end, we develop stable Petrov-Galerkin discretizations based on well-posed variational formulations for two different model problems:

- Parametrized first-order linear transport equations serve as an example for a hyperbolic problem, where the parameter dependence of the equation poses an additional difficulty and we can test an application of the RB method.
- The kinetic Fokker-Planck equation as a prototype of a mesoscopic glioma equation then contains both a multidimensional kinetic transport term on the purely advective space and time variables and a diffusion term in the velocity.

To obtain well-posed and stable Petrov-Galerkin discretizations, it is crucial to choose discrete trial and test spaces such that a discrete inf-sup condition with a favorable estimate for the discrete inf-sup constant can be shown. To achieve this, the main idea of our discrete schemes is to build the spaces for the Petrov-Galerkin projection completely from stable trial/test function pairs determined by the variational formulation. Since in these pairs the trial function can be easily computed from the test function (but not vice versa), we first choose a *fixed discrete test space* and then compute the *problem-dependent discrete trial space* consisting of the respective trial function counterparts in the stable pairs. With this strategy, we obtain discretizations that are inherently stable and efficiently computable. We show in the course of the thesis that this strategy leads to favorable solutions and is especially useful also in model reduction contexts.

The parametrized first-order linear transport equation considered in chapter 3 is of the form

$$\partial_t u_\mu(t,x) + \mathbf{b}_\mu(t,x) \cdot \nabla_x u_\mu(t,x) + c_\mu(t,x)\, u_\mu(t,x) = f_\mu(t,x), \tag{1.1}$$

for all parameter values μ in a compact set $\mathcal{P} \subset \mathbb{R}^p$, for all times $t \in (0,T)$ and all $x \in D \subset \mathbb{R}^d$ accompanied with appropriate initial and boundary conditions.

We consider weak solutions to (1.1) in an ultraweak space-time variational formulation that was already proposed, for instance, in [29, 43, 51, 52]. The ultraweak setting uses an L^2 trial space, while the test space is chosen with a problem-dependent norm containing the full adjoint transport operator. With this choice, well-posedness of the scheme with optimal inf-sup and continuity constants can be shown.

To obtain discrete solutions by a stable Petrov-Galerkin projection, related approaches usually choose a discrete trial space and then seek a discrete test space that ensures discrete inf-sup stability. However, the optimally stable test space results from applying the inverse adjoint transport operator to the trial space – which means, solving the PDE for every basis function, which is in general infeasible. Therefore, stable test spaces can only be approximated by different strategies [29, 43, 51, 52].

As mentioned above, we use the reverse approach and propose to first choose an appropriate test space. The trial space that results in optimal inf-sup stability is then given by an application of the adjoint transport operator. Since this application of a differential operator (instead of solving PDEs) can be easily and exactly computed for standard discrete spaces as e.g. an FE space, we obtain a numerical scheme which is optimally stable (with an inf-sup constant of one also for variable coefficients and independently of the mesh size) and easy to implement. We prove convergence for our scheme as $\delta \to 0$ but do not derive convergence rates in δ. Instead, we investigate the achieved rates numerically, obtaining convergence rates similar to the ansatz proposed in [43].

Our proposed scheme is especially advantageous when looking at the parametrized setting with the aim to solve (1.1) for a large quantity of parameter values or for individual parameter values in a real-time context with computational and/or time constraints. In such a setting, the RB method (see for instance [76,83,112] and references therein) has become a well-known and successful model reduction scheme applied to many different types of parametrized PDEs. The main idea of the RB method is to first build reduced spaces from precomputed snapshots, i.e., solutions to the PDE for appropriately chosen parameter values, in a so-called offline phase. Then, reduced solutions are defined as (Petrov-)Galerkin projections onto the reduced spaces, which can then be evaluated significantly cheaper and faster than the original discrete solutions in the online phase, making many-query or real-time requirements possible. However, for the parametrized transport problems like (1.1) in a Petrov-Galerkin framework and with function spaces dependent on the transport direction, the application of the RB method is not directly straightforward: On the one hand, the variational framework highly depends on the parameter, on the other hand, a Petrov-Galerkin projection onto reduced spaces is again not automatically stable.

We apply our numerical scheme to the parameter-dependent case by generating fixed basis functions for the reduced test space by a greedy algorithm and then applying the (parameter-dependent) transport operator to construct the then also parameter-dependent reduced trial space. Thereby, we automatically obtain an optimally stable reduced scheme without any additional stabilization procedures that are necessary, for instance, when basing the reduced scheme on fixed trial space functions, as in [45, 139]. Therefore, our scheme is easy to implement and our numerical examples show that we obtain very efficient reduced models.

In chapter 4, we consider a kinetic Fokker-Planck equation of the form

$$\partial_t u((t,x),v) + v \cdot \nabla_x u((t,x),v) = \Delta_v \left(\frac{u((t,x),v)}{q(x,v)} \right) \quad \text{in } \Omega = I_t \times \Omega_x \times \Omega_v \tag{1.2}$$

with suitable inflow boundary conditions. The equation describes a particle density u dependent on time $t \in I_t$, position $x \in \Omega_x \subset \mathbb{R}^d$, $d \in \{2,3\}$, and velocity $v \in \Omega_v = S^{d-1}$.

While for the transport equation suitable variational formulations have already been proposed in the literature, for the kinetic Fokker-Planck equation variational formulations were mainly considered in approaches focusing on the properties of the weak solution without an orientation towards a subsequent discretization. We therefore establish a variational framework for (1.2) focusing especially on estimates for the inf-sup constant. We introduce a variational formulation in all dimensions based on Bochner-type function spaces as similarly used in [3, 34]. We then analyze the well-posedness by combining ideas developed for space-time variational formulations of parabolic equations [65, 119, 129] with our formulation for the transport equation. To show the dual inf-sup condition, we construct specific function pairs in the trial and test spaces, where the trial space function is derived from the test space function by the application of the kinetic transport operator and the inverse Laplace-Beltrami operator.

The construction of the discrete spaces for the Petrov-Galerkin projection is then based on exactly these function pairs: We first choose an arbitrary discrete test space and then define the discrete trial space by the same application of kinetic transport and inverse Laplace-Beltrami operator, such that the spaces consist of the stable function pairs from the continuous inf-sup proof.

This approach automatically yields a well-posed discrete problem with the same stability constant as for the continuous problem independently of the choice of the test space and thus of the mesh size. Our choice ensures that the spaces can be efficiently computed in the course of the numerical scheme: As for the transport equation, we apply the transport operator to the test space functions. Here, we additionally have to solve elliptic problems in the velocity domain due to the inverse Laplace-Beltrami operator. However, as these problems are low-dimensional and can be carried out in parallel, the computation of the trial space functions is not dominant in the computational costs of the solution process of the full high-dimensional equation.

1.3 Overview of the literature

There is a large variety of approaches that are concerned with the numerical solution of transport-dominated and kinetic equations and the development of suitable model reduction methods for the parametrized case. We here give an overview of methods most closely related to our work.

Finite Element Methods for transport-dominated problems Elliptic equations can be easily and successfully discretized by FE methods, where the discrete solution is usually defined as Galerkin projection of a variational formulation with a coercive bilinear form. This simple framework is however not suited for transport-dominated problems as convection-dominated convection-diffusion equations, or, especially, hyperbolic problems as, for instance, first-order transport equations: Here, simple Galerkin projections lead to strong instabilities or cannot even be used from a theoretical point of view, since standard bilinear forms for first-order hyperbolic equations are not coercive.

Therefore, many different methods have been developed to overcome especially the stability problems that arise when discretizing transport-dominated problems.

One class of stabilized (continuous) FE methods is the *streamline upwind Petrov-Galerkin method* (SUPG), also called *streamline diffusion FE method* (SDFEM). Introduced by Hughes and Brooks [23, 88], the idea of the SUPG/SDFEM method is to add artificial diffusion only in the transport direction to the equation as a stabilization. A modified framework is the *Galerkin least-squares method* (GLS), introduced by Hughes et al. e.g. for advection-diffusion equations in [89].

The *Least-squares Finite Element method* (LSFEM) is based on the formulation of weak solutions as minimizers of suitable energy functionals, most often the norm of the residual in a Hilbert space. See [20] for an extensive overview and [18] for an introduction to LSFEM for hyperbolic problems. For linear transport, LSFEM methods have been proposed and analyzed e.g. in [19,48,101,110], see subsection 3.2.4 for a short description of the used variational framework. LSFEM for the neutron transport equation is covered in [102].

Instead of using conforming continuous discrete spaces, *discontinuous Galerkin* methods (DG) are based on nonconforming discrete spaces that may be discontinuous across element interfaces. The DG method was majorly developed by Cockburn, Shu, et al., and is widely used for various problem classes. Just to mention a few works, *hp*-adaptive versions of the DG method were applied to first-order transport equations, for instance, in [16, 84, 86]. In [55], a DG approximation in space and a conforming Petrov-Galerkin approximation in time was proposed. We refer to [122] for an extensive review of the DG method and further references.

The *Discontinuous Petrov-Galerkin* (DPG) method was introduced more recently by Demkowicz and Gopalakrishnan in [51, 52, 140] and has become very popular for many different problems in the recent years. The method is both a specific DG method and a residual minimization, i.e., least-squares, method. The methodology builds on the observation that "optimal" test spaces for a Petrov-Galerkin method can be derived from trial spaces by inverting the Riesz operator of the test space. To efficiently compute approximate optimal test functions, a mesh-dependent variational formulation is introduced, which allows for "inter-element discontinuities" of the test space. Then, the optimal test functions can be approximated by localized "element-by-element" computations [52]. There is a large variety of works developing DPG methods on the one hand for different equations and on the other hand with different approaches concerning the choice of the respective norms, variational formulations, and exact approximation of the respective test functions. For linear first-order transport equations, different discretizations have already been introduced in the original works [51, 52]. With the focus of using problem-specific energy norms, an optimally stable framework has been developed in [29]. A DPG scheme focusing on uniform inf-sup stability in view of mesh-dependent variational formulations is proposed in [22], in [46] the method is enhanced by a posteriori error estimators and an adaptive strategy. In [94], the method of [22] is combined with an approximation of the flux and a discretization of a transport equation along the characteristics, while in [42] the ideas from [22, 46] are used as a basis for a model order reduction scheme for the radiative transfer equation.

Using a similar optimally conditioned variational framework as the DPG method, Dahmen, Welper, et al. introduced Petrov-Galerkin discretizations using conforming trial and test spaces for advection-diffusion [37] and first-order transport equations [43]. Instead of using non-conforming stable discrete test spaces, so-called δ-proximal near-optimal test spaces are defined. For a feasible solution process, the problem is reformulated as a saddle point problem, see also subsection 3.2.4. RB methods in this framework are proposed for parametrized transport-dominated problems in [45]. In [44], the setting from [43] is enhanced by an adaptive scheme with anisotropic meshes using *shearlets*.

Another method related to DPG formulations and the setting in [43] is the *Discrete-Dual Minimal-Residual* (DDMRes) method. In [103], the first-order transport equation is considered in an L^p-setting, and discrete solutions are defined by a residual minimization problem for special discrete dual norms.

In [9], the *saddle point least squares* (SPLS) method was proposed for abstract inf-sup stable problems and applied to a div-curl-system. There, the weak problem is reformulated as a saddle point problem (similar to [43]). Different types of trial and test spaces for this saddle point problem are explored, where, different to [43] and more similar to our method in chapter 3, first a test space and then a dependent trial space are chosen, see also subsection 3.2.4.

Model order reduction methods for parametrized transport-dominated problems Projection-based model reduction methods provide a valuable tool for the efficient solution of parametrized problems, see [15,38] for general overviews and [14] for a literature review. Among the different methodologies, the RB method has become widely used to tackle parametrized PDEs in a multi-query or real-time context, see [83, 112] and [76] (part of [15]) for specific introductions to the RB method.

After the RB method was first formulated for the easiest case of elliptic and (semi-discretized) parabolic equations in a coercive variational framework, much work has been done to develop RB approximations also for problems described by inf-sup stable mixed variational formulations.

The inf-sup stability of reduced models was first considered for saddle-point problems such as the Stokes and Navier-Stokes systems. Stable RB methods can be derived by including the computation of so-called *supremizers* which enrich the test spaces to obtain stability, see [11, 70, 114, 115].

In [128, 129], an RB method for parabolic PDEs based on a space-time variational formulation is proposed. Due to favorable lower bounds for the inf-sup constant that were first developed in [119], error estimators for the space-time formulation lead to much sharper bounds than respective results based on semidiscretizations of parabolic equations [75]. The time-dependence of the discrete spaces can be chosen such that the space-time Petrov-Galerkin projection is equivalent to a Crank-Nicholson time-stepping scheme. Then, fixed reduced spaces only in the spatial variable can be built without the need of a further stabilization. This approach was also generalized to nonlinear equations as the Burgers equation and the Boussinesq equation [137, 138]. In [82], space-time RB methods for the heat and wave equation have been considered from a matrix-based perspective.

Based on the stabilized Petrov-Galerkin framework for convection-dominated convection diffusion equations [37] and first-order transport equations [43], in [45] a reduced model based on the *double greedy algorithm* is developed: The main idea of the algorithm is to combine a greedy algorithm for the reduced trial space with iterative extensions of the test space by supremizer functions that are also determined in a greedy-like fashion. With this strategy, one obtains stable reduced models, where the test space is of larger dimension than the trial space and the solutions are determined by an associated saddle point problem. For details, see also subsection 3.3.6. A different perspective on the stabilization in Petrov-Galerkin RB frameworks is given in [139], where a preconditioner for parametrized matrices is developed which is based on the interpolation of the matrix inverse. This preconditioner can be used in to generate a stable test space for a given reduced trial space or enhance the computation of residual-based error estimators.

Apart from these works focusing on Petrov-Galerkin projections of inf-sup stable formulations, RB methods for transport-dominated problems have also been developed in other frameworks. In [77, 78], RB methods are used for finite volume discretizations of linear and nonlinear parabolic or hyperbolic evolution equations. In [109], an RB method for a convection-dominated convection-diffusion equation with SUPG stabilization is proposed. A nonlinear model reduction scheme based on least squares formulations for discrete parametrized problems (*Gauss-Newton with approximated tensors* (GNAT) method) has been developed in [32, 33]. An RB method with a posteriori error estimator for the wave equation is proposed in [73]. The model reduction method for nonlinear hyperbolic equations presented in [1] is based on the collection of snapshots in a “dictionary” combined with an L^1-minimization scheme in the online phase.

The aforementioned works propose various ways to handle model reduction methods based on the projection on linear spaces built from solution snapshots. However, for transport-dominated problems such as a linear transport equation with a moving jump discontinuity linear approximations do not perform well: Since for these problems the Kolmogorov-n-width (a measure for the approximability of the solution manifold by linear spaces) decays only very slowly, linear approximations generally need a large model order to obtain satisfying accuracy, see [107]. Therefore, in the recent years there has been an increasing interest in developing nonlinear model reduction methods that try to incorporate the movement of shocks or fronts into the method e.g. by spatial transforms. Examples include the *method of freezing* [106], *approximate Lax pairs* [68], a characteristics-based method for nonlinear conservation laws [126], a method based on optimal transport [93], methods using shifts like [113] and the *transformed snapshot interpolation* method [131–133], a method using a spatial splitting of snapshots [31], and methods using a registration of snapshots to generate a spatial transform [125, 127].

Discretizations for kinetic (Fokker-Planck) equations Kinetic equations that describe particle densities in phase space consisting of all possible physical states arise in various contexts and forms. Hence, many different numerical methods for the simulation of kinetic phenomena have evolved, see [50, 54] for general overviews.

In this work, we are especially interested in numerical methods for the kinetic Fokker-Planck equation, which is characterized by a kinetic transport operator in space and a diffusive term (only) in the velocity variable[1]. On the one hand, weak solutions and variational formulations for different types of kinetic Fokker-Planck equations have been defined and analyzed in various works, see e.g. [3,10,34,49,92,121]. However, these approaches focus on the properties of the weak solution without an orientation towards a subsequent discretization. On the other hand, discretizations of kinetic Fokker-Planck equations are often not based on the direct connection to a weak solution or do not specifically consider stability estimates.

In [99], an FE discretization of a kinetic Fokker-Planck equation is described, where the well-posedness of the discrete problem is however not analyzed. In [96], an approximation of a kinetic Fokker-Planck equation in one space dimension by plane wave expansions and finite differences is proposed. In the context of neuronal networks, a Fokker-Planck equation is discretized with finite differences in [30].

Another well-established approach to discretize kinetic equations is the method of moments, applied to Fokker-Planck equations, for instance, in [67, 97, 117, 118]. A related approach applying hierarchical model reduction methods such as [108] to a kinetic Fokker-Planck equation was developed in the author’s master thesis and published in [26].

A link between moment methods and stable discretizations based on variational formulations has been made for different variants of the radiative transfer equation in [57, 58]: A mixed variational formulations based on a parity splitting of functions is developed and a subsequent stable discretization based on a Galerkin projection onto

[1] Note that the general “Fokker-Planck equation” which describes the time evolution of a probability density function commonly amounts to a (non-kinetic) “standard” convection-diffusion equation (with diffusion in all “non-time variables”). By the “kinetic Fokker-Planck equation” we here denote equations with diffusion only in the velocity variable.

suitable combined FE-P_N spaces is proposed. It can be shown that the standard FE-P_N moment discretization is stable in the case of strictly positive absorption [57] but may be unstable for vanishing absorption [58]. This framework is applied to a generalized Fokker-Planck equation with a velocity operator of the form $(\mathrm{Id} - \alpha\Delta_v)^{-1}$ in [80].

The kinetic Fokker-Planck equation fits into the general class of PDEs with nonnegative characteristic form. For these, discontinuous Galerkin methods [85, 87, 124] and also sparse tensor approximations [120] have been developed.

We close by highlighting some interesting approaches for related kinetic equations beyond Fokker-Planck which use discretizations related to our discussion: For the neutron transport/radiative transfer equation, stabilized FE approaches have been proposed using an LSFEM discretization in [102] and using the DPG method combined with an iterative scheme in [42]. For the related Vlasov-Fokker-Planck system there are, for instance, works based on finite differences [116, 135] and also stabilized FE in form of streamline-diffusion DG approximations [4, 5]. To reduce the complexity of the discretization for these high-dimensional equations, a sparse FE method based on the stabilized setting from [102] has been introduced in [134] and enhanced in [74]. Other tensor-based methods have, for instance, been proposed for the Vlasov-Poisson system: In [98], an approximation in tensor train format is developed. A tensor decomposition using the *proper generalized decomposition* (PGD) method is developed in [59], while in [60] a low-rank decomposition in hierarchical Tucker format using a projector-splitting approach is developed.

1.4 Outline of this thesis

This thesis is organized as follows. In the following chapter 2 we introduce the motivating glioma tumor model, give a short overview on the inf-sup theory characterizing well-posed variational formulations and Petrov-Galerkin projections, and introduce some nonstandard function spaces that we use in this thesis.

Chapter 3 is devoted to the parametrized transport equation (1.1). In section 3.1 we present an optimally stable ultraweak variational formulation of first order linear transport equations. Section 3.2 is devoted to the finite-dimensional, discrete case where we introduce an optimally stable Petrov-Galerkin method. Parametrized transport problems are considered in section 3.3 within the framework of the RB method. We describe the fairly easy computational realization of the new approach in section 3.4 and report on several numerical experiments in section 3.5.

In chapter 4 we consider the kinetic Fokker-Planck equation (1.2). After a more detailed description of the setting in section 4.1, we introduce suitable Bochner-type function spaces and establish density and trace properties in section 4.2. We then derive the variational formulation and prove the existence and uniqueness results in section 4.3. In section 4.4, we introduce the discrete scheme, show well-posedness and describe an efficient computation. These properties of the proposed method are finally confirmed for a numerical example in section 4.5.

We close this thesis with concluding remarks and a short outlook in chapter 5.

2 Background

In this chapter, we discuss different concepts that build the background of this thesis. First, in section 2.1 we introduce the glioma tumor model that motivated our work on transport equations and especially on the kinetic Fokker-Planck equation. We then give an overview of the well-posedness theory of weak solutions to PDEs and their approximations by Petrov-Galerkin projections on abstract functions spaces in section 2.2. This theory will be the basis of the developed variational formulations and discretizations in chapters 3 and 4. In section 2.3, we introduce different nonstandard Sobolev-type function spaces that are important for or closely related to the appropriate function spaces for our frameworks in chapters 3 and 4.

2.1 Modeling of glioma tumor spreading

To understand and simulate the complex behavior of glioma brain tumors as, for instance, the most aggressive form *glioblastoma multiforme*, various mathematical models have been developed. Special models to describe the spreading of glioma tumor cells in the brain are multiscale models. These aim at describing the tumor cells on different levels to incorporate the relevant biological processes and then arrive at a model for the whole tumor. First, subcellular dynamics between a glioma tumor cell and the surrounding tissue are taken into account on the *microscale*. Then, the tumor cells are viewed as particles in a kinetic model, i.e., the density of tumor cells in phase space (the high-dimensional space of all possible particle states) is described by a kinetic equation on the *mesoscale*. By suitable limiting and scaling procedures, a *macroscale* model for the whole tumor in terms of a PDE in space and time is derived. Here, we briefly give an overview of one such multiscale model that was developed in [61, 62, 90, 95]; we use the notation of [90].

On the subcellular microscale level, the cell movement as well as the proliferation is governed by the bindings of cell receptors especially to the *extracellular matrix* (ECM). In a simplified model, the fraction of bound receptors is represented by a state variable $y \in [0,1]$. The complex structure of the ECM is approximated by the macroscopic volume fraction of tissue fibers $Q(x)$. Then, the reaction of receptor bindings can be described by the ODE

$$\dot{y} = k^+(1-y)Q + k^- y, \tag{2.1}$$

where $k^+, k^- > 0$ denote the (constant) reaction rates for binding and unbinding of the receptors. These bindings and unbindings happen at a much faster rate than the cell movement, so that we can assume that y equilibrates rapidly at a steady-state, which then may change only slowly dependent on the cell movement (see e.g. [40]). This steady state of (2.1) is $y^* = k^+Q/(k^+Q + k^-)$, and we define the new variable $z = y - y^* \in [-y^*, 1-y^*] =: \Omega_z$ as the deviation of y from the steady state.

We write $f(s) := (k^+ s)/(k^+ s + k^-)$, which means that $y^* = f(Q(x))$. We consider cells in position x moving with velocity v, i.e., $\frac{\mathrm{d}x}{\mathrm{d}t} = v$, therefore the steady state y^* changes in time (in the time scale of moving cells) with

$$\frac{\mathrm{d}y^*}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t} f(Q(x)) = f'(Q(x)) \nabla_x Q(x) \frac{\mathrm{d}x}{\mathrm{d}t} = f'(Q(x)) v \cdot \nabla_x Q(x).$$

Therefore, we have for z

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \dot{y} - \frac{\mathrm{d}y^*}{\mathrm{d}t} = -k^+ Q(x) z + k^- z - f'(Q(x)) v \cdot \nabla Q(x),$$

where we use that y^* is the steady state of (2.1). The change of the velocity v is modeled in [90] in two different ways. On the one hand, one can use a velocity-jump process – a Poisson process modeling cells with constant speed changing their direction according to a *turning kernel* K and with a *turning rate* λ. Such a velocity-jump process is also used e.g. in [61, 62, 95]. With the assumption that the cells align with the tissue fibers, the kernel is chosen as

$$K(x, v, v') = q(x, v),$$

where q is the tissue fiber orientation distribution. The turning rate is dependent on the fraction of bound receptors and is chosen as $\lambda(z) = \lambda_0 + \lambda_1 z > 0$.

With this, the whole equation on the mesoscale for the cell density in phase space $\Omega = [0, T] \times \Omega_x \times \Omega_z \times \Omega_v$ can be established. Incorporating the velocity jump process for the changes in direction the mesoscale equation reads (see [90, p. 22, (2.10)], [61]): Find $p(t, x, v, z)$ such that

$$\begin{aligned} \partial_t p + v \cdot \nabla_x p - \partial_z \left(\left((k^+ Q + k^-) z + f'(Q) v \cdot \nabla_x Q \right) p \right) & \\ = (\lambda_0 + \lambda_1 z) \left(q(x, v) \int_{\Omega_v} p(v') \, \mathrm{d}v' - p(v) \right). & \end{aligned} \tag{2.2}$$

As an alternative to the velocity-jump process, a Wiener process on the velocity domain is proposed in [90]. Such a Gaussian process may take into account very fast reorientations of the cell, for instance, due to irregular shape changes, better than a jump process (see [90, pp. 19-20]). The turning rate and equilibrium distribution are supposed to be the same as for the jump process case, leading to a stochastic process for the velocity change

$$\mathrm{d}v = \sqrt{\frac{2(\lambda_0 + \lambda_1 z)}{q(x, v)}} \, \mathrm{d}W_t,$$

with $(W_t)_{t \geq 0}$ being an appropriate Wiener process on the velocity domain (for details see [90, pp. 20-21]). On the mesoscale, we then alternatively obtain the Fokker-Planck equation ([90, p. 23, (2.11)]):

$$\partial_t p + v \cdot \nabla_x p - \partial_z \left(\left((k^+ Q + k^-) z + f'(Q) v \cdot \nabla_x Q \right) p \right) = (\lambda_0 + \lambda_1 z) \Delta_v \left(\frac{p}{q(x, v)} \right), \tag{2.3}$$

where Δ_v is the Laplace-Beltrami operator on the velocity domain Ω_v.

These basic models on the mesoscale can be enhanced by many other effects. In [62, 63], proliferation is modeled, while in [41, 53] also the influence of acidity levels of the tissue on the cell motility is included. A model taking the therapy into account is developed in [91].

After developing the mesoscale model describing the density of tumor cells in the high-dimensional phase space, macroscale equations for the tumor evolution depending only on space and time can be derived by suitable limiting procedures, mostly a parabolic scaling. Thereby, the mesoscale equation (2.2) can be scaled to a macroscale reaction-diffusion equation with a *tumor diffusion tensor* and a *tumor drift velocity* dependent on the tissue fiber orientation q, and an additional haptotactic-like drift term including the gradient of the fiber volume fraction Q, see [61, 90]. A scaling of (2.3) leads to a comparable reaction-diffusion equation with slightly different coefficients, see [90]. For the more elaborate models, the macroscopic equations contain, for instance, additional (logistic) growth terms for the proliferation (see [62, 63]), and drift terms accounting for the influence of acidity (see [41, 53]). In the detailed models of the last two works, the acidity, necrosis (in [41]), vascularization (in [53]), and possibly also q and Q can all be considered to be changing in time, as well. Hence, the macroscale models contain systems of equations for the different quantities.

In the mentioned works, numerical simulations of the different variants of the macroscale equations are performed. To that end, the tissue volume fraction Q and the fiber orientation distribution q are computed from diffusion tensor imaging (DTI) scans of the human brain, making patient-specific simulations possible (see [61] and [90, Chap. 3] for details). The PDE or system of PDEs is then discretized by the discontinuous Galerkin method (e.g. in [61]) or finite volume method (e.g. in [41, 53, 62]). Numerical experiments show that the models can reproduce many observed phenomena such as the finger-like invasion patterns of glioma [61].

While such simulations of the macroscopic limit equations have been performed and compared in many different varieties, there are less results for a different approach to simulate the tumor spread: In the described settings the mesoscopic equations are the most detailed models for the tumors, while the macroscopic limits are analytical approximations relying on the choice of the limiting procedure, for instance, parabolic or hyperbolic scaling. To assess the validity of these limits, one could also opt for directly discretizing the mesoscopic equations themselves. The major challenge here lies in the large dimension of the phase space – for the mesoscopic equations (2.2) and (2.3) the phase space $\Omega = [0, T] \times \Omega_x \times \Omega_z \times \Omega_v$ is seven-dimensional when considering a full model in 3D space. In [40], a discretization of (2.2) by the *method of moments* is developed and compared to the macroscopic limit by a parabolic scaling from [61]. Dependent on the model parameters, a convergence of the moment approximation to the macroscopic diffusion limit can be observed, while in other regimes indeed a low order moment approximation can be more accurate than the diffusion limit.

These results show that it can be valuable to concentrate on the simulation of the mesoscopic kinetic equations. While the jump process version of the mesoscopic model (2.2) has been simulated in [39, 40], in chapter 4 we will be concerned with the discretization of kinetic Fokker-Planck equations inspired by (2.3).

2.2 Abstract well-posedness theory

All discretizations considered in this thesis will be based on Petrov-Galerkin projections of weak solutions defined by suitable variational formulations. Therefore, we here briefly recall the well-posedness theory on the infinite dimensional as well as on the discrete level using abstract function spaces.

2.2.1 Inf-sup theory

In this section, we introduce an abstract problem on (possibly) infinite-dimensional function spaces. The well-posedness of this problem is analyzed with functional analytical tools in the so-called inf-sup theory. In its center lies the Banach-Nečas-Babuška theorem (also called inf-sup or generalized Lax-Milgram theorem, see e.g. [65, Thm. 2.6]), that goes back to works of Nečas [104] and Babuška [8]. We here collect the respective results needed in the following chapters. For more extensive overviews we refer, for instance, to [65, sect. 2.1, A.2] and [105, sect. 2.3].

Let $(\mathcal{X}, \|\cdot\|_{\mathcal{X}}$ and $(\mathcal{Y}, \|\cdot\|_{\mathcal{Y}})$ be two reflexive Banach spaces with dual spaces $\mathcal{X}'$ and $\mathcal{Y}'$. We denote the dual pairings by $\langle\cdot,\cdot\rangle_{\mathcal{X}',\mathcal{X}}$ and $\langle\cdot,\cdot\rangle_{\mathcal{Y}',\mathcal{Y}}$.

Let $b : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be a continuous bilinear form with continuity constant[1]

$$\sup_{w\in\mathcal{X}} \sup_{p\in\mathcal{Y}} \frac{b(w,p)}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \gamma. \tag{2.4}$$

We consider the associated operator $B : \mathcal{X} \to \mathcal{Y}'$ and adjoint operator $B^* : \mathcal{Y} \to \mathcal{X}'$ defined by $\langle Bw, p\rangle_{\mathcal{Y}',\mathcal{Y}} = b(w,p) = \langle w, B^*p\rangle_{\mathcal{X}',\mathcal{X}}$ for all $w \in \mathcal{X}$, $p \in \mathcal{Y}$. Then, from (2.4) it directly follows that B and B^* are bounded linear operators with $\|B\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} = \|B^*\|_{\mathcal{L}(\mathcal{Y},\mathcal{X}')} = \gamma$.

Let $f \in \mathcal{Y}'$. Then, the variational problem is defined as follows: Seek $u \in \mathcal{X}$ such that

$$b(u,p) = f(p) \quad \forall p \in \mathcal{Y}. \tag{2.5}$$

We want to find conditions that ascertain the *well-posedness* of the variational problem in the sense of Hadamard, i.e., that (2.5) admits a unique solution which depends continuously on f. From the definition of B, we see that (2.5) is equivalent to the operator equation

$$Bu = f \quad \text{ in } \mathcal{Y}'.$$

Therefore, existence and uniqueness of u in (2.5) for arbitrary right-hand sides $f \in \mathcal{Y}'$ is equivalent to the bijectivity of B.

There are many different equivalent conditions for injectivity and surjectivity of Banach operators. A useful characterization is given in the following proposition:

Proposition 2.2.1. *For $B \in \mathcal{L}(\mathcal{X}, \mathcal{Y}')$ as above, the following statements are equivalent:*

(i) *$B : \mathcal{X} \to \mathcal{Y}'$ is surjective, i.e., $\mathrm{im}(B) = \mathcal{Y}'$.*

(ii) *$B^* : \mathcal{Y} \to \mathcal{X}'$ is injective, i.e., $\ker(B^*) = \{0\}$, and $\mathrm{im}(B^*)$ is closed in $\mathcal{X}'$.*

(iii) *There exists $\beta > 0$ such that*

$$\|B^*p\|_{\mathcal{X}'} \geq \beta\|p\|_{\mathcal{Y}} \quad \forall p \in \mathcal{Y}.$$

[1] We slightly abuse notation: Here and throughout the thesis all inf and sup are taken over nonzero vectors.

(*iv*) *There exists* $\beta > 0$ *such that*

$$\inf_{p\in\mathcal{Y}} \sup_{w\in\mathcal{X}} \frac{\langle w, B^*p\rangle_{\mathcal{X},\mathcal{X}'}}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} \geq \beta.$$

Proof. (iii) and (iv) are clearly reformulations of the same condition using the definition of $\|\cdot\|_{\mathcal{Y}'}$, hence they are equivalent. The equivalence of (i), (ii), and (iii) is proved in [21, Thm 2.20, p. 47]: For $(i) \Rightarrow (iii)$, it is shown that the set $M := \{p \in \mathcal{Y} : \|B^*p\|_{\mathcal{X}'} \leq 1\}$ is bounded in $\mathcal{Y}$. To that end, dual pairings of $f \in \mathcal{Y}'$ and $p \in M$ are estimated by writing any f as Bu for some $u \in \mathcal{X}$. For $(iii) \Rightarrow (ii)$, injectivity can be seen by setting $p = 0$ in (iii) by linearity of B^* and closedness of $\mathcal{X}'$. To show closedness of $\operatorname{im}(B^*)$, one sees from (iii) that the preimage of a Cauchy sequence in $\operatorname{im}(B^*)$ is a Cauchy sequence in $\mathcal{Y}$. $(ii) \Rightarrow (i)$ follows from Banach's closed range theorem (see [21, Thm. 2.19, p. 46]). □

The same result also holds when interchanging the roles of B and B^*:

Proposition 2.2.2. *For* $B \in \mathcal{L}(\mathcal{X}, \mathcal{Y}')$ *as above, the following statements are equivalent:*

(*i*) $B^* : \mathcal{Y} \to \mathcal{X}'$ *is surjective.*

(*ii*) $B : \mathcal{X} \to \mathcal{Y}'$ *is injective and* $\operatorname{im}(B)$ *is closed in* $\mathcal{Y}'$.

(*iii*) *There exists* $\beta \geq 0$ *such that*

$$\|Bw\|_{\mathcal{Y}'} \geq \beta\|w\|_{\mathcal{X}} \quad \forall w \in \mathcal{X}.$$

(*iv*) *There exists* $\beta \geq 0$ *such that*

$$\inf_{w\in\mathcal{X}} \sup_{p\in\mathcal{Y}} \frac{\langle Bw, p\rangle_{\mathcal{Y}',\mathcal{Y}}}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} \geq \beta.$$

Proof. Since $\mathcal{X}$ and $\mathcal{Y}$ are reflexive, it holds $(B^*)^* = B$. Therefore, the claim directly follows from Proposition 2.2.1. □

In the last two propositions, the inf-sup constant and the dual inf-sup constant were defined independently of each other, which raises the question whether these constants are related. In fact, it turns out that the inf-sup and dual inf-sup constants are equal, if the corresponding operators are bijective.

Proposition 2.2.3. *Let* $B \in \mathcal{L}(\mathcal{X}, \mathcal{Y}')$ *be bijective. Then,* $B^* \in \mathcal{L}(\mathcal{Y}, \mathcal{X}')$ *is bijective, the inf-sup condition and dual inf-sup condition hold with equal constants*

$$\inf_{w\in\mathcal{X}} \sup_{p\in\mathcal{Y}} \frac{b(w,p)}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \inf_{p\in\mathcal{Y}} \sup_{w\in\mathcal{X}} \frac{b(w,p)}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \beta > 0,$$

and the inverse operator and inverse adjoint operator are continuous with

$$\|B^{-1}\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} = \|B^{-*}\|_{\mathcal{L}(\mathcal{X}',\mathcal{Y})} = \frac{1}{\beta}.$$

Proof. As B is bijective, consider the inverse operator $B^{-1} : \mathcal{Y}' \to \mathcal{X}$ and its adjoint $(B^{-1})^* : \mathcal{X}' \to \mathcal{Y}$. It then holds for all $u' \in \mathcal{X}'$ and $w \in \mathcal{X}$

$$\langle B^*(B^{-1})^*u', w\rangle_{\mathcal{X}',\mathcal{X}} = \langle (B^{-1})^*u', Bw\rangle_{\mathcal{Y},\mathcal{Y}'} = \langle u', B^{-1}Bw\rangle_{\mathcal{X}',\mathcal{X}} = \langle u', w\rangle_{\mathcal{X}',\mathcal{X}},$$

and it holds for all $q' \in \mathcal{Y}'$ and $p \in \mathcal{Y}$

$$\langle (B^{-1})^* B^* p, q' \rangle_{\mathcal{Y},\mathcal{Y}'} = \langle B^* p, B^{-1} q' \rangle_{\mathcal{X}',\mathcal{X}} = \langle p, B B^{-1} q' \rangle_{\mathcal{Y},\mathcal{Y}'} = \langle p, q' \rangle_{\mathcal{Y},\mathcal{Y}'}.$$

Hence, B^* is bijective with inverse operator $(B^*)^{-1} = (B^{-1})^* : \mathcal{X}' \to \mathcal{Y}$, which we will call from now on B^{-*}. As B is bijective, due to Proposition 2.2.2 there holds an inf-sup condition with constant β_B. We obtain

$$\begin{aligned}\beta_B = \inf_{w\in\mathcal{X}} \sup_{p\in\mathcal{Y}} \frac{b(w,p)}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} &= \inf_{w\in\mathcal{X}} \sup_{p\in\mathcal{Y}} \frac{\langle Bw, p\rangle_{\mathcal{Y}',\mathcal{Y}}}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \inf_{w\in\mathcal{X}} \frac{\|Bw\|_{\mathcal{Y}'}}{\|w\|_{\mathcal{X}}} \\ &= \inf_{z\in\mathcal{Y}'} \frac{\|z\|_{\mathcal{Y}'}}{\|B^{-1}z\|_{\mathcal{X}}} = \left(\sup_{z\in\mathcal{Y}'} \frac{\|B^{-1}z\|_{\mathcal{X}}}{\|z\|_{\mathcal{Y}'}} \right)^{-1} = \left(\|B^{-1}\|_{\mathcal{L}(\mathcal{Y}',\mathcal{X})} \right)^{-1}.\end{aligned}$$

Similarly, for B^* there holds an inf-sup condition with constant β_{B^*} due to Proposition 2.2.1 and we obtain

$$\beta_{B^*} = \inf_{p\in\mathcal{Y}} \sup_{w\in\mathcal{X}} \frac{b(w,p)}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \inf_{p\in\mathcal{Y}} \sup_{w\in\mathcal{X}} \frac{\langle w, B^*p\rangle_{\mathcal{X},\mathcal{X}'}}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \left(\|B^{-*}\|_{\mathcal{L}(\mathcal{X}',\mathcal{Y})} \right)^{-1}.$$

As B^{-1} and B^{-*} are adjoint operators, they have the same norm, which means that in fact $\beta_B = \beta_{B^*} =: \beta$, and the claim follows. □

With these results we can now state the most commonly used result to characterize a well-posed variational formulation:

Theorem 2.2.4 (Well-posedness). *Let $B \in \mathcal{L}(\mathcal{X}, \mathcal{Y}')$ and $f \in \mathcal{Y}'$ as above. If either the inf-sup and surjectivity conditions for B*

$$(B1) \qquad \inf_{w\in\mathcal{X}} \sup_{p\in\mathcal{Y}} \frac{b(w,p)}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \beta > 0$$

$$(B2) \qquad \sup_{w\in\mathcal{X}} b(w,p) > 0 \quad \forall p \in \mathcal{Y}, p \neq 0$$

or the inf-sup and surjectivity conditions for the adjoint operator B^* (called dual inf-sup condition *and* dual surjectivity condition*)*

$$(B^*1) \qquad \inf_{p\in\mathcal{Y}} \sup_{w\in\mathcal{X}} \frac{b(w,p)}{\|w\|_{\mathcal{X}}\|p\|_{\mathcal{Y}}} = \beta > 0$$

$$(B^*2) \qquad \sup_{p\in\mathcal{Y}} b(w,p) > 0 \quad \forall w \in \mathcal{X}, w \neq 0$$

hold, then the variational problem (2.5) has a unique solution $u \in \mathcal{X}$ which satisfies the a priori estimate

$$\|u\|_{\mathcal{X}} \leq \frac{1}{\beta} \|f\|_{\mathcal{Y}'}.$$

Proof. By Proposition 2.2.2, $(B1)$ leads to injectivity of B with $\mathrm{im}(B)$ closed. From $(B2)$ we see that for all $0 \neq p \in \mathcal{Y}$, we have $B^*p \neq 0$ in $\mathcal{X}'$, i.e., B^* is injective. By Banach's closed range theorem (see e.g. [21, Thm. 2.19, p. 46]), $\mathrm{im}(B^*)$ is closed since $\mathrm{im}(B)$ is closed. Therefore, by Proposition 2.2.1, B is surjective, and thus bijective. Analogously, one can show that (B^*1) and (B^*2) imply bijectivity of B^*.

In both cases, by Proposition 2.2.3 both B and B^* are bijective, and we obtain from the primal or dual inf-sup condition with constant β that $\|B^{-1}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')} = \frac{1}{\beta}$. Hence, there exists a unique solution $u \in \mathcal{X}$ to (2.5) satisfying the a priori estimate

$$\|u\|_{\mathcal{X}} = \|B^{-1}f\|_{\mathcal{X}} \leq \|B^{-1}\|_{\mathcal{L}(\mathcal{X},\mathcal{Y}')}\|f\|_{\mathcal{Y}'} \leq \frac{1}{\beta}\|f\|_{\mathcal{Y}'}. \qquad \square$$

Remark 2.2.5. Theorem 2.2.4 is an extended version of the Banach-Nečas-Babuška theorem (going back to Nečas [104] and Babuška [8], see also e.g. [65, Thm. 2.6]). We reformulated the theorem to make it possible to show the dual inf-sup and surjectivity conditions instead of the “standard” primal conditions.

Remark 2.2.6. It can be seen from the definition of the conditions that both inf-sup conditions imply the respective other surjectivity condition, i.e., $(B1) \implies (B^*2)$ and $(B2) \implies (B^*1)$. Therefore, well-posedness also follows if both inf-sup conditions $(B1)$ and (B^*1) can be shown. However, as this is usually more elaborate, the most commonly used conditions are the ones given in Theorem 2.2.4.

2.2.2 Petrov-Galerkin projections

We now develop the well-posedness theory for discrete problems. We build approximations to the solution of (2.5) by a Petrov-Galerkin projection. To that end, let $\mathcal{X}^\delta \subset \mathcal{X}$ and $\mathcal{Y}^\delta \subset \mathcal{Y}$ be two discrete spaces with the same dimension $\dim(\mathcal{X}^\delta) = \dim(\mathcal{Y}^\delta) = N^\delta < \infty$. Then, the Petrov-Galerkin approximation $u^\delta \in \mathcal{X}^\delta$ is defined by

$$b(u^\delta, p^\delta) = f(p^\delta) \quad \forall p^\delta \in \mathcal{Y}^\delta. \tag{2.6}$$

Well-posedness of the Petrov-Galerkin approximation follows similarly to the infinite-dimensional case:

Proposition 2.2.7 (Well-posedness)**.** *Let $\mathcal{X}^\delta \subset \mathcal{X}$ and $\mathcal{Y}^\delta \subset \mathcal{Y}$ with $\dim(\mathcal{X}^\delta) = \dim(\mathcal{Y}^\delta) = N^\delta < \infty$. Then, the discrete inf-sup condition*

$$\inf_{w^\delta \in \mathcal{X}^\delta} \sup_{p^\delta \in \mathcal{Y}^\delta} \frac{b(w^\delta, p^\delta)}{\|w^\delta\|_{\mathcal{X}}\|p^\delta\|_{\mathcal{Y}}} \geq \beta_\delta > 0 \tag{2.7}$$

and the discrete dual inf-sup condition

$$\inf_{p^\delta \in \mathcal{Y}^\delta} \sup_{w^\delta \in \mathcal{X}^\delta} \frac{b(w^\delta, p^\delta)}{\|w^\delta\|_{\mathcal{X}}\|p^\delta\|_{\mathcal{Y}}} \geq \beta_\delta > 0 \tag{2.8}$$

are equivalent. If the conditions hold, then there exists a unique solution $u^\delta \in \mathcal{X}^\delta$ to (2.6) that satisfies the stability estimate

$$\|u^\delta\|_{\mathcal{X}} \leq \frac{1}{\beta_\delta}\|f\|_{(\mathcal{Y}^\delta)'} \leq \frac{1}{\beta_\delta}\|f\|_{\mathcal{Y}'}.$$

Proof. We can use the results from subsection 2.2.1 for $b|_{\mathcal{X}^\delta \times \mathcal{Y}^\delta}$ with the associated operators $B^\delta : \mathcal{X}^\delta \to (\mathcal{Y}^\delta)'$ and $B^{*,\delta} : \mathcal{Y}^\delta \to (\mathcal{X}^\delta)'$. If (2.7) holds, from Proposition 2.2.2 we see that $B^\delta : \mathcal{X}^\delta \to (\mathcal{Y}^\delta)'$ is injective and $B^{*,\delta} : \mathcal{Y}^\delta \to (\mathcal{X}^\delta)'$ is surjective. As

$\dim(\mathcal{X}^\delta) = \dim(\mathcal{Y}^\delta) = N^\delta < \infty$, B^δ and $B^{*,\delta}$ are therefore already bijective. Analogously, (2.8) implies by Proposition 2.2.1 that B^δ is surjective and $B^{*,\delta}$ is injective. Hence, again both operators are bijective. Since in both cases the operators are bijective, Proposition 2.2.3 shows that both (2.7) and (2.8) hold with the same constant β_δ, i.e., (2.7) and (2.8) are equivalent.

If the conditions hold and thus B^δ is bijective, (2.6) has a unique solution $u^\delta \in \mathcal{X}^\delta$. Proposition 2.2.3 shows that $\|B^{-1}\|_{\mathcal{L}(\mathcal{X}^\delta,(\mathcal{Y}^\delta)')} = \frac{1}{\beta_\delta}$, and we obtain the stability estimate from Theorem 2.2.4

$$\|u^\delta\|_{\mathcal{X}} \leq \|(B^\delta)^{-1}\|_{\mathcal{L}((\mathcal{Y}^\delta)',\mathcal{X}^\delta)}\|f\|_{(\mathcal{Y}^\delta)'} = \frac{1}{\beta_\delta}\|f\|_{(\mathcal{Y}^\delta)'} \leq \frac{1}{\beta_\delta}\|f\|_{\mathcal{Y}'}. \qquad \square$$

Remark 2.2.8. Note that the discrete inf-sup condition (or dual inf-sup condition) generally does *not* follow from the respective conditions for the infinite-dimensional spaces even in this conforming setting, i.e., $\mathcal{X}^\delta \subset \mathcal{X}$ and $\mathcal{Y}^\delta \subset \mathcal{Y}$. Therefore, in the construction of a numerical scheme, the discrete inf-sup condition has to be taken into account. This is different from Galerkin approximations of variational formulations based on a coercive bilinear form with the same trial and test space, where well-posedness for any conforming discrete space follows from the well-posedness of the infinite-dimensional problem by the Lax-Milgram theorem. On the other hand, while we need two conditions for the well-posedness of the weak solution in Theorem 2.2.4 (inf-sup and surjectivity, or, alternatively, inf-sup *and* dual inf-sup), for the Petrov-Galerkin projection, one condition (e.g., inf-sup *or* dual inf-sup) already ensures well-posedness due to the finite dimension of the spaces.

The setting also gives rise to an a priori error estimate similar to Céa's lemma:

Proposition 2.2.9 (Quasi-best approximation)**.** *Let the variational problem* (2.5) *be well-posed and let the discrete inf-sup condition* (2.7) *hold. Let* $u \in \mathcal{X}$ *be the solution to* (2.5) *and* $u^\delta \in \mathcal{X}^\delta$ *be the Petrov-Galerkin approximation defined in* (2.6)*. Then, we have the quasi-best approximation property ´*

$$\|u - u^\delta\|_{\mathcal{X}} \leq \left(1 + \frac{\gamma}{\beta_\delta}\right) \inf_{w^\delta \in \mathcal{X}^\delta} \|u - w^\delta\|_{\mathcal{X}}. \tag{2.9}$$

If $\mathcal{X}$ *and* $\mathcal{Y}$ *are Hilbert spaces we further have*

$$\|u - u^\delta\|_{\mathcal{X}} \leq \frac{\gamma}{\beta_\delta} \inf_{w^\delta \in \mathcal{X}^\delta} \|u - w^\delta\|_{\mathcal{X}}. \tag{2.10}$$

Proof. The first estimate is from Babuška [8], the enhancement was proved by Xu and Zikatanov [136]. From the definitions of u and u^δ we obtain the Galerkin orthogonality

$$b(u - u^\delta, p^\delta) = b(u, p^\delta) - b(u^\delta, p^\delta) = f(p^\delta) - f(p^\delta) = 0 \quad \forall p^\delta \in \mathcal{Y}^\delta. \tag{2.11}$$

Let $w^\delta \in \mathcal{X}^\delta$. Using the inf-sup condition (2.7) and the Galerkin orthogonality (2.11) yields

$$\beta_\delta \|u^\delta - w^\delta\|_{\mathcal{X}} \leq \sup_{p^\delta \in \mathcal{Y}^\delta} \frac{b(u^\delta - w^\delta, p^\delta)}{\|p^\delta\|_{\mathcal{Y}}} = \sup_{p^\delta \in \mathcal{Y}^\delta} \frac{b(u - w^\delta, p^\delta)}{\|p^\delta\|_{\mathcal{Y}}} \leq \gamma \|u - w^\delta\|_{\mathcal{X}}. \tag{2.12}$$

Hence, it holds

$$\|u - u^\delta\|_\mathcal{X} \leq \|u - w^\delta\|_\mathcal{X} + \|w^\delta - u^\delta\|_\mathcal{X} \leq \left(1 + \frac{\gamma}{\beta_\delta}\right) \|u - w^\delta\|_\mathcal{X},$$

and thus, since $w^\delta \in \mathcal{Y}^\delta$ was arbitrary, (2.9).

For the sharper estimate, Xu and Zikatanov [136] consider the (Petrov-Galerkin) projection operator $P^\delta : \mathcal{X} \to \mathcal{X}^\delta$ defined for each $w \in \mathcal{X}$ by $b(P^\delta w, p^\delta) = b(w, p^\delta)$ for all $p^\delta \in \mathcal{Y}^\delta$, i.e., $P^\delta u = u^\delta$. It is shown that

$$\|P^\delta\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} = \|I - P^\delta\|_{\mathcal{L}(\mathcal{X},\mathcal{X})}$$

(see [136, Lemma 5]). Then, we obtain for any $w^\delta \in \mathcal{X}^\delta$

$$\|u - u^\delta\|_\mathcal{X} = \|(I - P^\delta)(u - w^\delta)\|_\mathcal{X} \leq \|I - P^\delta\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \|u - w^\delta\|_\mathcal{X} = \|P^\delta\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \|u - w^\delta\|_\mathcal{X}.$$

Setting $w^\delta = 0$ in (2.12) we have $\|P^\delta u\|_\mathcal{X} = \|u^\delta\|_\mathcal{X} \leq \frac{\gamma}{\beta_\delta}\|u\|_\mathcal{X}$, i.e., $\|P^\delta\|_{\mathcal{L}(\mathcal{X},\mathcal{X})} \leq \frac{\gamma}{\beta_\delta}$, which yields (2.10). □

Remark 2.2.10. Through this estimate we see that the discrete inf-sup constant of the chosen discrete spaces is crucial for the quality of the Petrov-Galerkin approximation: For a small approximation error, it is desirable to choose discrete spaces with a good approximation property as well as with a discrete inf-sup constant β^δ as large as possible.

2.3 Sobolev spaces for transport and kinetic equations

Sobolev spaces are Banach spaces of functions possessing weak derivatives and are thus the suitable spaces to build variational formulations for PDEs with the framework given in section 2.2. For the basic concepts and properties of the standard Sobolev spaces $W^{k,p}(\Omega)$ we refer to the literature, e.g., [2, 66]. Here we briefly introduce some specific Sobolev-type spaces used for different transport and kinetic equations that will be important when introducing and discussing the appropriate function spaces used for the variational formulations of the linear transport equation in chapter 3 and the kinetic Fokker-Planck equation in chapter 4.

2.3.1 Spaces for first-order transport equations

For elliptic equations, one usually defines weak solutions in the standard isotropic Sobolev space $H^1(\Omega)$ for some domain $\Omega \subset \mathbb{R}^d$. Here, a function $u \in H^1(\Omega)$ possesses a weak gradient, i.e., a weak derivative $\partial_i u \in L^2(\Omega)$ for $i = 1, \ldots, n$.

While this is the right choice for equations with diffusion (in all dimensions), transport equations without diffusion only contain directional derivatives of the solution in the (possibly space-dependent) transport direction. Therefore, special anisotropic Sobolev spaces of functions that possess exactly these directional derivatives are the appropriate choice for the weak formulation. The major well-posedness theory for weak solutions to first-order transport equations in an L^2-setting was developed by C. Bardos in [12]. We here briefly collect definitions and properties of standard spaces used to describe first-order transport equations, following [71], where some of the results from [12] are slightly generalized to less regular data functions.

Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary $\partial\Omega$, and let $\mathbf{n}$ be the unit outer normal to $\partial\Omega$, which is defined for almost all $x \in \partial\Omega$ and is measurable on $\partial\Omega$. Let $\mathbf{b} : \bar{\Omega} \to \mathbb{R}^d$ be a Lipschitz vector field, which will denote the transport direction. We define the inflow and outflow boundaries Γ_- and Γ_+ by

$$\Gamma_- = \{x \in \partial\Omega \setminus \Xi : \mathbf{b} \cdot \mathbf{n} < 0\} \subset \partial\Omega,$$
$$\Gamma_+ = \{x \in \partial\Omega \setminus \Xi : \mathbf{b} \cdot \mathbf{n} > 0\} \subset \partial\Omega,$$

where Ξ denotes the set where $\mathbf{n}$ is not defined. The characteristic or no-flow boundary Γ_0 is likewise defined by

$$\Gamma_0 = \{x \in \partial\Omega \setminus \Xi : \mathbf{b} \cdot \mathbf{n} = 0\} \subset \partial\Omega.$$

We assume the following hypothesis for the boundary parts:

(H1) The sets $\Gamma_\pm$ and $(\overline{\Gamma_\pm})^\circ$ (where the closure and interior are taken with respect to $\partial\Omega$) have the same surface Lebesgue measure.

If $\partial\Omega$ is piecewise C^1, then (H1) is automatically satisfied. From now on, we redefine the inflow and outflow boundaries as $\Gamma_\pm := (\overline{\Gamma_\pm})^\circ$.

We define the space

$$H(\Omega, \mathbf{b}) := \{w \in L^2(\Omega) : \mathbf{b} \cdot \nabla w \in L^2(\Omega)\}, \tag{2.13}$$

and, given $K \subset \partial\Omega$ we define the boundary space $L^2(K, |\mathbf{b} \cdot \mathbf{n}|)$ with norm

$$\|w\|_{L^2(K,|\mathbf{b}\cdot\mathbf{n}|)} := \left(\int_K |w|^2 |\mathbf{b} \cdot \mathbf{n}| \, \mathrm{d}s \right)^{\frac{1}{2}}.$$

It then holds the following proposition from [71]:

Proposition 2.3.1 ([71, Proposition I.1]). *(i) $H(\Omega, \mathbf{b})$ is a Hilbert space under the graph norm $\|\cdot\|_{H(\Omega,\mathbf{b})}$ given by*

$$\|w\|^2_{H(\Omega,\mathbf{b})} := \|w\|^2_{L^2(\Omega)} + \|\mathbf{b} \cdot \nabla w\|^2_{L^2(\Omega)}.$$

(ii) $C^\infty(\bar{\Omega})$, and hence $C^{0,1}(\Omega)$, is dense in $H(\Omega, \mathbf{b})$.

(iii) The trace mapping $\chi_{\partial\Omega} : w \mapsto (\mathbf{b}\cdot\mathbf{n})w$ on $C^\infty(\bar{\Omega})$ admits an extension as a bounded linear operator from $H(\Omega, \mathbf{b})$ to $H^{-\frac{1}{2}}(\partial\Omega)$[2].

(iv) The integration by parts or Green's formula for $w \in C^\infty(\bar{\Omega})$, $p \in H^1(\Omega)$

$$\int_\Omega (\mathbf{b} \cdot \nabla w) p \, \mathrm{d}x = \int_{\partial\Omega} (\mathbf{b} \cdot \mathbf{n}) w p \, \mathrm{d}s - \int_\Omega w \nabla \cdot (\mathbf{b} p) \, \mathrm{d}x$$

admits an extension for all $w \in H(\Omega, \mathbf{b})$, $p \in H^1(\Omega)$, where the surface integral is replaced by the duality pairing $\langle \cdot, \cdot \rangle_{H^{-\frac{1}{2}}(\partial\Omega), H^{\frac{1}{2}}(\partial\Omega)}$.

(v) Under hypotheses (H1), the trace mappings $u \mapsto u|_{\Gamma_\pm}$ are linear and continuous from $H(\Omega, \mathbf{b})$ to $L^2_{\mathrm{loc}}(\Gamma_\pm, |\mathbf{b} \cdot \mathbf{n}|)$

[2] $H^{-\frac{1}{2}}(\partial\Omega)$ is the dual space of $H^{\frac{1}{2}}(\partial\Omega)$. Note that $H^1(\Omega)$ functions have a trace in $H^{\frac{1}{2}}(\partial\Omega)$, see e.g. [2, Thm. 7.39].

With this result, we see that it makes sense to assign trace values to a function $w \in H(\Omega, \mathbf{b})$. However, in general the trace is indeed only locally in L^2 as the following example from [12] shows:

Example 2.3.2. Consider $\Omega = (-1,1) \times (0,1) \subset \mathbb{R}^2$ and $\mathbf{b}(x_1, x_2) = (-1, x_1)^T$. Then, for $\alpha \in (-\frac{3}{4}, -\frac{1}{2})$ the function $w(x_1, x_2) = (x_2 + \frac{x_1^2}{2})^\alpha$ lies in $H(\Omega, \mathbf{b})$, but $w|_{\Gamma_\pm} \notin L^2(\Gamma_\pm, |\mathbf{b} \cdot \mathbf{n}|)$.

We now define spaces with trace zero on the inflow or outflow domain:

$$\begin{aligned} H_{\Gamma_-}(\Omega, \mathbf{b}) &= \{w \in L^2(\Omega) : \mathbf{b} \cdot \nabla w \in L^2(\Omega), w|_{\Gamma_-} = 0\}, \\ H_{\Gamma_+}(\Omega, \mathbf{b}) &= \{w \in L^2(\Omega) : \mathbf{b} \cdot \nabla w \in L^2(\Omega), w|_{\Gamma_+} = 0\}. \end{aligned} \tag{2.14}$$

In [71] (and, similarly, also in [12]) it is shown that functions in $H_{\Gamma_\pm}(\Omega, \mathbf{b})$ can be approximated by smooth functions. To that end, intermediate approximations are defined. In the following, we replicate the statements from [71] with short sketches of the given proofs:

Lemma 2.3.3. *$L^\infty(\Omega) \cap H_{\Gamma_\pm}(\Omega, \mathbf{b})$ is dense in $H_{\Gamma_\pm}(\Omega, \mathbf{b})$.*

Proof. (Sketch, see [71, Prop. I.1, Lemma I.7] for details). Define a "value cutoff function" $g_M \in C^{0,1}(\mathbb{R})$, $g_M(t) = t$ for $|t| < M$, $g_M(t) = \text{sign}(t) M$ for $|t| \geq M$. Then, for $w \in H_{\Gamma_\pm}(\Omega, \mathbf{b})$ show that $g_M(u)$ converges weakly to u as $M \to \infty$. Use Mazur's lemma to show the claim. □

In a second step, functions in $L^\infty(\Omega) \cap H_{\Gamma_\pm}(\Omega, \mathbf{b})$ are approximated by smooth functions that vanish near the possible singularities at the boundary of Γ_- or Γ_+: To that end, define

$$\widetilde{H_{\Gamma_\pm}}(\Omega, \mathbf{b}) = \{w \in L^\infty(\Omega) \cap H_{\Gamma_\pm}(\Omega, \mathbf{b}) : w(x) = 0 \text{ in a neighborhood of } \partial(\Gamma_\pm)\},$$

where $\partial(\Gamma_\pm)$ is the boundary of $\Gamma_\pm$ in $\partial\Omega$. We require an additional assumption for these boundaries:

(H2) The boundary $\partial(\Gamma_\pm)$ of $\Gamma_\pm$ in $\partial\Omega$ has a finite $(d-2)$-dimensional Hausdorff measure.

Lemma 2.3.4. *If (H2) holds, then $\widetilde{H_{\Gamma_\pm}}(\Omega, \mathbf{b})$ is dense in $H_{\Gamma_\pm}(\Omega, \mathbf{b})$.*

Proof. (Sketch, see [71, Lemma I.9] and [12, p. 202-203] for details). Define a specific cutoff function $\phi_\varepsilon \in C^{0,1}(\bar{\Omega})$ that satisfies $0 \leq \phi_\varepsilon \leq 1$ and

$$\begin{aligned} \phi_\varepsilon &= 0 \quad \text{in } \{x \in \bar{\Omega} : \text{dist}(x, \partial(\Gamma_\pm)) \leq \tfrac{\varepsilon}{2}\}, \\ \phi_\varepsilon &= 1 \quad \text{in } \{x \in \bar{\Omega} : \text{dist}(x, \partial(\Gamma_\pm)) \geq \varepsilon\}, \\ |\nabla \phi_\varepsilon| &\leq \tfrac{C}{\varepsilon} \quad \text{in } \bar{\Omega}. \end{aligned} \tag{2.15}$$

Using (H2) one can show that $\|\nabla \phi_\varepsilon\|_{L^2(\Omega)}$ is bounded uniformly in ε, since the gradient is bounded by $\frac{C}{\varepsilon}$ and is nonzero only on a set with measure $C\varepsilon^2$ due to (H2).

Taking $w \in L^\infty(\Omega) \cap H_{\Gamma_\pm}(\Omega, \mathbf{b})$, it then holds that $\phi_\varepsilon w \in \widetilde{H_{\Gamma_\pm}}(\Omega, \mathbf{b})$, that $\phi_\varepsilon w \to w$ in $L^2(\Omega)$ as $\varepsilon \to \infty$, and that

$$\|\mathbf{b} \cdot \nabla(\phi_\varepsilon w) - \mathbf{b} \cdot \nabla w\|_{L^2(\Omega)} \leq \underbrace{\|(\phi_\varepsilon - 1)\mathbf{b} \cdot \nabla w\|_{L^2(\Omega)}}_{\to 0} + \underbrace{\|w\mathbf{b} \cdot \nabla \phi_\varepsilon\|_{L^2(\Omega)}}_{\leq C\|w\|_{L^\infty(\Omega)}},$$

using again (H2). Thus, $\phi_\varepsilon w \to w$ in $H_{\Gamma_\pm}(\Omega, \mathbf{b})$ weakly as $\varepsilon \to \infty$, and the claim follows again by Mazur's lemma. □

Finally, the approximation by smooth functions is shown:

Proposition 2.3.5. *$C^\infty(\bar{\Omega}) \cap H_{\Gamma_\pm}(\Omega, \mathbf{b})$ is dense in $H_{\Gamma_\pm}(\Omega, \mathbf{b})$.*

Proof. (Sketch, see [71, Lemma I.10, Corollary I.11] and [12, Lemma 2.2] for details). Let $w \in \widetilde{H_{\Gamma_-}}(\Omega, \mathbf{b})$ which vanishes in an ε-neighborhood of $\partial(\Gamma_-)$. By a suitable partition of unity, we separate w into a sum of functions with support on $\partial\Omega$ strictly contained in $\Gamma_+ \cup \Gamma_0$, and in Γ_-, respectively. By separate mollifications of these functions, we can then define a mollification $w_\eta \in C^\infty(\bar{\Omega})$ of w such that w_η vanishes in an $\frac{\varepsilon}{2}$-neighborhood of $\partial(\Gamma_-)$ (since w vanishes in an ε-neighborhood of $\partial(\Gamma_-)$), and such that w_η vanishes on $\Gamma_+ \cup \Gamma_0$ (since w vanishes on $\Gamma_+ \cup \Gamma_0$, one can define an approximation also satisfying the boundary condition, see [12] and [66, Sect. 5.5, Thm. 2]). Therefore, $w_\eta \in \widetilde{H_{\Gamma_-}}(\Omega, \mathbf{b})$ and $w_\eta \to w$ as $\eta \to 0$, and the claim follows for $H_{\Gamma_-}(\Omega, \mathbf{b})$. The proof for $H_{\Gamma_+}(\Omega, \mathbf{b})$ follows analogously. □

Remark 2.3.6. In chapter 3, we will define the spaces used for our variational formulation slightly differently than the "standard" transport spaces defined in this section. However, it turns out that under suitable conditions on the data functions, the test space $\mathcal{Y}_\mathrm{t}$ is the same space as $H_{\Gamma_\pm}(\Omega, \mathbf{b})$ with an equivalent norm, see Remark 3.1.8.

Finally, using the density result from Proposition 2.3.5, one can show that, while functions in $H(\Omega, \mathbf{b})$ generally only have local traces in L^2, functions in $H_{\Gamma_\pm}(\Omega, \mathbf{b})$ *do* have global L^2-traces:

Proposition 2.3.7. *Under conditions (H1) and (H2), the trace mappings $u \mapsto u|_{\Gamma_\pm}$ are linear and continuous from $H_{\Gamma_\pm}(\Omega, \mathbf{b})$ to $L^2(\partial\Omega, |\mathbf{b} \cdot \mathbf{n}|)$, and the integration by parts or Green's formula*

$$\int_\Omega (\mathbf{b} \cdot \nabla w) w \,\mathrm{d}x = \frac{1}{2} \int_{\partial\Omega} w^2 (\mathbf{b} \cdot \mathbf{n}) \,\mathrm{d}s$$

holds for all $w \in H_{\Gamma_\pm}(\Omega, \mathbf{b})$.

Proof. By approximation with a function $w \in C^\infty(\bar{\Omega}) \cap H_{\Gamma_\pm}(\Omega, \mathbf{b})$, one can use integration by parts and exploit that $(\mathbf{b} \cdot \mathbf{n})$ does not change sign on $\operatorname{supp}(w|_{\partial\Omega})$. For details see Proposition 3.1.6, where the same statement is shown for a slightly different transport operator and norm. □

2.3.2 Spaces for kinetic equations without velocity derivatives

Kinetic equations are usually defined on a phase space domain consisting of space, possibly time, and velocity domains. The basis of a kinetic equation is the kinetic transport operator $v \cdot \nabla_x$ with a velocity-dependent directional derivative in space and possibly a time derivative ∂_t. For the standard kinetic equations of neutron transport and radiative transfer, no velocity derivatives appear, since the velocity dynamics are described by an integral operator.

Let $\Omega = \Omega_{t,x} \times \Omega_v$, where $\Omega_{t,x} \subset \mathbb{R}^{d+1}$ is the space-time domain consisting of the time interval $I_t = (0,T)$ and the spatial domain $\Omega_x \subset \mathbb{R}^d$, and $\Omega_v \subset \mathbb{R}^d$ (often, we have $\Omega_v = S^{d-1} \subset \mathbb{R}^d$ for particles with constant speed) is the velocity domain. We assume that Ω_x and thus $\Omega_{t,x}$ has a piecewise C^1 boundary.

An extensive overview of suitable spaces for kinetic transport equations in a general L^p-framework is given in [47, chapt. XXI], using results from Cessenat [35, 36]. Similar results to [35, 36] were also obtained by Germogenova [69]. We define the space

$$W^p_{\mathrm{nt}}(\Omega) = \{w \in L^p(\Omega_{t,x} \times \Omega_v) : \left(\begin{smallmatrix} 1 \\ v \end{smallmatrix}\right) \cdot \nabla_{t,x} w \in L^p(\Omega_{t,x} \times \Omega_v)\}, \tag{2.16}$$

where $\left(\begin{smallmatrix} 1 \\ v \end{smallmatrix}\right) \cdot \nabla_{t,x} w = \partial_t w + v \cdot \nabla_x w$ is the space-time kinetic transport operator. We here describe the time-dependent case with integrated space-time kinetic operator, all results are equally valid for the time-independent space on $\Omega_x \times \Omega_v$ and with the kinetic operator $v \cdot \nabla_x$.

The in- and outflow domains are defined analogously to subsection 2.3.1 [3]:

$$\Gamma_\pm := \{((t,x),v) \in \partial\Omega_{t,x} \times \Omega_v \, : \, \left(\begin{smallmatrix} 1 \\ v \end{smallmatrix}\right) \cdot \mathbf{n}(t,x) \gtrless 0\} \subset \partial\Omega,$$

where $\partial\Omega_{t,x}$ denotes the space-time boundary consisting of the spatial boundary $I_t \times \partial\Omega_x$ and the initial and final time $\{0\} \times \Omega_x$ and $\{1\} \times \Omega_x$.

For an L^2-based setting, the appropriate function spaces can be described as special cases of the spaces introduced in the last section. In fact, by setting $\mathbf{b}_{\mathrm{kin}}(t,x,v) = (1,v,0)$, we obtain

$$H_{\mathrm{nt}}(\Omega) := W^2_{\mathrm{nt}}(\Omega) = H(\Omega, \mathbf{b}_{\mathrm{kin}}) = \left\{ w \in L^2(\Omega) : \left(\begin{smallmatrix} 1 \\ v \\ 0 \end{smallmatrix}\right) \cdot \nabla_{t,x,v} w \in L^2(\Omega) \right\}.$$

Therefore, all results from subsection 2.3.1 hold for $H_{\mathrm{nt}}(\Omega) := W^2_{\mathrm{nt}}(\Omega)$.

The results, however, largely extend to the general L^p-case. One can show that $C^\infty(\bar{\Omega}_{t,x} \times \Omega_v)$ is dense in $W^p_{\mathrm{nt}}(\Omega)$ (cf. [47, p. 221]), and we have a local trace result:

Proposition 2.3.8. *The functions of $W^p_{\mathrm{nt}}(\Omega)$ have a trace in $L^p_{\mathrm{loc}}(\Gamma_\pm, |\left(\begin{smallmatrix} 1 \\ v \end{smallmatrix}\right) \cdot \mathbf{n}|)$ and in $L^p_{\mathrm{loc}}(\Gamma_\pm)$.*

Proof. See [47, Chap. XXI, sect. 2.2, Theorem 1 and Corollary 1]. □

Again, generally functions in $W^p_{\mathrm{nt}}(\Omega)$ do not have global traces in $L^p(\partial\Omega, |\left(\begin{smallmatrix} 1 \\ v \end{smallmatrix}\right) \cdot \mathbf{n}|)$. An L^2-counterexample specifically for the time-independent kinetic space is given in [102]:

Example 2.3.9. (Taken from [102, pp. 562-563].) Let $\Omega_x = B^1(0) \subset \mathbb{R}^3$ be the unit ball in $\mathbb{R}^3$ and $\Omega_v = S^2$ be the unit sphere. Let $v_0 = (0,0,-1)^T$, so that $\Gamma_-(v_0)$ (the spatial outflow boundary for fixed $v = v_0$) is the upper hemisphere. Using cylindrical coordinates for Ω_x, define $r = \sqrt{x_1^2 + x_2^2}$ and let $w_q(x, v_0) = (1-r)^{-q}$, which is defined

[3] Here we abuse the notation by ignoring the points in which $\mathbf{n}$ is not defined. We use the same handling as in subsection 2.3.1 and note that (H1) is fulfilled since $\partial\Omega_{t,x}$ is assumed to be piecewise C^1.

such that $v_0 \cdot \nabla_x w = 0$. For each $v \in S^2$, we define $w_q(\cdot, v)$ as the respective rotation of $w_q(\cdot, v_0)$. Then[4],

$$\|w_q\|^2_{L^2(\Omega_x \times \Omega_v)} \leq \frac{4\sqrt{2}\pi}{\frac{3}{2} - 2q},$$

which is finite only if $q < \frac{3}{4}$. On the other hand,

$$\int_{\Gamma_-(v_0)} w_q^2 |v_0 \cdot \mathbf{n}| \, \mathrm{d}s \leq \frac{2\sqrt{2}\pi}{1 - 2q},$$

which is finite only if $q < \frac{1}{2}$. Therefore, for $\frac{1}{2} < q < \frac{3}{4}$ we have $w_q \in H_{\mathrm{nt}}(\Omega)$, but $w_q \notin L^2(\partial\Omega, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|)$.

Functions in $W^p_{\mathrm{nt}}(\Omega)$ which do have a global trace in $L^p(\partial\Omega, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|)$ can be characterized in the following way:

Proposition 2.3.10. *Let*

$$\widetilde{W^p_{\mathrm{nt}}}(\Omega) := \{w \in W^p_{\mathrm{nt}}(\Omega) : w|_{\Gamma_-} \in L^p(\Gamma_-, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|) \text{ and } w|_{\Gamma_+} \in L^p(\Gamma_+, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|)\}.$$

Then,

$$\begin{aligned}\widetilde{W^p_{\mathrm{nt}}}(\Omega) &= \{w \in W^p_{\mathrm{nt}}(\Omega) : w|_{\Gamma_-} \in L^p(\Gamma_-, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|)\} \\ &= \{w \in W^p_{\mathrm{nt}}(\Omega) : w|_{\Gamma_+} \in L^p(\Gamma_+, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|)\}.\end{aligned}$$

Proof. See [36]. □

From this we immediately see that the spaces with zero in- or outflow trace

$$W^p_{\mathrm{nt},\Gamma_\pm}(\Omega) = \{w \in W^p_{\mathrm{nt}}(\Omega) : w = 0 \text{ on } \Gamma_\pm\}$$

have finite global traces, analogously to the result shown for $H_{\Gamma_\pm}(\Omega, \mathbf{b})$ in subsection 2.3.1.

Remark 2.3.11. In chapter 4 we will build a variational formulation for the kinetic Fokker-Planck equation, which has a diffusion-term in the velocity variable instead of an integral operator. To that end, we will introduce the function space $H^1_{\mathrm{fp}}(\Omega)$, which is related to $H_{\mathrm{nt}}(\Omega) = W^2_{\mathrm{nt}}(\Omega)$, but differs in the regularity in the velocity dimension. While many concepts work similarly for both spaces, the question of existence of global traces is more problematic for $H^1_{\mathrm{fp}}(\Omega)$, see Appendix A.

2.3.3 Sobolev spaces on manifolds and the Laplace-Beltrami operator

Since the velocity domain for the Fokker-Planck equation covered in chapter 4 is the unit sphere $\Omega_v = S^{d-1}$ for the spatial dimension d, we here recall some results for Sobolev spaces and the Laplace-Beltrami operator on surfaces. We follow the concise overview of the definitions of surfaces, tangential gradients, and the Laplace-Beltrami operator given in [56]. For an extensive overview on the theory of Sobolev spaces on general manifolds we refer to [81].

Surfaces can be described either as *parametrized surfaces*, cf. [56, Sect. 2.1] or as *hypersurfaces*, cf. [56, Sect. 2.2]. Both definitions may be useful dependent on the application. Here, we recall the definition of hypersurfaces:

[4] see [102] for the detailed computation

Definition 2.3.12 ([56, Def. 2.1])**.** Let $k \in \mathbb{N} \cup \{\infty\}$. $\Gamma \subset \mathbb{R}^{d+1}$ is called a C^k-*hypersurface* if, for each point $x_0 \in \Gamma$, there exists an open set $U \subset \mathbb{R}^{d+1}$ containing x_0 and a function $F \in C^k(U)$ with the property that $\nabla F \neq 0$ on $\Gamma \cap U$ and such that

$$U \cap \Gamma = \{x \in U \,|\, F(x) = 0\}. \tag{2.17}$$

The linear space

$$\begin{aligned} T_x\Gamma = \{\tau \in \mathbb{R}^{d+1} \,|\, \exists \gamma : (-\varepsilon, \varepsilon) \to \mathbb{R}^{d+1} & \text{ differentiable}, \\ \gamma((-\varepsilon, \varepsilon)) \subset \Gamma, \gamma(0) = x & \text{ and } \gamma'(0) = \tau\} \end{aligned} \tag{2.18}$$

is called the *tangent space* to Γ at $x \in \Gamma$.

Since for all γ as in (2.18) and the function F from (2.17) it holds

$$0 = \frac{\mathrm{d}}{\mathrm{d}t} F(\gamma(t)) = \langle \nabla F(\gamma(t)), \gamma'(t) \rangle = \langle \nabla F(\gamma(t)), \tau \rangle,$$

we have that $T_x\Gamma \perp \nabla F(x)$. Since $\nabla F \neq 0$, the implicit function theorem yields that $T_x\Gamma$ is an n-dimensional subspace of $\mathbb{R}^{n+1}$. Therefore, we have $T_x\Gamma = \nabla F(x)^\perp$.

A vector $\mathbf{n}(x) \in \mathbb{R}^{n+1}$ is called a *unit normal vector* at $x \in \Gamma$ if $\mathbf{n}(x) \perp T_x\Gamma$ and $|n(x)| = 1$. We thus have

$$\mathbf{n}(x) = \frac{\nabla F(x)}{|\nabla F(x)|} \quad \text{or} \quad -\frac{\nabla F(x)}{|\nabla F(x)|}.$$

Definition 2.3.13 (Tangential gradient, [56, Def. 2.3])**.** Let $\Gamma \subset \mathbb{R}^{d+1}$ be a C^1-hypersurface and let $f : \Gamma \to \mathbb{R}$ be differentiable at $x \in \Gamma$. We define the tangential gradient of f at $x \in \Gamma$ by

$$\nabla_\Gamma f(x) = \nabla \bar{f}(x) - \nabla \bar{f}(x) \cdot \mathbf{n}(x)\mathbf{n}(x) = P(x)\nabla \bar{f}(x),$$

where $P(x)_{ij} = \delta_{ij} - n_i(x)n_j(x)$, $i, j = 1, \ldots, d+1$. Here $\bar{f}$ is a smooth extension of $f : \Gamma \to \mathbb{R}$ to a $(d+1)$-dimensional neighborhood U of the surface Γ, so that $\bar{f}|_\Gamma = f$. ∇ denotes the gradient in $\mathbb{R}^{d+1}$ and $\mathbf{n}(x)$ is a unit normal at x.

The *Laplace-Beltrami operator* applied to a twice differentiable function $f \in C^2(\Gamma)$ is given by

$$\Delta_\Gamma f = \nabla_\Gamma \cdot \nabla_\Gamma f = \sum_{i=1}^{d+1} \underline{D}_i \underline{D}_i f,$$

with the notation $\nabla_\Gamma f(x) = (\underline{D}_1 f(x), \ldots, \underline{D}_{n+1} f(x))$.

Example 2.3.14. We are mainly interested in $\Gamma = \Omega_v = S^2 \subset \mathbb{R}^3$. When using spherical coordinates $(r, \theta, \phi) \in [0, \infty) \times [0, \pi] \times [0, 2\pi)$ with

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} r \sin\theta \cos\phi \\ r \sin\theta \sin\phi \\ r \cos\theta \end{pmatrix},$$

we expect the tangential gradient of the sphere to be the angular part of the full gradient.

This can indeed be computed in detail using the above definitions: In spherical coordinates, the full gradient on $\mathbb{R}^3$ is

$$\nabla f(r,\theta,\phi) = \mathbf{e}_r \frac{\partial f}{\partial r} + \mathbf{e}_\theta \frac{1}{r}\frac{\partial f}{\partial \theta} + \mathbf{e}_\phi \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi},$$

with the orthonormal basis vectors

$$\mathbf{e}_r = \begin{pmatrix} \sin\theta\cos\phi \\ \sin\theta\sin\phi \\ \cos\theta \end{pmatrix}, \quad \mathbf{e}_\theta = \begin{pmatrix} \cos\theta\cos\phi \\ \cos\theta\sin\phi \\ -\sin\theta \end{pmatrix}, \quad \mathbf{e}_\phi = \begin{pmatrix} -\sin\phi \\ \cos\phi \\ 0 \end{pmatrix}.$$

The surface $\Omega_v = S^2$ is simply given as

$$\{(r,\theta,\phi) \in [0,\infty) \times [0,\pi] \times [0,2\pi) : F(r,\theta,\phi) := r - 1 = 0\}.$$

Hence, we have $\nabla F = \mathbf{e}_r$ and thus $\mathbf{n} = \mathbf{e}_r$. Then, the tangential gradient is given as

$$\nabla_{\Omega_v} f(r,\theta,\phi) = \nabla f(r,\theta,\phi) - (\nabla f(r,\theta,\phi)\cdot \mathbf{e}_r)\mathbf{e}_r = \mathbf{e}_\theta \frac{1}{r}\frac{\partial f}{\partial \theta} + \mathbf{e}_\phi \frac{1}{r\sin\theta}\frac{\partial f}{\partial \phi},$$

as $\mathbf{e}_r, \mathbf{e}_\theta$, and $\mathbf{e}_\phi$ are orthonormal. For the Laplace-Beltrami operator, we obtain by using the definitions of $\mathbf{e}_\theta$ and $\mathbf{e}_\phi$

$$\Delta_{\Omega_v} f(r,\theta,\phi) = \nabla_{\Omega_v} f(r,\theta,\phi)\cdot\nabla_{\Omega_v} f(r,\theta,\phi) = \frac{1}{r^2}\frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{r^2\sin\theta}\frac{\partial f}{\partial\theta} + \frac{1}{r^2\sin^2\theta}\frac{\partial^2 f}{\partial\phi^2}.$$

From the theory of tangential gradients it is also possible to derive an integration by parts formula for surfaces, which can be proven by an appropriate extension of the function to a neighborhood of Γ (see [56] for details on this extension):

Theorem 2.3.15 ([56, Thm. 2.10])**.** *Assume that Γ is a hypersurface in $\mathbb{R}^{d+1}$ with smooth boundary $\partial\Gamma$ and that $f \in C^1(\bar{\Gamma})$. Then*

$$\int_\Gamma \nabla_\Gamma f \,\mathrm{d}s = \int_\Gamma f H \mathbf{n}\,\mathrm{d}s + \int_{\partial\Gamma} f \boldsymbol{\mu}\,\mathrm{d}s$$

Here, $H(x)$ denotes the mean curvature of Γ at the point x and $\boldsymbol{\mu}$ denotes the co-normal vector which is normal to $\partial\Gamma$ and tangent to Γ. $\mathrm{d}s$ denotes either the d-dimensional surface measure on Γ or the $(d-1)$-dimensional surface measure on $\partial\Gamma$. A compact hypersurface Γ does not have a boundary, $\partial\Gamma = \emptyset$, and the last term on the right-hand side vanishes.

For $\Omega_v = S^2$ the mean curvature is constant, with the definition given in [56] we have $H(x) = 2$. Furthermore, we have $\partial S^2 = \emptyset$.

With this, we can define weak derivatives and Sobolev spaces. $L^p(\Gamma)$ is simply defined through the surface measure $\mathrm{d}s$, i.e., $\|f\|_{L^p(\Gamma)} = (\int_\Gamma |f|^p \,\mathrm{d}s)^{1/p}$.

Definition 2.3.16 ([56, Def. 2.11])**.** A function $f \in L^1(\Gamma)$ has the weak derivative $v_i = \underline{D}_i f \in L^1(\Gamma)$ for $i \in \{1,\dots,d+1\}$ if, for every function $\phi \in C^1(\Gamma)$ with compact support $\overline{\{x \in \Gamma \,|\, \phi(x) \neq 0\}} \subset \Gamma$, we have the relation

$$\int_\Gamma f \underline{D}_i \phi \,\mathrm{d}s = -\int_\Gamma \phi v_i \,\mathrm{d}s + \int_\Gamma f\phi H n_i \,\mathrm{d}s.$$

The Sobolev space $H^{1,p}(\Gamma)$ is defined by

$$H^{1,p}(\Gamma) = \{f \in L^p(\Gamma) \,|\, \underline{D}_i f \in L^p(\Gamma), i = 1, \dots, d+1\}$$

with norm

$$\|f\|_{H^{1,p}(\Gamma)} = \left(\|f\|^p_{L^p(\Gamma)} + \|\nabla_\Gamma f\|^p_{L^p(\Gamma)}\right)^{\frac{1}{p}}.$$

Define $H^{k,p}$ analogously as usual and denote $H^k(\Gamma) = H^{k,2}(\Gamma)$.

We also have the Poincaré's inequality and Green's formula:

Theorem 2.3.17 (Poincaré's inequality on surfaces, [56, Thm. 2.12])**.** *Assume that $\Gamma \in C^3$ and $1 \leq p < \infty$. Then there is a constant $c > 0$ such that, for every function $f \in H^{1,p}(\Gamma)$ with $\int_\Gamma f dA = 0$, we have the inequality*

$$\|f\|_{L^p(\Gamma)} \leq c\|\nabla_\Gamma f\|_{L^p(\Gamma)}. \tag{2.19}$$

Theorem 2.3.18 ([56, Thm. 2.14])**.** *Assume that Γ is a hypersurface in $\mathbb{R}^{d+1}$ with smooth boundary $\partial\Gamma$ and that $f, g \in H^{1,p}(\Gamma)$. Then*

$$\int_\Gamma \nabla_\Gamma f \cdot \nabla_\Gamma g \, \mathrm{d}s = -\int_\Gamma f \Delta_\Gamma g \, \mathrm{d}s + \int_{\partial\Gamma} f \nabla_\Gamma g \cdot \boldsymbol{\mu} \, \mathrm{d}s.$$

In [81], many other properties for Sobolev spaces on general Riemannian manifolds, e.g., a chain rule formula and different Sobolev embedding theorems are proven. We close this section with the following density result that we will need in chapter 4.

Theorem 2.3.19 ([81, Thm. 2.4, p. 25])**.** *Given (M, g) a smooth, complete Riemannian manifold, the set $C_0^\infty(M)$ of smooth functions with compact support in M is dense in $H^{1,p}(M)$ for any $p \geq 1$.*

3 The transport equation

In this chapter, we are concerned with the numerical solution of the (parametrized) time-dependent linear first-order transport equation

$$\partial_t u_\mu(t,x) + \mathbf{b}_\mu(t,x) \cdot \nabla_x u_\mu(t,x) + c_\mu(t,x)\, u_\mu(t,x) = f_\mu(t,x), \tag{3.1}$$

for all parameter values μ in a compact set $\mathcal{P} \subset \mathbb{R}^p$, for all times $t \in (0,T)$ ($T > 0$ being some final time) and all $x \in D \subset \mathbb{R}^d$ accompanied with appropriate initial and boundary conditions.

We first introduce a suitable variational formulation for (3.1) in section 3.1: Orienting on similar settings in [29,43,51,52], we use an ultraweak formulation requiring only L^2-regularity of the weak solution. We choose the corresponding test space such that the variational problem is well-posed and optimally conditioned, meaning that the inf-sup and continuity constants are both unity.

In section 3.2, we then propose an optimally stable Petrov-Galerkin discretization of the variational problem. To that end, we first choose an appropriate discrete test space $\mathcal{Y}_\mathrm{t}^\delta$ and subsequently compute the corresponding trial space $\mathcal{X}_\mathrm{t}^\delta$. Doing so, the optimal trial space $\mathcal{X}_\mathrm{t}^\delta$ arises from the application of the differential operator on the basis functions of $\mathcal{Y}_\mathrm{t}^\delta$. This is different from the related approaches, where stable test spaces are computed from chosen trial space functions by approximately solving PDEs locally [29,51,52] or in a saddle point formulation [43]. With our strategy, we obtain efficiently computable discrete spaces with a discrete inf-sup constant equal to one independently of the mesh size. We include a comparison of our scheme with other methods in subsection 3.2.4.

The optimally stable discretization scheme is applied to parametrized transport problems in the context of the RB method in section 3.3. The ultraweak optimally stable variational formulation here leads to parameter-dependent test spaces $\mathcal{Y}_{\mathrm{t},\mu}$. Using the discretization from section 3.2, we obtain a Petrov-Galerkin discretization based on discrete test spaces with fixed basis functions and parameter-dependent norm and discrete trial spaces with parameter-dependent basis functions, but fixed L^2-norm. Thereby, the scheme automatically leads to discrete inf-sup constants of one simultaneously for all parameter values. We then apply the RB method: We introduce a reduced scheme with possible offline-/online decomposition and propose a suitable greedy algorithm for the basis generation. Here again, the optimally stable setting ensures that no additional stabilization is necessary for the reduced spaces, different from related approaches as the *double greedy* algorithm in [45]. However, due to the parameter-dependent test space norms the standard RB error estimators cannot be computed efficiently. We therefore introduce a hierarchical error estimator as an alternative.

In section 3.4, we describe the fairly easy computational realization of the new approach. In the solution process, we first solve a coercive least-squares problem in the test space. Afterwards, the solution of the Petrov-Galerkin problem can be computed by applying the transport operator to the least-squares test space solution.

Finally, we report on several numerical experiments in section 3.5. In subsection 3.5.1, we examine the approach for the non-parametric case. We assess the convergence rates for problems with different smoothness, investigate limitations of the chosen spaces and possible post-processing strategies, and compare our approach with the discretization from [43], where problem-dependent stable test spaces (instead of trial spaces) are computed. In subsection 3.5.2, we describe the experiments for the RB method in the parametric case. We assess the convergence rates of the greedy algorithm for different test cases, compare our approach to the *double greedy* algorithm introduced in [45], and test the proposed hierarchical error estimator.

The results presented in this chapter have been published in [28].

3.1 An optimally stable ultraweak (space-time) formulation

In this section we present an ideally conditioned variational framework for linear first order transport equations using results from [43] and [6, 7]. To that end, let $\Omega \subset \mathbb{R}^n$, $n \geq 1$, be a bounded polyhedral domain with Lipschitz boundary, where we note that Ω may also be a space-time domain, as will be shown in Example 3.1.9 at the end of this section. Moreover, $\mathbf{n}$ shall denote the outward normal of $\Gamma := \partial\Omega$. Next, we introduce the advection field $\mathbf{b}(\cdot) \in C^1(\bar{\Omega})^n$ and the reaction coefficient $c(\cdot) \in C^0(\bar{\Omega})$, noting that for some statements the regularity assumption on $\mathbf{b}(\cdot)$ may be relaxed. We assume throughout this chapter that

$$c(z) - \tfrac{1}{2}\nabla \cdot \mathbf{b}(z) \geq 0 \quad \text{for } z \in \Omega \ \text{ almost everywhere.}$$

Then, we consider the first order transport equation

$$\begin{aligned} B_{\mathrm{t},\circ} u(z) := \mathbf{b}(z) \cdot \nabla u(z) + c(z) u(z) &= f_\circ(z), & z &\in \Omega, \\ u(z) &= g(z), & z &\in \Gamma_- \equiv \Gamma_{\text{inflow}}, \end{aligned} \tag{3.2}$$

where $f_\circ \in C^0(\bar{\Omega})$, $g \in C^0(\overline{\Gamma_-})$, and $\Gamma_\pm := \{z \in \partial\Omega : \ \mathbf{b}(z) \cdot \mathbf{n}(z) \gtrless 0\}$.

For functions $v, w \in C^0(\bar{\Omega}) \cap C^1(\Omega)$ we obtain

$$(B_{\mathrm{t},\circ} v, w)_{L^2(\Omega)} = (v, B^*_{\mathrm{t},\circ} w)_{L^2(\Omega)} + \int_{\Gamma_-} vw(\mathbf{b}\cdot\mathbf{n}) \,\mathrm{d}s + \int_{\Gamma_+} vw(\mathbf{b}\cdot\mathbf{n}) \,\mathrm{d}s,$$

where $B^*_{\mathrm{t},\circ} w = -\mathbf{b}\cdot\nabla w + w(c - \nabla\cdot\mathbf{b})$ denotes the formal adjoint of $B_{\mathrm{t},\circ}$.[1] To account for the nonhomogeneous boundary conditions, we introduce as in [43] the spaces $C^1_{\Gamma_\pm}(\Omega) := \{v \in C^0(\bar{\Omega}) \cap C^1(\Omega) \ : \ v|_{\Gamma_\pm} = 0\}$ and obtain

$$(B_{\mathrm{t},\circ} v, w)_{L^2(\Omega)} = (v, B^*_{\mathrm{t},\circ} w)_{L^2(\Omega)}, \quad v \in C^1_{\Gamma_-}(\Omega), w \in C^1_{\Gamma_+}(\Omega).$$

Thus, we may define the domain of $B^*_{\mathrm{t},\circ}$ as $\mathrm{dom}(B^*_{\mathrm{t},\circ}) = C^1_{\Gamma_+}(\Omega)$. For the derivation of a stable variational formulation we require as in [43] the following two assumptions.

Assumption 3.1.1. We assume that the following conditions hold:
(B1) There exists a dense subspace $\mathrm{dom}(B^*_{\mathrm{t},\circ}) \subseteq L^2(\Omega)$ on which $B^*_{\mathrm{t},\circ}$ is injective.

[1] Considering (3.2) with $g(z) \equiv 0$ and thus homogeneous Dirichlet boundary conditions we define the formal adjoint $B^*_{\mathrm{t},\circ}$ of $B_{\mathrm{t},\circ}$ by $(B_{\mathrm{t},\circ} v, w)_{L^2(\Omega)} = (v, B^*_{\mathrm{t},\circ} w)_{L^2(\Omega)}$ for all $v, w \in C^\infty_0(\Omega)$.

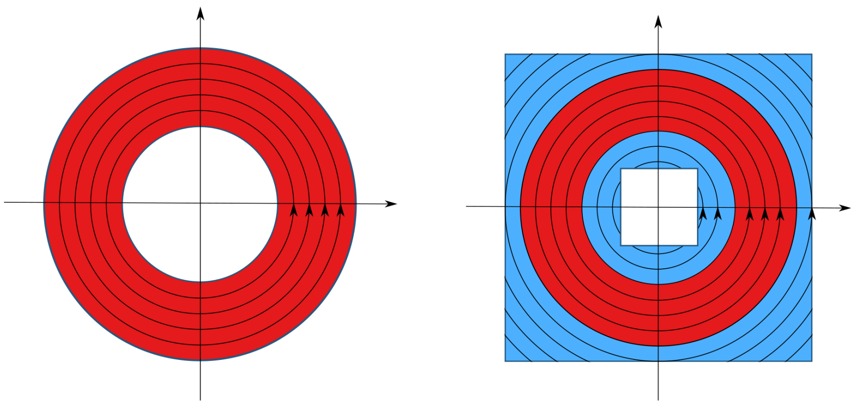

Figure 3.1: Domains Ω_1 (left) and Ω_2 with characteristic curves and $\operatorname{supp} u$ in red.

(B2) The *range* $\operatorname{ran}(B^*_{\mathrm{t},\circ}) := \{B^*_{\mathrm{t},\circ}v : v \in \operatorname{dom}(B^*_{\mathrm{t},\circ})\}$ of $B^*_{\mathrm{t},\circ}$ is densely embedded in $L^2(\Omega)$.

These essential assumptions on the well-posedness of the problem are however not fulfilled for arbitrary coefficient functions and domains. One commonly used condition where well-posedness can be shown rather easily (see e.g. [65, Prop. 5.9], [43, Remark 2.2(ii)], [72]) is that $c - \frac{1}{2}\nabla \cdot \mathbf{b} \geq \kappa$ in Ω for some $\kappa > 0$.

For a vanishing reaction term, however, we can construct basic counterexamples where Assumption 3.1.1 is not fulfilled: We consider $\Omega \subset \mathbb{R}^2$ with advection field $\mathbf{b}(x, y) = (-y, x)$ and no reaction $c \equiv 0$. It holds $\nabla \cdot \mathbf{b} = 0$, thus, the adjoint operator is simply $B^*_{\mathrm{t},\circ}v = -\mathbf{b} \cdot \nabla v$. For the annular domain

$$\Omega_1 = \{(x, y) \in \mathbb{R}^2 : 0.25 < x^2 + y^2 < 1\}$$

(see Figure 3.1, left), it holds $\mathbf{b} \neq 0$ on $\overline{\Omega}_1$. The boundary has the form of two circles: $\Gamma = \partial\Omega_1 = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 0.25\} \cup \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$, outward normal is $\mathbf{n} = (x, y)$ on $\{x^2 + y^2 = 1\}$ and $\mathbf{n} = -2(x, y)$ on $\{x^2 + y^2 = 0.25\}$. Since $\mathbf{b} \cdot \mathbf{n} = (-y, x) \cdot C(x, y) = C(-xy + xy) = 0$ for a constant $C \in \mathbb{R}$, $C \neq 0$, the whole boundary belongs to Γ_0. Therefore, $v \equiv 1 \in C^1_{\Gamma_+}(\Omega_1)$ but $\mathbf{b} \cdot \nabla v = 0$, i.e. $B^*_{\mathrm{t},\circ}$ is not injective on $C^1_{\Gamma_+}(\Omega_1)$.

Even with a polyhedral domain with polygonal boundary and $\Gamma_0 \subsetneq \partial\Omega$ the problem may not be well-posed: Consider to that end

$$\Omega_2 = (-1, 1)^2 \setminus (-0.25, 0.25)^2$$

(see Figure 3.1, right) with $\mathbf{b}$ and c as before. We thus have again $\mathbf{b} \neq 0$ on $\overline{\Omega}_2$. Then, let $0 \neq \psi \in C^1([0, 1])$ with $\operatorname{supp} \psi \subset [0.5, 0.9]$ and consider

$$u(x, y) = \psi(\sqrt{x^2 + y^2}).$$

With this definition, $u|_{\partial\Omega_2} = 0$, i.e., $u \in C^1_{\Gamma_+}(\Omega_2)$. The characteristic curves of $\mathbf{b}$ are circle-shaped of the form $\gamma(t) = r(\cos(\phi + t), \sin(\phi + t))$ for a starting point $(x, y) =$

$r(\cos\phi, \sin\phi)$. The rotational invariant function u is thus constant on the characteristic curves, therefore it holds $\mathbf{b}\cdot\nabla u = 0$, and $B^*_{\mathrm{t},\circ}$ is again not injective[2].

These examples show that closed characteristic curves of the advection field that do not reach the boundary may lead to ill-posed problems. To give sufficient conditions for the well-posedness of pure advection equations, we reuse the notion of Ω-filling flows developed in [6, 7].

Definition 3.1.2 (Ω-filling flow, [6])**.** Let $\mathbf{b} \in C^1(\bar{\Omega})^n$, and let the flow associated with $\mathbf{b}$ be described by the integral curves $\xi : (s, x) \in [\sigma_x, \tau_x] \times \bar{\Omega} \to \xi(s, x) \in \bar{\Omega}$ that solve

$$\frac{d\xi}{\mathrm{d}s} = \mathbf{b}(\xi), \quad \xi(0, x) = x.$$

Then the flow associated with $\mathbf{b}$ is called Ω-filling, if there exists $T > 0$ such that for almost every $x \in \bar{\Omega}$ there exist $x_0 \in \Gamma_-$ and $0 \le t \le T$ such that

$$x = \xi(t, x_0).$$

In other words, the trajectories of the flow associated with the vector field $\mathbf{b}$ starting from the inflow boundary do fill $\bar{\Omega}$ except perhaps for a set of measure zero in a finite bounded time T.

Similar to [6, Lem. 7], we show the following lemma.

Lemma 3.1.3. *If the flow associated with* $\mathbf{b} \in C^1(\bar{\Omega})^n$ *is* Ω*-filling, then there exists* $\rho \in L^\infty(\Omega)$ *such that*

$$\begin{aligned} \mathbf{b}\cdot\nabla\rho &= 2 \quad \text{in } \Omega, \\ \rho &= 0 \quad \text{on } \Gamma_-. \end{aligned} \tag{3.3}$$

Moreover, we have $\|\rho\|_{L^\infty(\Omega)} \le 2T$ *and* $\rho \ge 0$ *almost everywhere in* Ω*.*

Proof. The function ρ can be found by the method of characteristics: Since the flow associated with $\mathbf{b}$ is Ω-filling, for almost every $x \in \Omega$, there exist $x_0 \in \Gamma_-$ and $0 \le t \le T$ with $x = \xi(t, x_0)$. Define $\rho(x) = 2t$. Since $0 \le t \le T$, we get[3] $\rho \in L^\infty(\Omega)$, $\|\rho\|_{L^\infty(\Omega)} \le 2T$, and $\rho \ge 0$ almost everywhere in Ω. By definition, for $x_0 \in \Gamma_-$ we have $\xi(0, x_0) = x_0$, i.e. $\rho(x_0) = 0$, which means $\rho|_{\Gamma_-} = 0$. Furthermore, it holds for almost every $x \in \Omega$ that

$$\begin{aligned} \mathbf{b}(x)\cdot\nabla\rho(x) = \mathbf{b}(\xi(t, x_0))\cdot\nabla\rho(\xi(t, x_0)) &= \frac{\mathrm{d}}{\mathrm{d}t}\xi(t, x_0)\cdot\nabla\rho(\xi(t, x_0)) \\ &= \frac{\mathrm{d}}{\mathrm{d}t}\rho(\xi(t, x_0)) = \frac{\mathrm{d}}{\mathrm{d}t}2t = 2, \end{aligned}$$

i.e., ρ fulfills (3.3). □

The following proposition gives a sufficient condition for $\mathbf{b}$ to have an Ω-filling flow:

[2] This second case is a counterexample to the claim in [43, Remark 2.2(i)], that the assumption $0 \neq \mathbf{b} \in C^1(\Omega)^n$ is already sufficient for Assumption 3.1.1 on a bounded, polyhedral domain $\Omega \subset \mathbb{R}^n$, $n > 1$ with Lipschitz boundary that consists of finitely many polyhedral faces again having Lipschitz boundaries

[3] ρ is in general not continuous: Consider, e.g., a nonconvex domain Ω where a characteristic curve is tangential to the boundary at some (isolated) $x \in \Gamma_0$, but not in a neighborhood of x. Then ρ is discontinuous along the characteristic curve starting from x.

Proposition 3.1.4 ([6, Prop. 7])**.** *If $\mathbf{b} \in C^1(\bar{\Omega})^n$ is bounded as well as its gradient in a neighborhood V of $\bar{\Omega}$, if there are a unit vector $\mathbf{k}$, a number $\alpha > 0$ such that*

$$\mathbf{b}(x) \cdot \mathbf{k} \geq \alpha \quad \forall x \in \bar{\Omega}, \tag{3.4}$$

and if Ω is bounded in the $\mathbf{k}$ direction then the flow is Ω-filling.

With these preliminaries, we can now show in a modified version of [43, Remark 2.2][4] that Assumption 3.1.1 holds for two different conditions on the data functions and domain:

Proposition 3.1.5. *Let one of the following two conditions hold:*

(i) The flow associated with $\mathbf{b}$ is Ω-filling

(ii) There exists $\kappa > 0$ with $c - \frac{1}{2}\nabla \cdot \mathbf{b} \geq \kappa$ in Ω.

Then, the operator $B^*_{\mathrm{t},\circ}$ satisfies Assumption 3.1.1. Moreover, we have the curved Poincaré inequality

$$\|v\|_{L^2(\Omega)} \leq c_p \|B^*_{\mathrm{t},\circ} v\|_{L^2(\Omega)}, \quad v \in C^1_{\Gamma_+}(\Omega). \tag{3.5}$$

In case of condition (i) the constant can be bounded by $c_p = 2T$; in case (ii) $c_p = \frac{1}{\kappa}$.

Proof. We first show (3.5), i.e., $\|v\|_{L^2(\Omega)} \leq c_p \|B^*_{\mathrm{t},\circ} v\|_{L^2(\Omega)}$. Let thus $v \in C^1_{\Gamma_+}(\Omega)$. If condition (i) holds, we can slightly adapt the proof of [7, Thm. 1]: Let ρ be given as in Lemma 3.1.3. Then,

$$\begin{aligned}
(B^*_{\mathrm{t},\circ} v, \rho v)_{L^2(\Omega)} &= (-\mathbf{b} \cdot \nabla v + v(c - \nabla \cdot \mathbf{b}), \rho v)_{L^2(\Omega)} \\
&= -\int_\Omega \mathbf{b} \cdot \nabla v \rho v \,\mathrm{d}x + \int_\Omega v^2 \rho (c - \nabla \cdot \mathbf{b}) \,\mathrm{d}x \\
&= -\int_\Omega \tfrac{1}{2} \rho \mathbf{b} \cdot \nabla v^2 \,\mathrm{d}x + \int_\Omega v^2 \rho (c - \nabla \cdot \mathbf{b}) \,\mathrm{d}x \\
&= \int_\Omega \tfrac{1}{2} \nabla \cdot (\rho \mathbf{b}) v^2 \,\mathrm{d}x + \int_\Omega v^2 \rho (c - \nabla \cdot \mathbf{b}) \,\mathrm{d}x,
\end{aligned}$$

where we have no boundary integral from the partial integration since the traces of v on Γ_+ and of ρ on Γ_- vanish. Further, we obtain

$$(B^*_{\mathrm{t},\circ} v, \rho v)_{L^2(\Omega)} = \int_\Omega v^2 (\tfrac{1}{2} \underbrace{\mathbf{b} \cdot \nabla \rho}_{=2} + \underbrace{\rho}_{\geq 0} \underbrace{(c - \tfrac{1}{2} \nabla \cdot \mathbf{b})}_{\geq 0}) \,\mathrm{d}x \geq \|v\|^2_{L^2(\Omega)}. \tag{3.6}$$

Using $\|\rho v\|_{L^2(\Omega)} \leq \|\rho\|_{L^\infty(\Omega)} \|v\|_{L^2(\Omega)} \leq 2T \|v\|_{L^2(\Omega)}$ we have

$$\|B^*_{\mathrm{t},\circ} v\|_{L^2(\Omega)} \geq \|\rho v\|^{-1}_{L^2(\Omega)} (B^*_{\mathrm{t},\circ} v, \rho v)_{L^2(\Omega)} \geq \frac{1}{2T} \|v\|_{L^2(\Omega)}. \tag{3.7}$$

[4] As shown in the counterexample above, Remark 2.2(i) in [43] is in general not sufficient for well-posedness. Therefore, we reuse only Remark 2.2(ii) and give a second condition based on Ω-filling flows.

For condition (ii), i.e., $c - \frac{1}{2}\nabla \cdot \mathbf{b} \geq \kappa > 0$, we obtain by integration by parts (see e.g. [130, Lem. 3.1.1])

$$\begin{aligned}(B^*_{\mathrm{t},\circ}v, v)_{L^2(\Omega)} &= \int_\Omega -v\mathbf{b} \cdot \nabla v \,\mathrm{d}x + \int_\Omega v^2(c - \nabla \cdot \mathbf{b}) \,\mathrm{d}x \\ &= -\tfrac{1}{2}\int_\Omega v\mathbf{b} \cdot \nabla v \,\mathrm{d}x + \tfrac{1}{2}\int_\Omega v\mathbf{b} \cdot \nabla v + v^2\nabla \cdot \mathbf{b} \,\mathrm{d}x - \tfrac{1}{2}\int_{\Gamma_-} v^2\mathbf{b} \cdot \mathbf{n} \,\mathrm{d}s \\ &\quad + \int_\Omega v^2(c - \nabla \cdot \mathbf{b}) \,\mathrm{d}x \\ &= \int_\Omega v^2(c - \tfrac{1}{2}\nabla \cdot \mathbf{b}) \,\mathrm{d}x - \tfrac{1}{2}\int_{\Gamma_-} v^2 \underbrace{\mathbf{b} \cdot \mathbf{n}}_{<0} \,\mathrm{d}s \geq \kappa\|v\|^2_{L^2(\Omega)} \qquad (3.8)\end{aligned}$$

and thus

$$\|B^*_{\mathrm{t},\circ}v\|_{L^2(\Omega)} \geq \kappa\|v\|_{L^2(\Omega)},$$

i.e., (3.5) holds for both cases.

Since (3.5) implies injectivity of $B^*_{\mathrm{t},\circ}$ on $C^1_{\Gamma_+}(\Omega)$, which is dense in $L^2(\Omega)$, assumption (B1) is fulfilled.

To prove assumption (B2), we slightly modify the proof of [6, Thm. 16]. To prove density of $\mathrm{ran}(B^*_{\mathrm{t},\circ})$ in $L^2(\Omega)$, we take $w \in L^2(\Omega)$ that is orthogonal to $\mathrm{ran}(B^*_{\mathrm{t},\circ})$ and show $w \equiv 0$. We thus have

$$(B^*_{\mathrm{t},\circ}v, w)_{L^2(\Omega)} = 0 \quad \forall v \in C^1_{\Gamma_+}(\Omega).$$

Let at first $v \in C^1_0(\Omega)$. We then have

$$0 = \int_\Omega -\mathbf{b} \cdot \nabla vw + (c - \nabla \cdot \mathbf{b})vw \,\mathrm{d}x = \int_\Omega -\nabla \cdot (\mathbf{b}v)w + cvw \,\mathrm{d}x \qquad (3.9)$$

By partial integration we see that $\mathbf{b} \cdot \nabla w + cw$ is a distribution of order 1 with

$$\langle \mathbf{b} \cdot \nabla w + cw, v\rangle = 0,$$

which already means $\mathbf{b} \cdot \nabla w + cw = 0$, i.e., $\mathbf{b} \cdot \nabla w = -cw \in L^2(\Omega)$. Therefore, $w \in H(\Omega, \mathbf{b})$ and by Proposition 2.3.1, w has a local trace $w|_{\Gamma_-} \in L^2_{\mathrm{loc}}(\Gamma_-, |\mathbf{b} \cdot \mathbf{n}|)$ and a global trace $w|_{\partial\Omega} \in H^{-1/2}(\partial\Omega)$, admitting a Green's formula/integration by parts with a test function in $H^1(\Omega)$. Let now $v \in C^1_{\Gamma_+}(\Omega)$. We then obtain from partial integration of (3.9), using $\mathbf{b} \cdot \nabla w + cw = 0$ and $v|_{\Gamma_+} = 0$ that

$$\int_{\Gamma_-} vw\mathbf{b} \cdot \mathbf{n} \,\mathrm{d}s = 0.$$

Since v is arbitrary on Γ_- and $\mathbf{b}\cdot\mathbf{n} < 0$ on Γ_- we thus have $w|_{\Gamma_-} = 0$, i.e., $w \in H_{\Gamma_-}(\Omega, \mathbf{b})$.

We now consider the curved Poincaré inequality (3.5) for the (nonadjoint) operator $B_{\mathrm{t},\circ}z = \mathbf{b} \cdot \nabla z + cz$: By setting $\tilde{\mathbf{b}} = -\mathbf{b}$ and $\tilde{c} = c - \nabla \cdot \mathbf{b}$, (3.5) reads

$$\|-\tilde{\mathbf{b}}\cdot\nabla z+(\tilde{c}-\nabla\cdot\tilde{\mathbf{b}})z\|_{L^2(\Omega)} = \|\mathbf{b}\cdot\nabla z+cz\|_{L^2(\Omega)} \geq (\tilde{c_p})^{-1}\|z\|_{L^2(\Omega)} \quad \forall z \in C^1_{\Gamma_-}(\Omega), \quad (3.10)$$

as Γ_- is the outflow boundary for $\tilde{\mathbf{b}} = -\mathbf{b}$. Since $C^1_{\Gamma_-}(\Omega)$ is dense in $H_{\Gamma_-}(\Omega, \mathbf{b})$ by Proposition 2.3.5 we obtain $0 = \|\mathbf{b} \cdot \nabla w + cw\|_{L^2(\Omega)} \geq (\tilde{c_p})^{-1}\|w\|_{L^2(\Omega)}$, and thus $w = 0$. Hence, (B2) is also fulfilled. □

We may now define as in [43]

$$\|v\|_* := \|B_{\mathrm{t},\circ}^* v\|_{L^2(\Omega)}$$

and note that due to (B1) $\|\cdot\|_*$ is a norm on $\mathrm{dom}(B_{\mathrm{t},\circ}^*)$. With this framework at hand, we can define as in [43] the test space by

$$\mathcal{Y}_{\mathrm{t}} := \mathrm{clos}_{\|\cdot\|_*}\{\mathrm{dom}(B_{\mathrm{t},\circ}^*)\},$$

which is a Hilbert space with inner product $(v,w)_{\mathcal{Y}_{\mathrm{t}}} := (B_{\mathrm{t}}^* v, B_{\mathrm{t}}^* w)_{L^2(\Omega)}$ and induced norm $\|v\|_{\mathcal{Y}_{\mathrm{t}}} := \|v\|_*$, $v,w \in \mathcal{Y}_{\mathrm{t}}$. Here, $B_{\mathrm{t}}^* : \mathcal{Y}_{\mathrm{t}} \to L^2(\Omega)$ denotes the continuous extension of $B_{\mathrm{t},\circ}^*$ from $\mathrm{dom}(B_{\mathrm{t},\circ}^*)$ to $\mathcal{Y}_{\mathrm{t}}$. Then, we can define $B_{\mathrm{t}} : L^2(\Omega) \to \mathcal{Y}_{\mathrm{t}}'$ again by duality, i.e., $B_{\mathrm{t}} := (B_{\mathrm{t}}^*)^*$. The variational formulation of (3.2) may then be based upon the bilinear form

$$b_{\mathrm{t}} : L^2(\Omega) \times \mathcal{Y}_{\mathrm{t}} \to \mathbb{R}, \qquad b_{\mathrm{t}}(v,w) := (v, B_{\mathrm{t}}^* w)_{L^2(\Omega)} = \int_\Omega v(-\mathbf{b}\cdot\nabla w + w(c - \nabla\cdot\mathbf{b}))\,\mathrm{d}x. \tag{3.11}$$

To incorporate the boundary conditions, we use the weighted L^2-spaces $L^2(K, |\mathbf{b}\cdot\mathbf{n}|)$ for $K \subset \partial\Omega$ defined in subsection 2.3.1. We then show that functions in $\mathcal{Y}_{\mathrm{t}}$ have a trace in $L^2(\Gamma_-, |\mathbf{b}\cdot\mathbf{n}|)$.[5]

Proposition 3.1.6. *Assume that one of the two conditions of Proposition 3.1.5 holds. Then, there exists a linear continuous mapping*

$$\gamma_- : \mathcal{Y}_{\mathrm{t}} \to L^2(\Gamma_-, |\mathbf{b}\cdot\mathbf{n}|),$$

such that

$$\|\gamma_-(v)\|_{L^2(\Gamma_-,|\mathbf{b}\cdot\mathbf{n}|)} \le C_{tr}\|v\|_{\mathcal{Y}_{\mathrm{t}}}, \quad v \in \mathcal{Y}_{\mathrm{t}}. \tag{3.12}$$

The constant is $C_{tr} = \sqrt{4T}$, or $C_{tr} = \sqrt{2\kappa^{-1}}$, respectively.

Proof. Integration by parts yields for $v \in C^1(\bar\Omega)$ (see also (3.8))

$$(B_{\mathrm{t},\circ}^* v, v)_{L^2(\Omega)} = \int_\Omega v^2(c - \tfrac{1}{2}\nabla\cdot\mathbf{b})\,\mathrm{d}x - \tfrac{1}{2}\int_{\Gamma_-} v^2 \mathbf{b}\cdot\mathbf{n}\,\mathrm{d}s.$$

By using the general assumption $c - \frac{1}{2}\nabla\cdot\mathbf{b} \ge 0$ and $\mathbf{b}\cdot\mathbf{n} < 0$ on Γ_-, we have for $v \in C^1_{\Gamma_+}(\Omega)$

$$\int_{\Gamma_-} v^2|\mathbf{b}\cdot\mathbf{n}|\,\mathrm{d}s \le 2|(B_{\mathrm{t},\circ}^* v, v)| \le 2\|v\|_{L^2(\Omega)}\|v\|_{\mathcal{Y}_{\mathrm{t}}} \le 2C\|v\|^2_{\mathcal{Y}_{\mathrm{t}}}, \tag{3.13}$$

where we have used (3.5) in the last estimate. The assertion for $v \in \mathcal{Y}_{\mathrm{t}}$ follows by density. □

Next, we define for any $f_\circ \in L^2(\Omega)$ and $g \in L^2(\Gamma_-, |\mathbf{b}\cdot\mathbf{n}|)$ a linear form $f \in \mathcal{Y}_{\mathrm{t}}'$ as

$$f(v) := (f_\circ, v)_{L^2(\Omega)} + \int_{\Gamma_-} g\gamma_-(v)|\mathbf{b}\cdot\mathbf{n}|\,\mathrm{d}s. \tag{3.14}$$

Then, we obtain the well-posedness of the variational formulation:

[5] Note that, due to a wrong estimate, the constant given in the corresponding result [43, Prop. 2.3] is generally not true. We therefore give a modified proof using (3.5) for the estimate in question.

Theorem 3.1.7 ([43, Thm. 2.4])**.** *Assume that one of the two conditions in Proposition 3.1.5 is valid and b_t and f are defined as in (3.11) and (3.14), respectively. Then, there exists a unique $u \in L^2(\Omega)$ such that*

$$b_\mathrm{t}(u,v) = f(v) \quad \forall v \in \mathcal{Y}_\mathrm{t}, \tag{3.15}$$

and the stability estimate $\|u\|_{L^2(\Omega)} \leq \|f\|_{\mathcal{Y}_\mathrm{t}'}$ holds. Moreover,

$$\gamma_\mathrm{t} := \sup_{w\in L^2(\Omega)} \sup_{v\in\mathcal{Y}_\mathrm{t}} \frac{b_\mathrm{t}(w,v)}{\|w\|_{L^2(\Omega)}\|v\|_{\mathcal{Y}_\mathrm{t}}} = 1,$$

$$\beta_\mathrm{t} := \inf_{w\in L^2(\Omega)} \sup_{v\in\mathcal{Y}_\mathrm{t}} \frac{b_\mathrm{t}(w,v)}{\|w\|_{L^2(\Omega)}\|v\|_{\mathcal{Y}_\mathrm{t}}} = 1,$$

i.e., inf-sup and continuity constants are unity and, equivalently,

$$\|B_\mathrm{t}\|_{\mathcal{L}(L^2(\Omega),\mathcal{Y}_\mathrm{t}')} = \|B_\mathrm{t}^*\|_{\mathcal{L}(\mathcal{Y}_\mathrm{t},L^2(\Omega))} = \|B_\mathrm{t}^{-1}\|_{\mathcal{L}(\mathcal{Y}_\mathrm{t}',L^2(\Omega))} = \|B_\mathrm{t}^{-*}\|_{\mathcal{L}(L^2(\Omega),\mathcal{Y}_\mathrm{t})} = 1,$$

where $B_\mathrm{t}^{-*} := (B_\mathrm{t}^*)^{-1} = (B_\mathrm{t}^{-1})^* : L^2(\Omega) \to \mathcal{Y}_\mathrm{t}$.

Proof. The proof follows the lines of the proof of [43, Thm. 2.4] invoking Proposition 3.1.6 instead of [43, Prop. 2.3]. □

Remark 3.1.8. As the last theorem shows, the specific choice of the test space norm $\|\cdot\|_{\mathcal{Y}_\mathrm{t}} = \|B_\mathrm{t}^* \cdot \|_{L^2(\Omega)}$ results in an optimally stable variational formulation. However, as already noted in [43, Prop. 2.6], our test space is the same space as the more "standard" space $H_{\Gamma_+}(\Omega, \mathrm{b})$ (see subsection 2.3.1) with equivalent norms: One the one hand, we have

$$\begin{aligned}\|w\|_{\mathcal{Y}_\mathrm{t}}^2 = \|B_\mathrm{t}^* w\|_{L^2(\Omega)}^2 &= \|-\mathbf{b}\cdot\nabla w + w(c - \nabla\cdot\mathbf{b})\|_{L^2(\Omega)}^2 \\ &\leq 2\|-\mathbf{b}\cdot\nabla w\|_{L^2(\Omega)}^2 + 2\|(c-\nabla\cdot\mathbf{b})\|_{L^\infty(\Omega)}^2\|w\|_{L^2(\Omega)}^2 \\ &\leq 2\max\{1, \|(c-\nabla\cdot\mathbf{b})\|_{L^\infty(\Omega)}^2\}\|w\|_{H(\Omega,\mathbf{b})}^2.\end{aligned}$$

On the other hand, it holds due to (3.5)

$$\begin{aligned}\|w\|_{H(\Omega,\mathbf{b})}^2 &= \|w\|_{L^2(\Omega)}^2 + \|\mathbf{b}\cdot\nabla w\|_{L^2(\Omega)}^2 \\ &\leq \|w\|_{L^2(\Omega)}^2 + 2\|(c-\nabla\cdot\mathbf{b})w\|_{L^2(\Omega)}^2 + 2\|\mathbf{b}\cdot\nabla w - w(c-\nabla\cdot\mathbf{b})\|_{L^2(\Omega)}^2 \\ &\leq (c_p^2(1 + 2\|(c-\nabla\cdot\mathbf{b})\|_{L^\infty(\Omega)}^2) + 2)\|w\|_{\mathcal{Y}_\mathrm{t}}^2.\end{aligned}$$

Hence, the norms are equivalent, and thus $\mathcal{Y}_\mathrm{t} = \mathrm{clos}_{\|\cdot\|_{H(\Omega,\mathbf{b})}}(C_{\Gamma_+}^1(\Omega))$. Due to Proposition 2.3.5, $C_{\Gamma_+}^1(\Omega)$ is dense in $H_{\Gamma_+}(\Omega,\mathbf{b})$ (for which the boundary values were defined in the trace sense). Thus, we indeed have $\mathcal{Y}_\mathrm{t} = H_{\Gamma_+}(\Omega,\mathbf{b})$ with equivalent norms.

Example 3.1.9 (Time-dependent linear transport equations)**.** The setting described in the beginning of this section includes both time-independent and time-dependent linear first order transport problems: As remarked in [43], we can consider time as an additional transport direction in the space-time domain, i.e., $z = (t,x) \in \Omega := (0,T)\times D = I \times D$, $n = 1+d$, where $D \subset \mathbb{R}^d$ denotes the spatial domain. Next, we define the space-time transport direction $\mathbf{b} := (1, \mathbf{b}_x)^T \in C^1(\overline{I\times D})^{1+d}$, where $\mathbf{b}_x$ denotes the spatial advective field. Moreover, we introduce the space-time gradient operator as $\nabla := (\partial_t, \nabla_x)^T$, where ∇_x is the gradient on the spatial domain D. Accordingly, we set $\Gamma_\pm := \{(t,x) \in \Gamma : \mathbf{b}(t,x) \cdot \mathbf{n}(t,x) \gtrless 0\}$, where $\mathbf{n}(t,x)$ is again the outward normal of Γ. Then, we obtain exactly the form (3.1), namely

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

For the space-time boundary, we have $\Gamma = \overline{\Gamma_{\text{in}} \cup \Gamma_{\text{out}} \cup \Gamma_D}$, where $\Gamma_{\text{in}} := \{0\} \times D$, $\Gamma_{\text{out}} := \{T\} \times D$, $\Gamma_D := I \times \partial D$, along with its corresponding outward normals $\mathbf{n}_{\text{in}} := (-1,0)^T$, $\mathbf{n}_{\text{out}} := (1,0)^T$ and $\mathbf{n}_D := (0,\mathbf{n}_x)^T$, where $\mathbf{n}_x$ denotes the spatial outward normal (of D). Hence, $\mathbf{b} \cdot \mathbf{n}_{\text{in}} = -1$, $\mathbf{b} \cdot \mathbf{n}_{\text{out}} = 1$, and $\mathbf{b} \cdot \mathbf{n}_D = \mathbf{b}_x \cdot \mathbf{n}_x$, so that $\Gamma_- = \overline{\Gamma_{\text{in}} \cup \Gamma_{D_-}}$, where $\Gamma_{D_\pm} = I \times \partial D_\pm$ and $\partial D_\pm := \{x \in \partial D : \mathbf{b}_x(x) \cdot \mathbf{n}_x(x) \gtrless 0\}$. We emphasize that also nonhomogeneous initial values are thus prescribed in an essential manner. Note that for the time-dependent case condition (i) of Proposition 3.1.5 is always fulfilled, which can be seen by taking $\mathbf{k} = (1,0,\dots,0)^T$ in Proposition 3.1.4, since $\mathbf{k} \cdot \mathbf{b} \equiv 1$ (cf. [6]). As an alternative to this realization of a space-time formulation, one could also treat spatial and temporal variables separately. Such a *strong in time* variational formulation, however, results in a suboptimal inf-sup constant that may be positive only for short time intervals. For details see [28, SM2].

3.2 An optimally stable Petrov-Galerkin method

In this section we introduce computationally feasible and optimally stable (conforming) finite-dimensional trial and test spaces $\mathcal{X}_\mathrm{t}^\delta \subset \mathcal{X}_\mathrm{t} = L^2(\Omega)$ and $\mathcal{Y}_\mathrm{t}^\delta \subset \mathcal{Y}_\mathrm{t}$ for the approximation of the solution of (3.15). Here, we denote by δ a discretization parameter, where δ equals the mesh size h for spatial problems and $\delta = (\Delta t, h)$ for time-dependent problems in space and time with a time step Δt[6]. Then, the Petrov-Galerkin approximation of (3.15) reads

$$u^\delta \in \mathcal{X}_\mathrm{t}^\delta : \qquad b_\mathrm{t}(u^\delta, v^\delta) = f(v^\delta) \quad \forall v^\delta \in \mathcal{Y}_\mathrm{t}^\delta. \tag{3.16}$$

From Proposition 2.2.7, we know that (3.16) has a unique solution $u^\delta \in \mathcal{X}_\mathrm{t}^\delta$ if

$$\beta_\mathrm{t}^\delta := \inf_{w^\delta \in \mathcal{X}_\mathrm{t}^\delta} \sup_{v^\delta \in \mathcal{Y}_\mathrm{t}^\delta} \frac{b_\mathrm{t}(w^\delta, v^\delta)}{\|w^\delta\|_{L^2(\Omega)} \|v^\delta\|_{\mathcal{Y}_\mathrm{t}}} > 0, \tag{3.17}$$

where we additionally require stability of the scheme as $\delta \to 0$, i.e., that there is $\bar{\beta}_\mathrm{t} > 0$ such that

$$\beta_\mathrm{t}^\delta \geq \bar{\beta}_\mathrm{t} > 0, \qquad \forall \delta > 0.$$

The stability (or inf-sup) constant β_t^δ also plays a key role for the relation of the *error* $e^\delta := u - u^\delta$ and the *residual* $r^\delta \in \mathcal{Y}_\mathrm{t}'$ defined as

$$r^\delta(w) := f(w) - b_\mathrm{t}(u^\delta, w) = b_\mathrm{t}(e^\delta, w), \qquad w \in \mathcal{Y}_\mathrm{t},$$

[6] If we use a tensor product discretization in space, δ may also take the form $\delta = (h_1, \dots, h_d)$ or $\delta = (\Delta t, h_1, \dots, h_d)$, respectively.

as can be seen by the standard lines

$$\begin{aligned}\bar{\beta}_{\mathrm{t}}\,\|e^{\delta}\|_{L^2(\Omega)} \leq \sup_{w\in\mathcal{Y}_{\mathrm{t}}} \frac{b_{\mathrm{t}}(e^{\delta},w)}{\|w\|_{\mathcal{Y}_{\mathrm{t}}}} = \sup_{w\in\mathcal{Y}_{\mathrm{t}}} \frac{r^{\delta}(w)}{\|w\|_{\mathcal{Y}_{\mathrm{t}}}} = \|r^{\delta}\|_{\mathcal{Y}_{\mathrm{t}}'} \\ \leq \gamma_{\mathrm{t}} \sup_{w\in\mathcal{Y}_{\mathrm{t}}} \frac{\|e^{\delta}\|_{L^2(\Omega)}\,\|w\|_{\mathcal{Y}_{\mathrm{t}}}}{\|w\|_{\mathcal{Y}_{\mathrm{t}}}} = \gamma_{\mathrm{t}}\,\|e^{\delta}\|_{L^2(\Omega)}.\end{aligned}$$

In the optimal case, i.e., $\bar{\beta}_{\mathrm{t}} = \gamma_{\mathrm{t}} = 1$, error and residual coincide, i.e., $\|e^{\delta}\|_{L^2(\Omega)} = \|r^{\delta}\|_{\mathcal{Y}_{\mathrm{t}}'}$. Moreover, we have the quasi-best approximation result Proposition 2.2.9

$$\|e^{\delta}\|_{L^2(\Omega)} = \|u-u^{\delta}\|_{L^2(\Omega)} \leq \frac{\gamma_{\mathrm{t}}}{\bar{\beta}_{\mathrm{t}}} \inf_{v^{\delta}\in\mathcal{X}_{\mathrm{t}}^{\delta}} \|u-v^{\delta}\|_{L^2(\Omega)} = \frac{\gamma_{\mathrm{t}}}{\bar{\beta}_{\mathrm{t}}}\,\sigma_{L^2(\Omega)}(u;\mathcal{X}_{\mathrm{t}}^{\delta}),$$

where $\sigma_{L^2(\Omega)}(u;\mathcal{X}_{\mathrm{t}}^{\delta}) := \inf_{v^{\delta}\in\mathcal{X}_{\mathrm{t}}^{\delta}} \|u-v^{\delta}\|_{L^2(\Omega)}$ denotes the error of the best approximation to an element $u\in L^2(\Omega)$ in $\mathcal{X}_{\mathrm{t}}^{\delta}$ w.r.t. the $L^2(\Omega)$-norm. Since $u^{\delta}\in\mathcal{X}_{\mathrm{t}}^{\delta}$, it is trivially seen that $\sigma_{L^2(\Omega)}(u;\mathcal{X}_{\mathrm{t}}^{\delta}) \leq \|u-u^{\delta}\|_{L^2(\Omega)} = \|e^{\delta}\|_{L^2(\Omega)}$, so that in the optimal case $\bar{\beta}_{\mathrm{t}} = \gamma_{\mathrm{t}} = 1$ it holds that

$$\|r^{\delta}\|_{\mathcal{Y}_{\mathrm{t}}'} = \|e^{\delta}\|_{L^2(\Omega)} = \sigma_{L^2(\Omega)}(u;\mathcal{X}_{\mathrm{t}}^{\delta}), \tag{3.18}$$

i.e., the numerical approximation is the best approximation.

3.2.1 Optimally stable discrete spaces

To realize an optimally conditioned and thus optimally stable Petrov-Galerkin method, which is also computationally feasible, we suggest to first choose a conformal finite-dimensional test space $\mathcal{Y}_{\mathrm{t}}^{\delta}\subset\mathcal{Y}_{\mathrm{t}}$ and then to set

$$\mathcal{X}_{\mathrm{t}}^{\delta} := B_{\mathrm{t}}^{*}(\mathcal{Y}_{\mathrm{t}}^{\delta}) \subset L^2(\Omega). \tag{3.19}$$

For this pair of trial and test spaces we then obtain for every $w^{\delta}\in\mathcal{X}_{\mathrm{t}}^{\delta}$ that

$$\sup_{v^{\delta}\in\mathcal{Y}_{\mathrm{t}}^{\delta}} \frac{b_{\mathrm{t}}(w^{\delta},v^{\delta})}{\|w^{\delta}\|_{L^2(\Omega)}\|v^{\delta}\|_{\mathcal{Y}_{\mathrm{t}}}} = \frac{b_{\mathrm{t}}(w^{\delta},B_{\mathrm{t}}^{-*}w^{\delta})}{\|w^{\delta}\|_{L^2(\Omega)}\|B_{\mathrm{t}}^{-*}w^{\delta}\|_{\mathcal{Y}_{\mathrm{t}}}} = \frac{(w^{\delta},B_{\mathrm{t}}^{*}B_{\mathrm{t}}^{-*}w^{\delta})_{L^2(\Omega)}}{\|w^{\delta}\|_{L^2(\Omega)}\|B_{\mathrm{t}}^{*}B_{\mathrm{t}}^{-*}w^{\delta}\|_{L^2(\Omega)}} \equiv 1. \tag{3.20}$$

Here, we have exploited the fact that for all $w^{\delta}\in\mathcal{X}_{\mathrm{t}}^{\delta}$ for the supremizer $s_{w^{\delta}}^{\delta}\in\mathcal{Y}_{\mathrm{t}}^{\delta}$, defined as the solution of $(s_{w^{\delta}}^{\delta},v^{\delta})_{\mathcal{Y}_{\mathrm{t}}} = b_{\mathrm{t}}(w^{\delta},v^{\delta})$ for all $v^{\delta}\in\mathcal{Y}_{\mathrm{t}}^{\delta}$, we have $s_{w^{\delta}}^{\delta} = B_{\mathrm{t}}^{-*}w^{\delta}$ as B_{t}^{*} is boundedly invertible. From (3.20) we may thus conclude that indeed

$$\beta_{\mathrm{t}}^{\delta} = \gamma_{\mathrm{t}}^{\delta} = 1 \tag{3.21}$$

and the proposed method is optimally stable.

Moreover, we emphasize that the suggested approach is computationally feasible since B_{t}^{*} is a differential operator which can easily be applied – as long as the test space is formed by "easy" functions such as splines as in the case of FEs. Additionally, for our choice of test and trial space we may reformulate the discrete problem (3.16) as follows: Thanks to the definition of the trial space $\mathcal{X}_{\mathrm{t}}^{\delta}$ in (3.19), there exists for all $v^{\delta}\in\mathcal{X}_{\mathrm{t}}^{\delta}$ a unique $w^{\delta}\in\mathcal{Y}_{\mathrm{t}}^{\delta}$ such that $v^{\delta} = B_{\mathrm{t}}^{*}w^{\delta}$. Therefore, the problem (3.16) is equivalent to the problem

$$w^{\delta}\in\mathcal{Y}_{\mathrm{t}}^{\delta}: \qquad a(w^{\delta},v^{\delta}) := (B_{\mathrm{t}}^{*}w^{\delta},B_{\mathrm{t}}^{*}v^{\delta})_{L^2(\Omega)} = f(v^{\delta}) \quad \forall v^{\delta}\in\mathcal{Y}_{\mathrm{t}}^{\delta}, \tag{3.22}$$

which obviously is a symmetric and coercive problem, the normal equations, or a least-squares problem. Thus, problem (3.22) is well-posed and we identify the solution of (3.16) as $u^\delta := B_\mathrm{t}^* w^\delta$. This reformulation will also be used for the implementation of the framework. From (3.22) we see that for the setup of the linear system for w^δ the precise knowledge of the basis of $\mathcal{X}_\mathrm{t}^\delta = B_\mathrm{t}^* \mathcal{Y}_\mathrm{t}^\delta$ is not needed; it is needed only for the pointwise evaluation of u^δ when e.g. visualizing the solution. For further details on the computational realization we refer the reader to section 3.4.

Thanks to (3.21), we are, moreover, in the optimal case described in the beginning of this section and the numerical approximation $u^\delta \in \mathcal{X}_\mathrm{t}^\delta$ is thus the best approximation of $u \in L^2(\Omega)$ for our suggested choice of trial and test space. Hence, we obtain $\|e^\delta\|_{L^2(\Omega)} = \sigma_{L^2(\Omega)}(u, \mathcal{X}_\mathrm{t}^\delta) = \|r^\delta\|_{\mathcal{Y}_\mathrm{t}'}$. Due to (3.19) we have that for any $w^\delta \in \mathcal{X}_\mathrm{t}^\delta$ there exists a unique $v^\delta \in \mathcal{Y}_\mathrm{t}^\delta$ with $B_\mathrm{t}^* v^\delta = w^\delta$. In view of (B1) in Assumption 3.1.1, there also exists a unique $v \in \mathcal{Y}_\mathrm{t}$ such that $B_\mathrm{t}^* v = u$, namely $v^* = B_\mathrm{t}^{-*} u$. Therefore,

$$\begin{aligned}\|e^\delta\|_{L^2(\Omega)} &= \sigma_{L^2(\Omega)}(u, \mathcal{X}_\mathrm{t}^\delta) = \inf_{w^\delta \in \mathcal{X}_\mathrm{t}^\delta} \|u - w^\delta\|_{L^2(\Omega)} = \inf_{v^\delta \in \mathcal{Y}_\mathrm{t}^\delta} \|B_\mathrm{t}^* v - B_\mathrm{t}^* v^\delta\|_{L^2(\Omega)} \\ &= \inf_{v^\delta \in \mathcal{Y}_\mathrm{t}^\delta} \|v - v^\delta\|_{\mathcal{Y}_\mathrm{t}} = \sigma_{\mathcal{Y}_\mathrm{t}}(B_\mathrm{t}^{-*} u, \mathcal{Y}_\mathrm{t}^\delta).\end{aligned} \tag{3.23}$$

We may thus also infer from (3.23) the (strong) convergence of the approximation u^δ to u in $L^2(\Omega)$ provided that $\inf_{v^\delta \in \mathcal{Y}_\mathrm{t}^\delta} \|v - v^\delta\|_{\mathcal{Y}_\mathrm{t}}$ converges to 0 as $\delta \to 0$. Note that the latter can be ensured by choosing an appropriate test space $\mathcal{Y}_\mathrm{t}^\delta$, such as, say, a standard FE space.

We finally remark that in standard FE methods the error analysis is usually done in two steps: (1) relation of the error to the best approximation by a Céa-type lemma; (2) proving an asymptotic rate of convergence e.g. by using a Clément-type interpolation operator. As seen above, (1) also holds for our new trial spaces – in a nonstandard norm, however. Regarding the second step (2) there is hope that it might maybe be possible to derive convergence rates via the term $\inf_{v^\delta \in \mathcal{Y}_\mathrm{t}^\delta} \|v - v^\delta\|_{\mathcal{Y}_\mathrm{t}}$ (see (3.23)) and mapping properties of the operator B_t. This is, however, beyond the scope of this work. We will instead investigate the rate of convergence in numerical experiments in section 3.5.

Example 3.2.1 (Illustration of trial space)**.** We illustrate the trial space $\mathcal{X}_\mathrm{t}^\delta$ as defined in (3.19) for a very simple, one-dimensional problem. In detail, we consider $\Omega := (0,1)$, a constant transport term $b > 0$, and a variable reaction coefficient $c \in C^0([0,1])$; that means $B_{\mathrm{t},\circ} u(x) := b\, u'(x) + c(x)\, u(x)$, $x \in \Omega$, as well as $u(0) = g$ on $\Gamma_- = \{0\}$. We get $B_{\mathrm{t},\circ}^* v(x) := -b\, v'(x) + c(x)\, v(x)$. According to our proposed approach, we start by defining a test space $\mathcal{Y}_\mathrm{t}^h$. To this end, let $n_h \in \mathbb{N}$ and $h := \frac{1}{n_h}$, $I_i := [(i-1)h, ih) \cap \bar{\Omega}$, $i = 1, \dots, n_h$, $I_0 := \emptyset$. We use standard piecewise linear FEs, i.e.,

$$\eta_i(x) := \begin{cases} \frac{x}{h} + 1 - i, & \text{if } x \in I_{i-1}, \\ -\frac{x}{h} + 1 + i, & \text{if } x \in I_i, \\ \quad 0, & \text{else}, \end{cases}$$

for $i = 1, \dots, n_h$ and define $\mathcal{Y}_\mathrm{t}^h := \operatorname{span}\{\eta_1, \dots, \eta_{n_h}\}$. Then, we construct the optimal

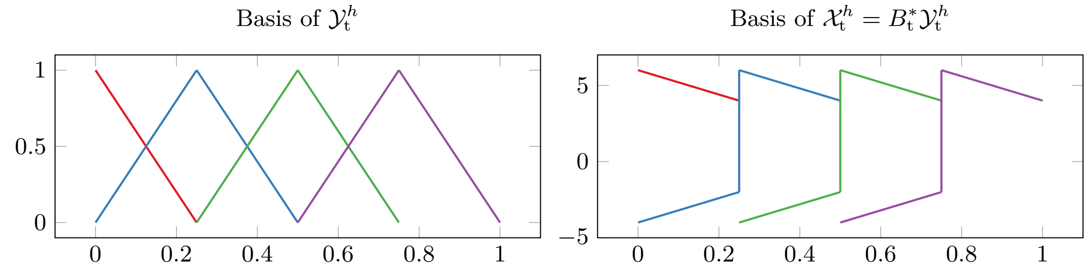

Figure 3.2: Basis functions of $\mathcal{Y}_t^h$ and $\mathcal{X}_t^h$ for $h = \frac{1}{4}$, $b \equiv 1$, $c \equiv 2$.

trial space in the above sense by $\mathcal{X}_t^h := \operatorname{span}\{\xi_1, \dots, \xi_{n_h}\}$, where we set

$$\xi_i(x) := B_t^* \eta_i(x) = -b\,\eta_i'(x) + c(x)\,\eta_i(x) = \begin{cases} -\frac{b}{h} + c(x)(\frac{x}{h} + 1 - i), & \text{if } x \in I_{i-1}, \\ \frac{b}{h} + c(x)(-\frac{x}{h} + 1 + i), & \text{if } x \in I_i, \\ \quad 0, & \text{else,} \end{cases}$$

for $i = 1, \dots, n_h$. Note, that for the special case of constant reaction $c(x) \equiv c$, the functions ξ_i are piecewise linear and discontinuous, see Figure 3.2.

3.2.2 Nonphysical restrictions at the boundary

From a computational perspective it is appealing to use discrete spaces that are tensor products of one-dimensional spaces; for details see section 3.4. However, this choice may result in nonphysical restrictions of functions in the trial space on certain parts of the outflow boundary.

To illustrate this, consider $\Omega = (0,1)^2$ and let $\mathbf{b} \equiv (b_1, b_2)^T \in \mathbb{R}^2$, $c \in \mathbb{R}$ with $b_1, b_2 > 0$, such that we have for the inflow boundary $\Gamma_- = (\{0\} \times (0,1)) \cup ((0,1) \times \{0\})$ and thus for the outflow boundary $\Gamma_+ = (\{1\} \times (0,1)) \cup ((0,1) \times \{1\})$. Let $\mathcal{Y}_{t,1D}^h$ be a univariate finite-dimensional space with $\mathcal{Y}_{t,1D}^h = \operatorname{span}\{\phi_1, \dots, \phi_{n_h}\} \subset H_{(1)}^1(0,1) := \{v \in H^1(0,1) : v(1) = 0\}$. Next, we define the discrete test space on $\Omega = (0,1)^2$ as the tensor product space

$$\mathcal{Y}_t^\delta := \mathcal{Y}_{t,1D}^h \otimes \mathcal{Y}_{t,1D}^h = \operatorname{span}\{\phi_i \otimes \phi_j : 1 \le i,j \le n_h\}, \qquad \delta = (h,h).$$

Then, the optimal trial functions are given for $i, j, = 1, \dots, n_h$ by

$$\psi_{i,j} := B_t^*(\phi_i \otimes \phi_j) = -b_1(\phi_i' \otimes \phi_j) - b_2(\phi_i \otimes \phi_j') + c(\phi_i \otimes \phi_j)$$

and we set $\mathcal{X}_t^\delta := \operatorname{span}\{\psi_{i,j} : 1 \le i,j \le n_h\}$. However, this simple tensor product ansatz results in $\psi_{i,j}(1,1) = 0$ for all i and j; i.e., any numerical approximation would vanish at the right upper corner $(1,1) \in \overline{\Omega}$. Needless to say, this is a nonphysical restriction at the boundary, even though point values do not matter for an L^2-approximation. It is obvious that the 2D case is only the simplest one in which this effect appears. In fact, in a general dD situation ($d \ge 2$), we would obtain that optimal trial functions constructed as the B_t^*-image of tensor products would vanish on $(d-2)$-dimensional sets along the boundary of $\overline{\Omega}$, leading to nonphysical boundary values. To reduce the impact of this effect, we suggest considering an additional "layer" around the computational domain by defining a tube of width $\alpha > 0$ around Γ_+ by

$$\Omega_+(\alpha) := \{x \in \mathbb{R}^n \setminus \Omega : \exists y \in \Gamma_+ : \|x - y\|_\infty < \alpha\}, \qquad \Omega(\alpha) := \Omega \cup \Omega_+(\alpha). \tag{3.24}$$

Then, we solve the original transport problem on the extended domain $\Omega(\alpha)$ using the associated pair of optimal trial and test spaces. As a result, the trial functions vanish on the exterior boundary of $\Omega_+(\alpha)$, but not on $\partial\Omega$. From a numerical perspective, by choosing $\alpha = mh$ for a (small) $m \in \mathbb{N}$ and the mesh size h, this adds m layers of grid cells and thus $\mathcal{O}(n_h^{d-1})$ degrees of freedom. On the larger domain $\Omega_+(\alpha)$, the numerical solution remains a best-approximation in the enlarged trial space. Due to the larger dimension, this is no longer true w.r.t. the original domain Ω. However, note that the additional unknowns are only $(d-1)$-dimensional. We will numerically investigate this effect in section 3.5.

3.2.3 Postprocessing

As already mentioned, we are particularly interested in using our framework for problems with nonregular solutions $u \in L^2(\Omega)$, which especially includes jump discontinuities that are transported through the domain. However, it is well known that (piecewise) polynomial L^2-approximations of such discontinuities result – especially for higher polynomial orders – in overshoots; this is the so-called Gibbs phenomenon. There are many works concerning postprocessing techniques to mitigate such effects, see, for instance, [122] and the references therein.

Here, we restrict ourselves to a rather simple postprocessing procedure aimed at limiting the solution near jump discontinuities. Let $\mathcal{Y}_{\mathrm{t}}^\delta \subset \mathcal{Y}_{\mathrm{t}}$ be a conforming FE test space on $\Omega \subset \mathbb{R}^n$ corresponding to a partition $\mathcal{T}_\delta = \{K_i\}_{i=1}^{n_{\mathcal{T}_\delta}}$ of $\Omega = \bigcup_{i=1}^{n_{\mathcal{T}_\delta}} K_i$ with polynomial order $p \geq 2$:

$$\mathcal{Y}_{\mathrm{t}}^\delta := \{v \in C^0(\Omega) : v|_K \in \mathbb{P}^p(K)\, \forall K \in \mathcal{T}_\delta, v|_{\Gamma_+} = 0\} \subset \mathcal{Y}_{\mathrm{t}}.$$

If $w^\delta \in \mathcal{Y}_{\mathrm{t}}^\delta$ denotes the solution to (3.22), the solution $u^\delta \in \mathcal{X}_{\mathrm{t}}^\delta = B_{\mathrm{t}}^* \mathcal{Y}_{\mathrm{t}}^\delta$ to (3.16) reads

$$u^\delta = B_{\mathrm{t}}^* w^\delta = -\sum_{i=1}^n b_i \partial_{x_i} w^\delta + (c - \nabla \cdot \mathbf{b}) w^\delta.$$

Since $w^\delta \in \mathcal{Y}_{\mathrm{t}}^\delta$ is an FE function, the partial derivatives $\partial_{x_i} w^\delta, i = 1, \ldots, n$ contain discontinuities across the cell boundaries, such that limiting these terms has the potential to mitigate overshoot effects. For all $K \in \mathcal{T}_\delta$, we have $\partial_{x_i} w^\delta|_K \in \mathbb{P}^p(K)$. Based upon this, we define

$$\widetilde{\partial_{x_i} w^\delta} \in L^2(\Omega) \quad \text{by} \quad \widetilde{\partial_{x_i} w^\delta}|_K := P_{\mathbb{P}^{(p-1)}(K)} \partial_{x_i} w^\delta|_K \quad \forall K \in \mathcal{T}_\delta,$$

where $P_{\mathbb{P}^{(p-1)}(K)}$ is the L^2-orthogonal projection onto the polynomials of order at most $(p-1)$ on K. We then define the postprocessed solution to (3.16) as

$$\tilde{u}^\delta := -\sum_{i=1}^n b_i \widetilde{\partial_{x_i} w^\delta} + (c - \nabla \cdot \mathbf{b}) w^\delta.$$

As a first attempt, one may perform the elementwise L^2-projection on all grid cells. However, for many problems it might be better (or even necessary) to choose a set of grid cells $\mathcal{T}_\delta^{\mathrm{jump}} \subset \mathcal{T}_\delta$ that contains all cells where overshoots due to the jumps indeed occur, and only perform the postprocessing for the cells $K \in \mathcal{T}_\delta^{\mathrm{jump}}$. For methods that are able to detect such cells we refer to [111].

Due to the construction of the postprocessed solution independent from the trial space $\mathcal{X}_{\mathrm{t}}^{\delta}$, it is not clear whether the postprocessed solution shows the same convergence rate as the standard solution. We will investigate the convergence behavior in numerical examples in section 3.5. We will test this approach for piecewise constant solutions u with jump discontinuities. For more complex problems, perhaps other, more sophisticated methods from the literature have to be used.

3.2.4 Comparison with other approaches

There are many different approaches to define stable FE solutions to the transport equation (3.2). As already mentioned, the variational formulation developed in section 3.1 is equivalent to the formulations used in [43] and for the DPG method [29,51,52]. On the discrete level, however, our scheme differs from the DPG method: While we here use a Petrov-Galerkin projection onto conforming spaces $\mathcal{X}_{\mathrm{t}}^{\delta} \subset L^2(\Omega)$ and $\mathcal{Y}_{\mathrm{t}}^{\delta} \subset \mathcal{Y}_{\mathrm{t}}$, in the DPG method, a mesh-dependent discrete bilinear form is defined so that discontinuous test functions can be used.

Our discrete scheme is more closely related to continuous and conforming formulations of the LSFEM for (3.2) as e.g. in [19,20,48] and to the framework presented in [43].

The LSFEM is based on the minimization of the residual usually in the L^2-norm. In [19], the L^2-residual minimization problem for (3.2) with homogeneous boundary condition $g = 0$ (after a possible lifting) is defined on the space $H_{\Gamma_-}(\Omega, \mathbf{b})$ (see (2.14)). Rewriting the minimization problem as a variational formulation, one seeks $u \in H_{\Gamma_-}(\Omega, \mathbf{b})$ such that

$$(\tilde{B}_{\mathrm{t}} u, \tilde{B}_{\mathrm{t}} v)_{L^2(\Omega)} = (f, \tilde{B}_{\mathrm{t}} v)_{L^2(\Omega)} \quad \forall v \in H_{\Gamma_-}(\Omega, \mathbf{b}), \tag{3.25}$$

where the transport operator $\tilde{B}_{\mathrm{t}} : H(\Omega, \mathbf{b}) \to L^2(\Omega)$ is the continuous extension of $B_{\mathrm{t},\circ}$ from $\mathrm{dom}(B_{\mathrm{t},\circ})$ to $H(\Omega, \mathbf{b})$[7]. In [20, sect. 10.3], the inhomogeneous boundary condition is included into the formulation, resulting in the variational problem: Find $u \in H(\Omega, \mathbf{b})$ such that

$$(\tilde{B}_{\mathrm{t}} u, \tilde{B}_{\mathrm{t}} v)_{L^2(\Omega)} + (u, v)_{L^2(\Gamma_-, |\mathbf{b}\cdot\mathbf{n}|)} = (f, \tilde{B}_{\mathrm{t}} v)_{L^2(\Omega)} + (g, \tilde{B}_{\mathrm{t}} v)_{L^2(\Gamma_-, |\mathbf{b}\cdot\mathbf{n}|)} \quad \forall v \in H(\Omega, \mathbf{b}). \tag{3.26}$$

Alternatively, in [48] a version of (3.26) with a different boundary functional is used.

We notice that the LSFEM method, viewed from the angle of the variational formulation, can be described as a "strong" weak formulation of (3.2) (i.e., with the transport operator on the trial space) and with test functions that are defined by applying the operator $\tilde{B}_{\mathrm{t}}$ to the trial functions. In this way, the method developed in subsection 3.2.1 is an "adjoint version" of LSFEM in using an ultraweak formulation and choosing the trial space as application of B_{t}^* onto the test space. From the viewpoint of the computation, the definition of w^δ in (3.22) resembles the problem (3.25) with B_{t}^* instead of $\tilde{B}_{\mathrm{t}}$ and with a different right-hand side, while we afterwards compute the solution by applying the adjoint operator, i.e., $u^\delta = B_{\mathrm{t}}^* w^\delta$.

The method developed in [43] is based on the ultraweak variational formulation of section 3.1, but differs from our method in the choice of the discrete spaces. As for our method, the goal is to achieve a residual minimization in the $\mathcal{Y}_{\mathrm{t}}'$-norm, which means that the discrete solution is the L^2-best approximation of the weak solution in $\mathcal{X}_{\mathrm{t}}^{\delta}$.

[7] $\tilde{B}_{\mathrm{t}} : H(\Omega, \mathbf{b}) \to L^2(\Omega)$ is the transport operator of a "strong" weak formulation, and is thus a restriction of the "ultraweak" operator $B_{\mathrm{t}} : L^2(\Omega) \to \mathcal{Y}_{\mathrm{t}}'$ defined in section 3.1.

Our discretization developed in subsection 3.2.1, which we call *optimal trial* approach, consists of choosing a test space $\mathcal{Y}_{\mathrm{t}}^{\delta} \subset \mathcal{Y}_{\mathrm{t}}$ which automatically determines the optimally stable trial space $\mathcal{X}_{\mathrm{t}}^{\delta} = B_{\mathrm{t}}^{*}\mathcal{Y}_{\mathrm{t}}^{\delta} \subset L^2(\Omega)$. In contrast, for the method in [43], which we call the *optimal test* method, one first chooses a trial space $\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta} \subset L^2(\Omega)$. The optimally stable test space realizing a discrete inf-sup constant of one and residual minimization in $\mathcal{Y}_{\mathrm{t}}'$ is then given by $(B_{\mathrm{t}}^{*})^{-1}\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta}$, which is not feasible for the numerical scheme. Therefore, an approximation, called δ-proximal test space[8], is defined by choosing an *auxiliary* or *test search space* $\mathcal{Z}_{\mathrm{t}}^{\delta} \subset \mathcal{Y}_{\mathrm{t}}$ of larger dimension than $\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta}$. The optimally stable problem in $\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta} \times (B_{\mathrm{t}}^{*})^{-1}\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta}$ is then substituted by the problem in $\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta} \times P_{\mathcal{Z}_{\mathrm{t}}^{\delta}}((B_{\mathrm{t}}^{*})^{-1}\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta})$, where $P_{\mathcal{Z}_{\mathrm{t}}^{\delta}}$ denotes the $\mathcal{Y}$-orthogonal projection onto $\mathcal{Z}_{\mathrm{t}}^{\delta}$. To circumvent the still prohibitively expensive computation of the basis functions of $P_{\mathcal{Z}^{\delta}}((B_{\mathrm{t}}^{*})^{-1}\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta})$, the variational problem (3.15) is reformulated into the saddle point problem: Find $(u, \hat{r}) \in L^2(\Omega) \times \mathcal{Y}$ such that

$$\begin{aligned} (B_{\mathrm{t}}^{*}\hat{r}, B_{\mathrm{t}}^{*}z)_{L^2(\Omega)} + \langle B_{\mathrm{t}}u, z\rangle_{\mathcal{Y}_{\mathrm{t}}',\mathcal{Y}_{\mathrm{t}}} &= \langle f, z\rangle_{\mathcal{Y}_{\mathrm{t}}',\mathcal{Y}_{\mathrm{t}}} && \forall z \in \mathcal{Y}_{\mathrm{t}}, \\ (v, B_{\mathrm{t}}^{*}\hat{r})_{L^2(\Omega)} &= 0 && \forall v \in L^2(\Omega), \end{aligned} \tag{3.27}$$

with the auxiliary variable $\hat{r} := (B_{\mathrm{t}}B_{\mathrm{t}}^{*})^{-1}(f - B_{\mathrm{t}}u) \in \mathcal{Y}_{\mathrm{t}}$. This, in turn, is solved *approximately* by an Uzawa algorithm, where the discretization is realized with the following iteration: Given $u^{\delta,k} \in \widehat{\mathcal{X}}_{\mathrm{t}}^{\delta}$, find $\hat{r}^{\delta,k} \in \mathcal{Z}_{\mathrm{t}}^{\delta}$ and $u^{\delta,k+1} \in \widehat{\mathcal{X}}_{\mathrm{t}}^{\delta}$ such that

$$\begin{aligned} (B_{\mathrm{t}}^{*}\hat{r}^{\delta,k}, B_{\mathrm{t}}^{*}z^{\delta})_{L^2(\Omega)} &= \langle f - B_{\mathrm{t}}^{*}u^{\delta,k}, z^{\delta}\rangle_{\mathcal{Y}_{\mathrm{t}}',\mathcal{Y}_{\mathrm{t}}} && \forall z^{\delta} \in \mathcal{Z}^{\delta}, \\ (u^{\delta,k+1}, v^{\delta})_{L^2(\Omega)} &= (u^{\delta,k}, v^{\delta})_{L^2(\Omega)} + (B_{\mathrm{t}}^{*}\hat{r}^{\delta,k}, v^{\delta})_{L^2(\Omega)} && \forall v^{\delta} \in \widehat{\mathcal{X}}_{\mathrm{t}}^{\delta}. \end{aligned} \tag{3.28}$$

Here, the first equation is based upon the same bilinear form as for (3.22). As the *optimal trial* method and the *optimal test* method are similar both in the continuous formulation as starting point and in the actual numerical problem to solve, we will compare the methods in the numerical experiments in subsection 3.5.1.

Finally, in [9] the SPLS method was proposed for abstract inf-sup stable problems and then subsequently applied to a div-curl-system. The authors first rewrite the abstract weak problem (2.5) in a least-squares form involving the operators associated to the scalar products of the trial and test space. This is equivalent to a saddle-point problem containing the bilinear form of the variational formulation and the test space scalar product, which corresponds to (3.27) when using the specific test space norm of our setting (from section 3.1 and [43]).

The authors state different versions of an Uzawa algorithm to solve the saddle point problem. Then, they propose pairs of discrete trial and test spaces, where the trial space is built from the test space by the application of the adjoint operator B^{*}, the inverse of the operator associated to the trial space scalar product and possibly a projection operator. Inserting the specific choice of spaces of our setting and using no additional projection, this corresponds to our choice of discrete spaces in subsection 3.2.1. Due to the saddle-point formulation, however, the exact discrete problem to solve differs. In fact, our solution procedure (3.22) corresponds to the first iteration of the Uzawa algorithm in [9] when applying the SPLS method to the setting of section 3.1.

[8] in [43], δ is an additional parameter different from the usage here

3.3 The reduced basis method for parametrized transport problems

In this section we generalize the above setting to problems depending on a parameter and apply the RB method for that purpose [76, 83, 112].

3.3.1 Parametrized transport problem

We consider a parametrized problem based upon a compact set of parameters $\mathcal{P} \subset \mathbb{R}^p$. In analogy to the above framework we define the domain Ω and the now possibly parameter-dependent quantities $\mathbf{b}_\mu \in C^1(\bar{\Omega})^n$ and $c_\mu \in C^0(\bar{\Omega})$ with $c_\mu - \frac{1}{2}\nabla \cdot \mathbf{b}_\mu \geq 0$ for all $\mu \in \mathcal{P}$. For all $\mu \in \mathcal{P}$ we define $f_{\mu;\circ} \in C^0(\bar{\Omega})$ and $g_\mu \in C^0(\bar{\Gamma}_-)$. Then we consider the parametric problem of finding $u_\mu : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} B_{\mu;\circ} u_\mu(z) := \mathbf{b}_\mu(z) \cdot \nabla u_\mu(z) + c_\mu(z) u_\mu(z) &= f_{\mu;\circ}(z), \quad && z \in \Omega, \\ u_\mu(z) &= g_\mu(z), && z \in \Gamma_-. \end{aligned}$$

Assumption 3.3.1. We assume that Ω, $\mathcal{P}$ and $\mathbf{b}_\mu$ are chosen such that the inflow and outflow boundaries $\Gamma_\pm := \{z \in \partial\Omega : \ \mathbf{b}_\mu(z) \cdot \mathbf{n}(z) \gtrless 0\}$ are *parameter-independent.*

Remark 3.3.2. As we shall see below, Assumption 3.3.1 is a direct consequence of a necessary density assumption to be formulated below. However, as stated in [45], for parameter-dependent $\Gamma_\pm(\mu)$ and a polyhedral domain Ω, it is always possible to decompose $\mathcal{P}$ into a finite number of subsets $\mathcal{P}_m$, $m = 1, \ldots, M$, with fixed parameter-independent corresponding inflow and outflow boundaries. Hence, one considers M subproblems on $\mathcal{P}_m$, $m = 1, \ldots, M$, with separate reduced models. Moreover, one could also consider parameter-dependent Ω_μ, $\Gamma_{\pm,\mu}$ that can be mapped onto a parameter-independent reference domain Ω with fixed inflow and outflow boundaries by varying the data.

Next, we require Assumption 3.1.1 for the formal adjoint $B^*_{\mu;\circ}$ for all $\mu \in \mathcal{P}$ such that we can apply the above framework separately for all $\mu \in \mathcal{P}$ in order to define the test space $\mathcal{Y}_{\mathrm{t},\mu}$ with parameter-dependent norm $\|v\|_{\mathcal{Y}_{\mathrm{t},\mu}} := \|B^*_\mu v\|_{L^2(\Omega)}$ as well as the extended operators $B_\mu : L^2(\Omega) \to \mathcal{Y}'_{\mathrm{t},\mu}$ and $B^*_\mu : \mathcal{Y}_{\mathrm{t},\mu} \to L^2(\Omega)$. Hence, we aim at determining solutions $u_\mu \in L^2(\Omega)$ such that

$$b_\mu(u_\mu, v) := (u_\mu, B^*_\mu v)_{L^2(\Omega)} = f_\mu(v) \quad \forall v \in \mathcal{Y}_{\mathrm{t},\mu}. \tag{3.29}$$

Note that, thanks to the definition of $\mathcal{Y}_{\mathrm{t},\mu}$, we have $\|B^{-*}_\mu\|_{\mathcal{L}(L^2(\Omega),\mathcal{Y}_{\mathrm{t},\mu})} = 1$, and therefore

$$\|u_\mu\|_{L^2(\Omega)} \leq \|f_\mu\|_{\mathcal{Y}'_{\mathrm{t},\mu}}. \tag{3.30}$$

We mention that the norms $\|\cdot\|_{\mathcal{Y}_{\mathrm{t},\mu}}$ cannot be expected to be pairwise equivalent for different $\mu \in \mathcal{P}$, which means that even the sets of two test spaces $\mathcal{Y}_{\mathrm{t},\mu_1}$, $\mathcal{Y}_{\mathrm{t},\mu_2}$, $\mu_1 \neq \mu_2$, can differ. Therefore, we define as in [45] the parameter-independent test space

$$\bar{\mathcal{Y}}_{\mathrm{t}} := \bigcap_{\mu \in \mathcal{P}} \mathcal{Y}_{\mathrm{t},\mu}, \tag{3.31}$$

where we assume that $\bar{\mathcal{Y}}_{\mathrm{t}}$ is dense in $\mathcal{Y}_{\mathrm{t},\mu}$ for all $\mu \in \mathcal{P}$.[9] Thanks to the compactness of $\mathcal{P}$, we may equip $\bar{\mathcal{Y}}_{\mathrm{t}}$ with the norm

$$\|v\|_{\bar{\mathcal{Y}}_{\mathrm{t}}} := \sup_{\mu \in \mathcal{P}} \|v\|_{\mathcal{Y}_{\mathrm{t},\mu}}.$$

The above theory of optimal trial and test spaces as well as well-posedness immediately extends to the parameter-dependent case in an obvious manner.

As usual, we assume that B_μ^* and f_μ are affine w.r.t. the parameter. In detail, we assume that there exist functions $\theta_b^q \in C^0(\bar{\mathcal{P}})$ for $q = 1, \dots, Q_b$ and $\theta_f^q \in C^0(\bar{\mathcal{P}})$ for $q_f = 1, \dots, Q_f$ and μ-independent operators $(B^q)^* \in \mathcal{L}(\bar{\mathcal{Y}}_{\mathrm{t}}, L^2(\Omega)), q = 1, \dots, Q_b$, and linear functionals $f^q \in \bar{\mathcal{Y}}_{\mathrm{t}}', q_f = 1, \dots, Q_f$, such that for all $\mu \in \mathcal{P}$ we have

$$B_\mu^* = \sum_{q=1}^{Q_b} \theta_b^q(\mu)\,(B^q)^* \in \mathcal{L}(\mathcal{Y}_{\mathrm{t},\mu}, L^2(\Omega)), \quad f_\mu = \sum_{q=1}^{Q_f} \theta_f^q(\mu)\, f^q \in \mathcal{Y}_{\mathrm{t},\mu}'. \tag{3.32}$$

Lemma 3.3.3. *Under the above assumptions, the set $\mathcal{M} := \{u_\mu \text{ solves } (3.29),\ \ \mu \in \mathcal{P}\}$ of solutions is a compact subset of $L^2(\Omega)$.*

Proof. Let $u_n(\mu_n)$ form a sequence in $\mathcal{M}$. Thanks to (3.30), (3.32), and the assumption that $\theta_f^q \in C^0(\bar{\mathcal{P}})$, $q = 1, \dots, Q_f$, there exists a subsequence $u_{n_k}(\mu_{n_k}) \in \mathcal{M}$ that converges weakly in $L^2(\Omega)$ to a limit $\tilde{u} \in L^2(\Omega)$. To infer compactness of $\mathcal{M}$, it thus remains to show that $\tilde{u} \in \mathcal{M}$. To that end, we employ the parameter values μ_{n_k} of the weakly converging subsequence $u_{n_k}(\mu_{n_k})$ to define a sequence $(\mu_{n_k})_k$ in $\mathcal{P}$. Thanks to the compactness of $\mathcal{P}$ this sequence has a weakly converging subsequence which we denote w.l.o.g. again by $(\mu_{n_k})_k$ that converges to a limit $\bar{\mu} \in \mathcal{P}$.

To show continuity of the mappings $\mu \mapsto B_\mu^*$ and $\mu \mapsto f_\mu$, we first note that we have for all $\mu \in \mathcal{P}$ and all $v \in \bar{\mathcal{Y}}_{\mathrm{t}}$ that

$$\|B_\mu^* v\|_{L^2(\Omega)} = \|v\|_{\mathcal{Y}_{\mathrm{t},\mu}} \leq \|v\|_{\bar{\mathcal{Y}}_{\mathrm{t}}} \quad \text{and} \quad \sup_{v \in \bar{\mathcal{Y}}_{\mathrm{t}}} \frac{|f_\mu(v)|}{\|v\|_{\bar{\mathcal{Y}}_{\mathrm{t}}}} \leq \sup_{v \in \mathcal{Y}_{\mathrm{t},\mu}} \frac{|f_\mu(v)|}{\|v\|_{\mathcal{Y}_{\mathrm{t},\mu}}} = \|f\|_{\mathcal{Y}_{\mathrm{t},\mu}'}$$

and thus $B_\mu^* \in \mathcal{L}(\bar{\mathcal{Y}}_{\mathrm{t}}, L^2(\Omega))$ and $f_\mu \in \bar{\mathcal{Y}}_{\mathrm{t}}'$. Thanks to the assumption that B_μ^* and f_μ are affine w.r.t. parameter we may thus infer as in [45] that for all $\mu_1, \mu_2 \in \mathcal{P}$ and all $v \in \bar{\mathcal{Y}}_{\mathrm{t}}$ we have

$$\begin{aligned} \|(B_{\mu_1}^* - B_{\mu_2}^*)v\|_{L^2(\Omega)} &\leq C_B \max_{q=1,\dots,Q_b} |\theta_b^q(\mu_1) - \theta_b^q(\mu_2)|\ \|v\|_{\bar{\mathcal{Y}}_{\mathrm{t}}}, \\ |f_{\mu_1}(v) - f_{\mu_2}(v)| &\leq C_f \max_{q=1,\dots,Q_f} |\theta_f^q(\mu_1) - \theta_f^q(\mu_2)|\ \|v\|_{\bar{\mathcal{Y}}_{\mathrm{t}}}, \end{aligned}$$

which yields the continuity of the mappings $\mathcal{P} \to \mathcal{L}(\bar{\mathcal{Y}}_{\mathrm{t}}, L^2(\Omega))$, $\mu \mapsto B_\mu^*$ and $\mathcal{P} \to \bar{\mathcal{Y}}_{\mathrm{t}}'$, $f_\mu \in \bar{\mathcal{Y}}_{\mathrm{t}}'$. As a consequence we have that for all $v \in \bar{\mathcal{Y}}_{\mathrm{t}}$ the sequences $(B_{\mu_{n_k}}^* v) \in L^2(\Omega)$ and $f_{\mu_{n_k}}(v) \in \mathbb{R}$ converge in the following sense

$$\|(B_{\mu_{n_k}}^* - B_{\bar{\mu}})v\|_{L^2(\Omega)} \to 0 \quad \text{and} \quad |f_{\mu_{n_k}}(v) - f_{\bar{\mu}}(v)| \to 0 \quad \text{for } \mu_{n_k} \to \bar{\mu}. \tag{3.33}$$

[9] This assumption, which is required, for instance, for Lemma 3.3.3, automatically implies that $\Gamma_\pm$ are parameter-independent (Assumption 3.3.1), since a homogeneous Dirichlet boundary condition on Γ_+ is included in the test spaces.

In particular, the sequence $(B^*_{\mu_{n_k}} v)$ hence converges strongly to $B_{\bar{\mu}} v$ in $L^2(\Omega)$.

We may thus infer that we have for all $v \in \bar{\mathcal{Y}}_{\mathrm{t}}$ that

$$(u_{n_k}(\mu_{n_k}), B^*_{\mu_{n_k}} v)_{L^2(\Omega)} - f_{\mu_{n_k}}(v) \longrightarrow (\tilde{u}, B^*_{\bar{\mu}} v)_{L^2(\Omega)} - f_{\bar{\mu}}(v)$$

and as a consequence $(\tilde{u}, B^*_{\bar{\mu}} v)_{L^2(\Omega)} = f_{\bar{\mu}}(v)$ for all $v \in \bar{\mathcal{Y}}_{\mathrm{t}}$. To conclude, it remains to prove that there holds $(\tilde{u}, B^*_{\bar{\mu}} v)_{L^2(\Omega)} = f_{\bar{\mu}}(v)$ for all $v \in \mathcal{Y}_{\mathrm{t},\bar{\mu}}$. To that end, consider an arbitrary function $v \in \mathcal{Y}_{\mathrm{t},\bar{\mu}}$. As $\bar{\mathcal{Y}}_{\mathrm{t}}$ is dense in $\mathcal{Y}_{\mathrm{t},\bar{\mu}}$, there exists a sequence v_n such that $\|v_n - v\|_{\mathcal{Y}_{\mathrm{t},\bar{\mu}}} \to 0$. Then, we have

$$\begin{aligned}(\tilde{u}, B^*_{\bar{\mu}} v)_{L^2(\Omega)} - f_{\bar{\mu}}(v) &= (\tilde{u}, B^*_{\bar{\mu}}(v - v_n))_{L^2(\Omega)} - f_{\bar{\mu}}(v - v_n) \\ &\leq \|\tilde{u}\|_{L^2(\Omega)} \|B^*_{\bar{\mu}}\|_{\mathcal{L}(\mathcal{Y}_{\mathrm{t},\bar{\mu}}, L^2(\Omega))} \|v - v_n\|_{\mathcal{Y}_{\mathrm{t},\bar{\mu}}} + \|f_{\bar{\mu}}\|_{\mathcal{Y}'_{\mathrm{t},\bar{\mu}}} \|v - v_n\|_{\mathcal{Y}_{\mathrm{t},\bar{\mu}}} \\ &\longrightarrow 0.\end{aligned}$$

We may thus infer that $\tilde{u} = u_{\bar{\mu}} \in \mathcal{M}$, which was to be proven. □

3.3.2 Discretization

For the discretization of the parametric problem, we introduce a parameter-independent discrete space $\mathcal{Y}^\delta_{\mathrm{t}} \subset \bar{\mathcal{Y}}_{\mathrm{t}}$. Next, for fixed $\mu \in \mathcal{P}$ we define the discrete test space and the corresponding trial space as

$$\mathcal{Y}^\delta_{\mathrm{t},\mu} := (\mathcal{Y}^\delta_{\mathrm{t}}, \|\cdot\|_{\mathcal{Y}_{\mathrm{t},\mu}}) \subset \mathcal{Y}_{\mathrm{t},\mu}, \qquad \mathcal{X}^\delta_{\mathrm{t},\mu} := B^*_\mu(\mathcal{Y}^\delta_{\mathrm{t}}) \subset L^2(\Omega).$$

Note that, for different $\mu \in \mathcal{P}$, the spaces $\mathcal{X}^\delta_{\mathrm{t},\mu}$ *differ as sets* but have the *common norm* $\|\cdot\|_{L^2(\Omega)}$, whereas the spaces $\mathcal{Y}^\delta_{\mathrm{t},\mu}$ consist of the *common set* $\mathcal{Y}^\delta_{\mathrm{t}}$ with *different norms* $\|\cdot\|_{\mathcal{Y}_{\mathrm{t},\mu}}$. By the same reasoning as for the nonparametric case (see (3.20)), we have an optimal discrete inf-sup constant for all $\mu \in \mathcal{P}$, i.e.,

$$\beta^\delta_\mu := \inf_{w^\delta \in \mathcal{X}^\delta_{\mathrm{t},\mu}} \sup_{v^\delta \in \mathcal{Y}^\delta_{\mathrm{t},\mu}} \frac{b_\mu(w^\delta, v^\delta)}{\|w^\delta\|_{L^2(\Omega)} \|v^\delta\|_{\mathcal{Y}_{\mathrm{t},\mu}}} = 1.$$

The discrete solution $u^\delta_\mu \in \mathcal{X}^\delta_{\mathrm{t},\mu}$ is then defined via

$$u^\delta_\mu \in \mathcal{X}^\delta_{\mathrm{t},\mu} : \quad b_\mu(u^\delta_\mu, v^\delta) = (u^\delta_\mu, B^*_\mu v^\delta)_{L^2(\Omega)} = f_\mu(v^\delta) \quad \forall v^\delta \in \mathcal{Y}^\delta_{\mathrm{t},\mu}. \tag{3.34}$$

As in subsection 3.2.1 we observe that problem (3.34) is equivalent to the problem

$$w^\delta_\mu \in \mathcal{Y}^\delta_{\mathrm{t},\mu} : \qquad a_\mu(w^\delta_\mu, v^\delta) := (B^*_\mu w^\delta_\mu, B^*_\mu v^\delta)_{L^2(\Omega)} = f_\mu(v^\delta) \quad \forall v^\delta \in \mathcal{Y}^\delta_{\mathrm{t},\mu} \tag{3.35}$$

and we may thus solve (3.35) and identify the solution of (3.34) as $u^\delta_\mu := B^*_\mu w^\delta_\mu$.

Remark 3.3.4. Since for all $\mu \in \mathcal{P}$ we have

$$\mathcal{X}^\delta_{\mathrm{t},\mu} = B^*_\mu(\mathcal{Y}^\delta_{\mathrm{t}}) = \sum_{q=1}^{Q_b} \theta^q_b(\mu)(B^q)^*(\mathcal{Y}^\delta_{\mathrm{t}}),$$

there holds

$$\mathcal{X}^\delta_{\mathrm{t},\mu} \subset \widehat{\mathcal{X}^\delta_{\mathrm{t}}} := (B^1)^*(\mathcal{Y}^\delta_{\mathrm{t}}) + \cdots + (B^{Q_b})^*(\mathcal{Y}^\delta_{\mathrm{t}}) \subset L^2(\Omega),$$

which means that the trial spaces for all $\mu \in \mathcal{P}$ are contained in a common discrete space with dimension $\dim \widehat{\mathcal{X}^\delta_{\mathrm{t}}} \leq Q_b \cdot \dim \mathcal{Y}^\delta_{\mathrm{t}}$.

Corollary 3.3.5. *Under the above assumptions the discrete solution set* $\mathcal{M}^\delta := \{u_\mu^\delta$ *solves* (3.34), $\mu \in \mathcal{P}\} \subset \widehat{\mathcal{X}_\mathrm{t}^\delta}$ *is a compact subset of* $\widehat{\mathcal{X}_\mathrm{t}^\delta}$.

Proof. The proof can be done completely analogously to the continuous setting exploiting that $\widehat{\mathcal{X}_\mathrm{t}^\delta}$ is a Hilbert space equipped with the L^2-inner product. □

3.3.3 Reduced scheme

We assume that we have determined a reduced test space[10] $Y^N \subset \mathcal{Y}_\mathrm{t}^\delta$ with dimension $N \in \mathbb{N}$ constructed, for instance, via a greedy algorithm (see subsection 3.3.4). Then, for each $\mu \in \mathcal{P}$ we introduce the reduced discretization with test space $Y_\mu^N := (Y^N, \|\cdot\|_{\mathcal{Y}_{\mathrm{t},\mu}}) \subset \mathcal{Y}_{\mathrm{t},\mu}^\delta$ and trial space $X_\mu^N := B_\mu^*(Y_\mu^N) \subset \mathcal{X}_{\mathrm{t},\mu}^\delta$. The reduced problem then reads

$$u_\mu^N \in X_\mu^N : \qquad b_\mu(u_\mu^N, v^N) = (u_\mu^N, B_\mu^* v^N)_{L^2(\Omega)} = f_\mu(v^N) \quad \forall v^N \in Y_\mu^N. \tag{3.36}$$

As in the high-dimensional case discussed in subsection 3.3.2, these pairs of spaces yield optimal inf-sup constants

$$\beta_\mu^N := \inf_{w^N \in X_\mu^N} \sup_{v^N \in Y_\mu^N} \frac{b_\mu(w^N, v^N)}{\|w^N\|_{L^2(\Omega)} \|v^N\|_{\mathcal{Y}_{\mathrm{t},\mu}}} = 1 \quad \forall \mu \in \mathcal{P}.$$

Hence, regardless of the choice of the “initial” reduced test space Y^N we get a perfectly stable numerical scheme without the need to stabilize. Note that this is a major difference from the related work [45], where, due to a different strategy in finding discrete spaces, a stabilization procedure is necessary. Using the least-squares-type reformulation (3.35), we can (similarly to (3.22)) first compute $w_\mu^N \in Y_\mu^N$ such that

$$a_\mu(w_\mu^N, v^N) = (B_\mu^* w_\mu^N, B_\mu^* v^N)_{L^2(\Omega)} = f_\mu(v^N) \quad \forall v^N \in Y_\mu^N, \tag{3.37}$$

and then set $u_\mu^N := B_\mu^* w_\mu^N$ as the solution of (3.36).

Offline-/Online-Decomposition

By employing the assumed affine parameter dependence of B_μ^* and f_μ, the computation of u_μ^N can be decomposed efficiently in an offline stage and an online stage: Let $\{v_i^N : i = 1, \ldots, N\}$ be a basis of the parameter-independent test space Y^N. In the offline stage, we precompute and store the following parameter-independent quantities:

$$\begin{aligned}
b_{q,i} &:= (B^q)^* v_i^N, && \text{for } q = 1, \ldots, Q_b,\ i = 1, \ldots, N, \\
A_{q_1,q_2;i,j} &:= (b_{q_1,i}, b_{q_2,j})_{L^2(\Omega)}, && \text{for } q_1, q_2 = 1, \ldots, Q_b,\ i,j = 1, \ldots, N, \\
f_{q,i} &:= f^q(v_i^N), && \text{for } q = 1, \ldots, Q_f,\ i = 1, \ldots, N.
\end{aligned}$$

[10] In order to have a clear distinction between high- and low-dimensional spaces, we use calligraphic letters for the high-dimensional and normal symbols for the reduced spaces.

In the online stage, given a new parameter $\mu \in \mathcal{P}$, we assemble for all $i, j = 1, \dots, N$

$$(\mathbf{A}_\mu^N)_{i,j} := (B_\mu^* v_i^N, B_\mu^* v_j^N)_{L^2(\Omega)} = \sum_{q_1=1}^{Q_b} \sum_{q_2=1}^{Q_b} \theta_{b_\mathrm{t}}^{q_1}(\mu) \theta_{b_\mathrm{t}}^{q_2}(\mu) A_{q_1,q_2;i,j},$$

$$(\mathbf{f}_\mu^N)_i := f_\mu(v_i^N) = \sum_{q=1}^{Q_f} \theta_f^q(\mu) f_{q,i}.$$

Next, we compute $w_\mu^N = \sum_{i=1}^N w_i(\mu)\, v_i^N \in Y^N$ as in (3.37) by solving the linear system $\mathbf{A}_\mu^N \mathbf{w}_\mu^N = \mathbf{f}_\mu^N$ of size N, where $\mathbf{w}_\mu^N := (w_i(\mu))_{i=1,\dots,N} \in \mathbb{R}^N$. The RB approximation is then determined as

$$u_\mu^N := B_\mu^* w_\mu^N = \sum_{i=1}^N w_i(\mu)\, B_\mu^* v_i^N = \sum_{i=1}^N \sum_{q=1}^{Q_b} w_i(\mu)\, \theta_{b_\mathrm{t}}^q(\mu)\, b_{q,i}.$$

3.3.4 Basis generation

While in the standard RB method a reduced trial space is generated from snapshots of the parametrized problem, the reduced discretization of our method is based upon one common reduced test space, while the reduced trial spaces are parameter-dependent. However, although we have to find a good basis of the reduced test space $Y^N \subset \mathcal{Y}_\mathrm{t}^\delta$, we still want to build the reduced model from snapshots of the problem. To that end, we again use the formulation (3.35): Given $\tilde{\mu} \in \mathcal{P}$, let $w_{\tilde{\mu}}^\delta \in \mathcal{Y}_{\mathrm{t},\tilde{\mu}}^\delta$ be the solution of (3.35), such that $u_{\tilde{\mu}}^\delta := B_{\tilde{\mu}}^* w_{\tilde{\mu}}^\delta \in \mathcal{X}_{\mathrm{t},\tilde{\mu}}^\delta$ is the solution of (3.34). If $w_{\tilde{\mu}}^\delta \in Y^N$, then we have $u_{\tilde{\mu}}^\delta \in X_{\tilde{\mu}}^N = B_{\tilde{\mu}}^* Y^N$, such that $u_{\tilde{\mu}}^N = u_{\tilde{\mu}}^\delta$ holds for the solution of (3.36). Note, however, that due to the parameter dependence of the trial spaces $u_{\tilde{\mu}}^\delta$ is only included in $X_{\tilde{\mu}}^N$, but in general $u_{\tilde{\mu}}^\delta \notin X_\mu^N$ for $\mu \neq \tilde{\mu}$ (instead, $B_\mu^* w_{\tilde{\mu}}^\delta \in X_\mu^N$). Building the reduced test space Y^N from "snapshots" of (3.35) is thus analogous to the standard RB strategy to build the reduced trial space from snapshots of the problem of interest: Although a single trial space X_μ^N is not solely spanned by snapshots, the model error $\|u_\mu^N - u_\mu^\delta\|_{L^2(\Omega)}$ is zero for all parameter values μ whose (3.35)-snapshot is included in Y^N.

Algorithm 1 describes an analogue of the standard RB strong greedy algorithm for our setting: Iteratively, we first evaluate the model errors of reduced solutions for all parameters μ in a train sample $\Xi \subset \mathcal{P}$. Then, we extend Y^N by the (3.35)-snapshot $w_{\mu^*}^\delta \in \bar{\mathcal{Y}}_\mathrm{t}^\delta$ corresponding to the worst-approximated parameter μ^*. This automatically extends $X_{\mu^*}^N$ by the (3.34)-snapshot $u_{\mu^*}^\delta \in \mathcal{X}_{\mathrm{t},\mu^*}^\delta$, such that from then on the model error for μ^* is zero.

Of course, this algorithm is computationally expensive, since we have to compute u_μ^δ for all $\mu \in \Xi$, which may not be feasible for very complex problems and a finely resolved $\Xi \subset \mathcal{P}$. It is hence desirable to use some kind of surrogate – ideally a reliable and efficient error estimator – instead of the true model error in the greedy algorithm. However, as will be seen in the next subsection, the standard error estimator is not offline-online decomposable in our setting – a problem already encountered in [45]. Therefore, we have to use error indicators instead when using the full model error is computationally not feasible. We note that until now we have not been able to prove convergence of the greedy algorithm due to the parameter-dependent trial spaces.

Algorithm 1 Strong greedy method

```
input: train sample $\Xi \subset \mathcal{P}$, tolerance $\varepsilon$
output: set of chosen parameters $S_N$, reduced test space $Y^N$
Initialize $S_0 \leftarrow \emptyset$, $Y^0 \leftarrow \{0\}$
for all $\mu \in \Xi$ do
    Compute $w_\mu^\delta$ and $u_\mu^\delta = B_\mu^* w_\mu^\delta$
end for
while true do
    if $\max_{\mu \in \Xi} \|u_\mu^\delta - u_\mu^N\|_{L^2(\Omega)} \leq \varepsilon$ then
        return
    end if
    $\mu^* \leftarrow \arg\max_{\mu \in \Xi} \|u_\mu^\delta - u_\mu^N\|_{L^2(\Omega)}$
    $S_{N+1} \leftarrow S_N \cup \{\mu^*\}$
    $Y^{N+1} \leftarrow \mathrm{span}\{w_\mu^\delta, \mu \in S_{N+1}\}$
    $N \leftarrow N + 1$
end while
```

Alternatively, to obtain a computationally more feasible offline stage one might let the strong greedy algorithm run on a small test set with relatively high tolerance and use a hierarchical a posteriori error estimator on the large(r) training set, which was proposed in a slightly different context in [123]. Another idea might be to keep a second test training set during the greedy algorithm. In order to estimate the dual norm of the residual more cheaply, one could then compute Riesz representations on the span of test training snapshots instead of the full discrete space.

3.3.5 Error analysis for the reduced basis approximation

In the online stage, for a given (new) parameter $\mu \in \mathcal{P}$ we are interested in efficiently estimating the model error $\|u_\mu^\delta - u_\mu^N\|_{L^2(\Omega)}$ to assess the quality of the reduced solution. As already mentioned above, due to the choice of the reduced spaces, the reduced inf-sup and continuity constants are unity. This means that the error, the residual, and the error of best approximation coincide also in the reduced setting (cf. (3.18)). To be more precise, defining for some $v \in L^2(\Omega)$ the discrete residual $r_\mu^\delta(v) \in (\mathcal{Y}_{\mathrm{t},\mu}^\delta)'$ as

$$\langle r_\mu^\delta(v), w^\delta \rangle_{(\mathcal{Y}_{\mathrm{t},\mu}^\delta)' \times \mathcal{Y}_{\mathrm{t},\mu}^\delta} := f(w^\delta) - (v, B_\mu^* w^\delta)_{L^2(\Omega)}, \quad w^\delta \in \mathcal{Y}_{\mathrm{t},\mu}^\delta,$$

we have

$$\|u_\mu^\delta - u_\mu^N\|_{L^2(\Omega)} = \|r_\mu^\delta(u_\mu^N)\|_{(\mathcal{Y}_{\mathrm{t},\mu}^\delta)'} = \inf_{v^N \in X_\mu^N} \|u_\mu^\delta - v^N\|_{L^2(\Omega)}.$$

In principle, $r_\mu^\delta(v) \in (\mathcal{Y}_{\mathrm{t},\mu}^\delta)'$ can be computed. However, due to the special choice of the parameter-dependent norm of $\mathcal{Y}_{\mathrm{t},\mu}^\delta$, i.e., $\|w\|_{\mathcal{Y}_{\mathrm{t},\mu}^\delta} = \|B_\mu^* w\|_{L^2(\Omega)}$, the computation of the dual norm involves applying the inverse operator $(B_\mu^*)^{-1}$ and is thus as computationally expensive as solving the discrete problem (3.34). Therefore, the computation of $\|r_\mu^\delta(u_\mu^N)\|_{(\mathcal{Y}_{\mathrm{t},\mu}^\delta)'}$ is not offline-online decomposable, so that the residual cannot be computed in an online-efficient manner.

As an alternative for the error estimation mainly in the online stage, we consider an online-efficient but nonrigorous *hierarchical error estimator* similar to the one proposed in [13]. Let $Y^N \subset Y^M \subset Y^\delta$ be nested reduced spaces with dimensions N and M, $N < M$ and denote for some $\mu \in \mathcal{P}$ by $u^N(\mu) \in X^N_\mu := B^*_\mu Y^N$, $u^M(\mu) \in X^M_\mu := B^*_\mu Y^M$ the corresponding solutions of (3.36). Then, we can rewrite the model error of u^N as

$$\|u^N - u^\delta\|_{L^2(\Omega)} = \|u^N - u^M + u^M - u^\delta\|_{L^2(\Omega)} \leq \|u^N - u^M\|_{L^2(\Omega)} + \|u^M - u^\delta\|_{L^2(\Omega)}.$$

Assuming that Y^M is large enough such that $\|u^M - u^\delta\|_{L^2(\Omega)} < \varepsilon \ll 1$, we can approximate the model error of u^N by

$$\|u^N - u^\delta\|_{L^2(\Omega)} \leq \|u^N - u^M\|_{L^2(\Omega)} + \varepsilon \approx \|u^N - u^M\|_{L^2(\Omega)},$$

which can be computed efficiently also in the online stage. In practice, Y^N and Y^M can be generated by the strong greedy algorithm with different tolerances ε_N and $\varepsilon_M \ll \varepsilon_N$. Of course, this approximation to the model error is in general not reliable, since it depends on the quality of Y^M. Reliable and rigorous variants of such an error estimator can be derived based on an appropriate saturation assumption, see [79]. Reference [79] also discusses a strategy for the use of hierarchical estimators in terms of Hermite spaces Y^M for the construction of a reduced model in the offline phase. We do not go into details here. Numerical investigations of the quality of the error estimator will be given in subsection 3.5.2.

3.3.6 Comparison with the *double greedy* algorithm

As already mentioned, our scheme is closely related to the *double greedy* framework [45], where an RB approximation of the parametrized transport equation based on the *optimal test* method discretization scheme of [43] (see also subsection 3.2.4) is developed. Hence, the scheme is based on the variational formulation described in subsection 3.3.1 but uses a discrete scheme based on the saddle-point problem (3.27). For the high-dimensional "truth solution" a discrete trial space $\widehat{\mathcal{X}}^\delta_{\mathrm{t}} \subset L^2(\Omega)$ and a "test search space"[11] $\mathcal{Z}^\delta_{\mathrm{t}} \subset \bar{\mathcal{Y}}_{\mathrm{t}}$ of larger dimension are chosen independently of the parameter (recall that $\bar{\mathcal{Y}}_{\mathrm{t}} := \bigcap_{\mu\in\mathcal{P}} \mathcal{Y}_{\mathrm{t},\mu}$, see (3.31)). Then, the "truth solution" $u^\delta_\mu \in \widehat{\mathcal{X}}^\delta_{\mathrm{t}}$ is defined by the parametrized version of (3.27): Given $\mu \in \mathcal{P}$, find $(\hat{r}^\delta_\mu, u^\delta_\mu) \in \mathcal{Z}^\delta_{\mathrm{t}} \times \widehat{\mathcal{X}}^\delta_{\mathrm{t}}$ such that

$$\begin{aligned} (B^*_{\mathrm{t},\mu}\hat{r}^\delta_\mu, B^*_{\mathrm{t},\mu}z^\delta)_{L^2(\Omega)} + \langle B_{\mathrm{t},\mu}u^\delta_\mu, z^\delta\rangle_{\mathcal{Y}'_{\mathrm{t},\mu},\mathcal{Y}_{\mathrm{t},\mu}} &= \langle f, z^\delta\rangle_{\mathcal{Y}'_{\mathrm{t},\mu},\mathcal{Y}_{\mathrm{t},\mu}} \quad && \forall z^\delta \in \mathcal{Z}^\delta_{\mathrm{t}}, \\ (v^\delta, B^*_{\mathrm{t},\mu}\hat{r}^\delta_\mu)_{L^2(\Omega)} &= 0 && \forall v^\delta \in \widehat{\mathcal{X}}^\delta_{\mathrm{t}}, \end{aligned} \tag{3.38}$$

see also [45, Eq. (3.22)]. For well-posedness and stability of (3.38), it is then assumed that $\mathcal{Z}^\delta_{\mathrm{t}}$ is large enough to ensure uniform stability in the parameter, i.e., b_μ is assumed to be inf-sup stable on $\widehat{\mathcal{X}}^\delta_{\mathrm{t}} \times P_{\mathcal{Z}^\delta_{\mathrm{t}}}((B^*_{\mathrm{t},\mu})^{-1}\widehat{\mathcal{X}}^\delta_{\mathrm{t}})$ with a uniform inf-sup estimate for all parameter values.

The reduced model is built on a pair of reduced trial and test (search) spaces $X^N \subset \widehat{\mathcal{X}}^\delta_{\mathrm{t}}$ of dimension N and $Z^{M(N)} \subset \mathcal{Z}^\delta_{\mathrm{t}}$ of dimension $M(N) \geq N$. Then, the reduced solution

[11] Note that the terminology and notation of the spaces in [45] differs from the usage in [43]. We here try to stay consistent with the notation introduced in subsection 3.2.4.

$u_\mu^N \in X^N$ is one solution component of the reduced saddle point problem given by (3.38) with the reduced spaces instead of the “truth” spaces.

In order to obtain an accurate and stable reduced model, the goal is to generate the pairs of spaces $(X^N, Z^{M(N)})$ such that X^N is built from snapshots $u_{\mu_i}^\delta, i = 1, \dots, N$ to obtain a good approximation quality and $Z^{M(N)}$ is chosen so that the uniform inf-sup condition

$$\inf_{\mu \in \mathcal{P}} \inf_{w^N \in X^N} \sup_{p^N \in Z^{M(N)}} \frac{b_\mu(w^N, p^N)}{\|w^N\|_{L^2(\Omega)} \|p^N\|_{\mathcal{Y}_{\mathrm{t},\mu}}} \geq \beta_{\min} \tag{3.39}$$

(cf. [45, Eq. 4.1 and 5.14]) is fulfilled. To that end, the *double greedy* algorithm ([45, Algorithm 5]) is proposed, in which two different algorithms are executed alternately:

- The algorithm *update-inf-sup* ([45, Algorithm 2]) takes a trial space X^{N+1} and a (not sufficiently stable) test space $Z^{M(N)}$. For a sample set of parameter values, all respective inf-sup constants are computed from the reduced system matrices by corresponding Cholesky or spectral factorizations and singular value decompositions. For the parameter value $\bar{\mu}$ with the smallest inf-sup constant, the infimizing trial space vector $w_{\bar{\mu}}^N \in X^N$ is computed. Then, the supremizer of $w_{\bar{\mu}}^N$, i.e.,

  $$\bar{p}^\delta = \operatorname*{argmax}_{p^\delta \mathcal{Z}_{\mathrm{t}}^\delta} \frac{b_{\bar{\mu}}(w_{\bar{\mu}}^N, p^\delta)}{\|p^\delta\|_{\mathcal{Y}_{\mathrm{t},\bar{\mu}}}},$$

  which is given as $\bar{p}^\delta = (B_{\bar{\mu}}^\delta)^{-*} w_{\bar{\mu}}^N$, is computed and added as a new basis function to the reduced test space, i.e., $Z^{M(N)+1} = \operatorname{span}\{Z^{M(N)}, \bar{p}^\delta\}$. This procedure is repeated k times until the test space $Z^{M(N+1)} := Z^{M(N)+k}$ is large enough such that the uniform inf-sup condition (3.39) is fulfilled.

- The algorithm *update-approximation* ([45, Algorithm 4]) takes spaces $X^N, Z^{M(N)}$ that satisfy (3.39) and then determines a new basis function for X^N by one step of a (“standard RB”) greedy algorithm: For a sample set of parameter values, an error indicator for the reduced solution is evaluated. Then, for the worst approximated parameter $\bar{\mu}$, the solution $u_{\bar{\mu}}^\delta$ of (3.38) is computed and is added to the trial space, i.e., $X^{N+1} = \operatorname{span}\{X^N, u_{\bar{\mu}}^\delta\}$.

By alternately executing *update-inf-sup* and *update-approximation*, the *double greedy* algorithm iteratively generates reduced function pairs with increasing approximation quality in the trial space and fixed lower bounds for the uniform inf-sup constant. However, unlike the reduced model from subsection 3.3.3, here the test space is generally of larger dimension than the trial space, since *update-inf-sup* may (need to) increase the dimension of the test space by more than one.

The *double greedy* algorithm includes an error indicator in *update-approximation*, while we so far only proposed a strong greedy algorithm using the true reduction errors in Algorithm 1. This indicator aims to approximate the residual $\|f - B_\mu u_\mu^N\|_{\mathcal{Y}_{\mathrm{t}}'}$ in dual norms of additional reduced test spaces that are computed by iterated re-runs of the *double greedy* algorithm (which the authors call *iterative tightening*). For details see [45, subsec. 5.1.2 and sec. 6.3].

As for the non-parametric case, we include a comparison of our new method and the *double greedy* algorithm in the numerical experiments in subsection 3.5.2.

3.4 Computational realization

In this section, we specify the implementation of the solution procedure developed in section 3.2. This is also used for the methods for parameter-dependent problems developed in section 3.3. In fact, due to our assumption of affine dependence in the parameter (3.32), the computational realization in the parametric setting is very similar to the standard setting and can be done following the offline-online decomposition described at the end of subsection 3.3.3, which is why we do not address it in this section.

To solve the discrete problem (3.16) we use the equivalent formulation (3.22); i.e., we first find $w^\delta \in \mathcal{Y}_\mathrm{t}^\delta$ such that $(B_\mathrm{t}^* w^\delta, B_\mathrm{t}^* v^\delta)_{L^2(\Omega)} = f(v^\delta)$ for all $v^\delta \in \mathcal{Y}_\mathrm{t}^\delta$, and then set $u^\delta := B_\mathrm{t}^* w^\delta \in \mathcal{X}_\mathrm{t}^\delta$. The solution procedure thus consists of first assembling and solving the problem for w^δ in $\mathcal{Y}_\mathrm{t}^\delta$ and second computing u^δ. The implementation is especially dependent on the exact form of the adjoint operator B_t^*. First, we address the case of constant data, which is easier to implement and slightly more computationally efficient than the general case which we discuss subsequently.

3.4.1 Implementation for constant data

We first consider constant data functions in the adjoint operator, which has thus the form $B_\mathrm{t}^* w := -\mathbf{b}\cdot\nabla w + cw$ for $0 \neq \mathbf{b} \in \mathbb{R}^n, c \in \mathbb{R}$. We have already seen in Example 3.2.1 that in the one-dimensional case, choosing a standard linear continuous FE space for the test space $\mathcal{Y}_\mathrm{t}^\delta$ yields a trial space $\mathcal{X}_\mathrm{t}^\delta$ with piecewise linear and discontinuous functions. This can be generalized to conforming FE test spaces with arbitrary dimension, grid, and polynomial order: If $v^\delta \in \mathcal{Y}_\mathrm{t}^\delta$ is globally continuous and polynomial on each grid cell, all terms of $B_\mathrm{t}^* v^\delta$, due to the constant data functions, are still polynomials of the same or lower order on the cells, while the gradient terms yield discontinuities on the cell boundaries. Denoting thus by $\mathcal{Y}_\mathrm{t}^\delta \subset \mathcal{Y}_\mathrm{t}$ a conforming FE space on a partition $\mathcal{T}_\delta = \{K_i\}_{i=1}^{n_{\mathcal{T}_\delta}}$ of $\Omega = \bigcup_{i=1}^{n_{\mathcal{T}_\delta}} K_i$ with polynomial order p, and by $\bar{\mathcal{X}}_\mathrm{t}^\delta \subset L^2(\Omega)$ the corresponding discontinuous FE space, i.e.,

$$\mathcal{Y}_\mathrm{t}^\delta := \{v \in C^0(\Omega) : v|_K \in \mathbb{P}^p(K)\, \forall K \in \mathcal{T}_\delta, v|_{\Gamma_+} = 0\} \subset \mathcal{Y}_\mathrm{t}, \tag{3.40}$$

$$\bar{\mathcal{X}}_\mathrm{t}^\delta := \{u \in L^2(\Omega) : u|_K \in \mathbb{P}^p(K)\, \forall K \in \mathcal{T}_\delta\} \subset L^2(\Omega), \tag{3.41}$$

we have $\mathcal{X}_\mathrm{t}^\delta = B_\mathrm{t}^* \mathcal{Y}_\mathrm{t}^\delta \subset \bar{\mathcal{X}}_\mathrm{t}^\delta$ and can determine the solution $u^\delta \in \mathcal{X}_\mathrm{t}^\delta$ in terms of the standard nodal basis of $\bar{\mathcal{X}}_\mathrm{t}^\delta$.

Let $\mathbf{B}_\mathrm{t}^* \in \mathbb{R}^{\bar{n}_x \times n_y}$ be the matrix representation of $B_\mathrm{t}^* : \mathcal{Y}_\mathrm{t}^\delta \to \bar{\mathcal{X}}_\mathrm{t}^\delta$ in the nodal bases $(\phi_1, \dots, \phi_{n_y})$ of $\mathcal{Y}_\mathrm{t}^\delta$ and $(\psi_1, \dots, \psi_{\bar{n}_x})$ of $\bar{\mathcal{X}}_\mathrm{t}^\delta$, meaning that the ith column of $\mathbf{B}_\mathrm{t}^*$ contains the coefficients of $B_\mathrm{t}^* \phi_i$ in the basis $(\psi_1, \dots, \psi_{\bar{n}_x})$, i.e., $B_\mathrm{t}^* \phi_i = \sum_{j=1}^{\bar{n}_x} [\mathbf{B}_\mathrm{t}^*]_{j,i} \psi_j$. Due to the form of the operator and the chosen spaces, the matrix $\mathbf{B}_\mathrm{t}^*$ can be computed rather easily, see the example in subsection 3.4.2. Then, the coefficient vector $\mathbf{u} = (u_1, \dots, u_{\bar{n}_x})^T$ of $u^\delta = \sum_{i=1}^{\bar{n}_x} u_i \psi_i \in \bar{\mathcal{X}}_\mathrm{t}^\delta$ can simply be computed from the coefficient vector $\mathbf{w} = (w_1, \dots, w_{n_y})$ of $w^\delta = \sum_{i=1}^{n_y} w_i \phi_i \in \mathcal{Y}_\mathrm{t}^\delta$ by $\mathbf{u} = \mathbf{B}_\mathrm{t}^* \mathbf{w}$.

To solve (3.22), we have to assemble the matrix corresponding to the bilinear form $a : \mathcal{Y}_\mathrm{t}^\delta \times \mathcal{Y}_\mathrm{t}^\delta$ with

$$a(w^\delta, v^\delta) = (B_\mathrm{t}^* w^\delta, B_\mathrm{t}^* v^\delta)_{L^2(\Omega)} = (w^\delta, v^\delta)_{\mathcal{Y}_\mathrm{t}},$$

i.e., the $\mathcal{Y}_\mathrm{t}$-inner product matrix of $\mathcal{Y}_\mathrm{t}^\delta$. One possibility for the assembly is to use the matrix $\mathbf{B}_\mathrm{t}^*$: Denoting by $\mathbf{M}_{\bar{\mathcal{X}}_\mathrm{t}^\delta} \in \mathbb{R}^{\bar{n}_x \times \bar{n}_x}$ the L^2-mass matrix of $\bar{\mathcal{X}}_\mathrm{t}^\delta$, i.e., $[\mathbf{M}_{\bar{\mathcal{X}}_\mathrm{t}^\delta}]_{i,j} =$

$(\psi_i, \psi_j)_{L^2(\Omega)}$, we see that for $\mathbf{Y}_{\mathrm{t}} := (\mathbf{B}_{\mathrm{t}}^*)^T \mathbf{M}_{\bar{\mathcal{X}}_{\mathrm{t}}^\delta} \mathbf{B}_{\mathrm{t}}^* \in \mathbb{R}^{n_y \times n_y}$ it holds that $[\mathbf{Y}_{\mathrm{t}}]_{i,j} = (B_{\mathrm{t}}^* \phi_i, B_{\mathrm{t}}^* \phi_j)_{L^2(\Omega)} = (\phi_i, \phi_j)_{\mathcal{Y}_{\mathrm{t}}}$.

The solution procedure thus consists of the following steps:

1. Assemble $\mathbf{B}_{\mathrm{t}}^*$ and $\mathbf{Y}_{\mathrm{t}}$.
2. Assemble the load vector $\mathbf{f} \in \mathbb{R}^{n_y}$, $[\mathbf{f}]_i := f(\phi_i), i = 1, \dots, n_y$.
3. Solve $\mathbf{Y}_{\mathrm{t}} \mathbf{w} = \mathbf{f}$.
4. Compute $\mathbf{u} = \mathbf{B}_{\mathrm{t}}^* \mathbf{w}$.

3.4.2 Assembling the matrices for spaces on rectangular grids

As a concrete example of how to assemble the matrices $\mathbf{B}_{\mathrm{t}}^*$ and $\mathbf{Y}$ we consider $\Omega = (0,1)^n$ and use a rectangular grid. We start with the one-dimensional case as already seen in Example 3.2.1. Let thus $\Omega = (0,1)$ and $b > 0$. Moreover, let $\mathcal{T}^h = \{[(i-1)h, ih)\}_{i=1}^{n_h}$ be the uniform one-dimensional grid with mesh size $h = 1/n_h$, fix a polynomial order $p \geq 1$, and define $\mathcal{Y}_{\mathrm{t,1D}}^{h,p}, \bar{\mathcal{X}}_{\mathrm{t,1D}}^{h,p}$ as in (3.40), (3.41). Let $(\phi_1, \dots, \phi_{n_y})$ and $(\psi_1, \dots, \psi_{\bar{n}_x})$ be the respective nodal bases of $\mathcal{Y}_{\mathrm{t,1D}}^{h,p}$ and $\bar{\mathcal{X}}_{\mathrm{t,1D}}^{h,p}$.

Moreover, let $\mathbf{I}_{\mathrm{1D}} \in \mathbb{R}^{\bar{n}_x \times n_y}$ be the matrix representation of the embedding $\mathrm{Id} : \mathcal{Y}_{\mathrm{t,1D}}^{h,p} \to \bar{\mathcal{X}}_{\mathrm{t,1D}}^{h,p}$ in the respective nodal bases, i.e., the i-th column of $\mathbf{I}_{\mathrm{1D}}$ contains the coefficients of $\phi_i \in \mathcal{Y}_{\mathrm{t,1D}}^{h,p} \subset \bar{\mathcal{X}}_{\mathrm{t,1D}}^{h,p}$ in the basis $(\psi_1, \dots, \psi_{\bar{n}_x})$, such that for $\mathbf{u} = \mathbf{I}_{\mathrm{1D}} \cdot \mathbf{w}$ it holds $\sum_{i=1}^{\bar{n}_x} u_i \psi_i = \sum_{i=1}^{n_y} w_i \phi_i$. Similarly, let $\mathbf{A}_{\mathrm{1D}} \in \mathbb{R}^{\bar{n}_x \times n_y}$ be the matrix representation of the differentiation $\frac{\mathrm{d}}{\mathrm{d}x} : \mathcal{Y}_{\mathrm{t,1D}}^{h,p} \to \bar{\mathcal{X}}_{\mathrm{t,1D}}^{h,p}, w^h \mapsto (w^h)'$. Additionally, as above, we define $\mathbf{M}_{\mathrm{1D}} \in \mathbb{R}^{\bar{n}_x \times \bar{n}_x}$, $[\mathbf{M}_{\mathrm{1D}}]_{i,j} = (\psi_i, \psi_j)_{L^2((0,1))}$ as the L^2-mass matrix of $\bar{\mathcal{X}}_{\mathrm{t1D}}^{h,p}$.

For $p = 1$, i.e., linear FEs, and a standard choice of the nodal bases the matrices $\mathbf{I}_{\mathrm{1D}}$, $\mathbf{A}_{\mathrm{1D}}$, and $\mathbf{M}_{\mathrm{1D}}$ read

$$\mathbf{I}_{\mathrm{1D}} := \begin{pmatrix} 1 & 0 & 0 & \cdots \\ 0 & 1 & 0 & \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \\ \vdots & & & \ddots \end{pmatrix}, \mathbf{A}_{\mathrm{1D}} := \frac{1}{h} \cdot \begin{pmatrix} -1 & 1 & 0 & \cdots \\ -1 & 1 & 0 & \\ 0 & -1 & 1 & \\ 0 & -1 & 1 & \\ \vdots & & & \ddots \end{pmatrix}, \mathbf{M}_{\mathrm{1D}} = h \cdot \begin{pmatrix} 1/3 & 1/6 & 0 & 0 & \cdots \\ 1/6 & 1/3 & 0 & 0 & \\ 0 & 0 & 1/3 & 1/6 & \\ 0 & 0 & 1/6 & 1/3 & \\ \vdots & & & & \ddots \end{pmatrix}.$$

With these three matrices we can then compose the matrices $\mathbf{B}_{\mathrm{1D}}^*$ and $\mathbf{Y}_{\mathrm{1D}}$ by

$$\mathbf{B}_{\mathrm{t,1D}}^* := -b \cdot \mathbf{A}_{\mathrm{1D}} + c \cdot \mathbf{I}_{\mathrm{1D}}, \qquad \mathbf{Y}_{\mathrm{1D}} := (\mathbf{B}_{\mathrm{t,1D}}^*)^T \mathbf{M}_{\mathrm{1D}} \mathbf{B}_{\mathrm{t,1D}}^*.$$

Next, we consider a rectangular domain of higher dimension, e.g., $\Omega = (0,1)^n, n \geq 2$. We choose in each dimension one-dimensional FE spaces $\mathcal{Y}_{\mathrm{t}}^i, \bar{\mathcal{X}}_{\mathrm{t}}^i$, $i = 1, \dots, n$ as in (3.40), (3.41) separately, and use the tensor product of these spaces $\mathcal{Y}_{\mathrm{t}}^\delta := \bigotimes_{i=1}^n \mathcal{Y}_{\mathrm{t}}^i, \bar{\mathcal{X}}_{\mathrm{t}}^\delta := \bigotimes_{i=1}^n \bar{\mathcal{X}}_{\mathrm{t}}^i$ as FE spaces on the rectangular grid formed by a tensor product of all one-dimensional grids. The system matrices can then be assembled from Kronecker products of the one-dimensional matrices corresponding to the spaces $\mathcal{Y}_{\mathrm{t}}^i, \bar{\mathcal{X}}_{\mathrm{t}}^i, i = 1, \dots, n$: We first assemble for $i = 1, \dots, n$ the matrices $\mathbf{I}_{\mathrm{1D}}^i$ and $\mathbf{A}_{\mathrm{1D}}^i$ corresponding to the pair of spaces $\mathcal{Y}_{\mathrm{t}}^i, \bar{\mathcal{X}}_{\mathrm{t}}^i$. Then, the matrix corresponding to the adjoint operator can be assembled by

$$\mathbf{B}_{\mathrm{t}}^* := -\sum_{i=1}^n b_i \mathbf{I}_{\mathrm{1D}}^1 \otimes \cdots \otimes \mathbf{I}_{\mathrm{1D}}^{(i-1)} \otimes \mathbf{A}_{\mathrm{1D}}^i \otimes \mathbf{I}_{\mathrm{1D}}^{(i+1)} \otimes \cdots \otimes \mathbf{I}_{\mathrm{1D}}^n + c \bigotimes_{i=1}^n \mathbf{I}_{\mathrm{1D}}^i, \tag{3.42}$$

e.g., for $n = 2$ we have

$$\mathbf{B}_{\mathrm{t},2\mathrm{D}}^{*} := -b_1(\mathbf{A}_{1\mathrm{D}}^{1} \otimes \mathbf{I}_{1\mathrm{D}}^{2}) - b_2(\mathbf{I}_{1\mathrm{D}}^{1} \otimes \mathbf{A}_{1\mathrm{D}}^{2}) + c(\mathbf{I}_{1\mathrm{D}}^{1} \otimes \mathbf{I}_{1\mathrm{D}}^{2}).$$

Similarly, the mass matrix $\mathbf{M}_{\bar{\mathcal{X}}_{\mathrm{t}}^{\delta}}$ of $\bar{\mathcal{X}}_{\mathrm{t}}^{\delta}$ can be computed from the one-dimensional mass matrices $\mathbf{M}_{1\mathrm{D}}^{i}$ of $\bar{\mathcal{X}}_{\mathrm{t}}^{i}, i = 1, \dots, n$, by $\mathbf{M}_{\bar{\mathcal{X}}_{\mathrm{t}}^{\delta}} := \bigotimes_{i=1}^{n} \mathbf{M}_{1\mathrm{D}}^{i}$, such that $\mathbf{Y}_{\mathrm{t}} := (\mathbf{B}_{\mathrm{t}}^{*})^{T} \mathbf{M}_{\bar{\mathcal{X}}_{\mathrm{t}}^{\delta}} \mathbf{B}_{\mathrm{t}}^{*}$ can also be directly assembled using the matrices $\mathbf{I}_{1\mathrm{D}}^{i}, \mathbf{A}_{1\mathrm{D}}^{i}, \mathbf{M}_{1\mathrm{D}}^{i}, i = 1, \dots, n$.

3.4.3 Implementation for nonconstant data

If the data functions $\mathbf{b}$ and c are not constant, we do not automatically get a standard FE space $\bar{\mathcal{X}}_{\mathrm{t}}^{\delta}$ in which the solution u^{δ} can be described; thus the implementation has to be adapted. A way to retain the implementation for constant data functions is to approximate the data by piecewise constants on each grid cell. Then, there holds again $u_{\delta} \in \bar{\mathcal{X}}_{\mathrm{t}}^{\delta}$, and we only have to slightly modify the implementation presented in subsection 3.4.1: Every nodal basis function $\psi_i \in \bar{\mathcal{X}}_{\mathrm{t}}^{\delta}, i = 1, \dots, \bar{n}_x$, has, due to the discontinuous FE space, a support of only one grid cell. Denoting by c^i the value of c on the grid cell of ψ_i, we define the diagonal matrix $\mathbf{c} \in \mathbb{R}^{\bar{n}_x \times \bar{n}_x}$, $[\mathbf{c}]_{i,i} := c^i$, and, similarly, the matrices $\mathbf{b}_j \in \mathbb{R}^{\bar{n}_x \times \bar{n}_x}$ corresponding to $b_j, j = 1, \dots, n$. We then simply change the scalars b_j and c in (3.42) to matrices $\mathbf{b}_j$ and $\mathbf{c}$, $j = 1, \dots, n$.

However, a piecewise constant approximation of the functions $\mathbf{b} \in C^1(\Omega)^n, c \in C^0(\Omega)$ may not lead to a sufficient accuracy of the solution. For general $\mathbf{b} \in C^1(\Omega)^n, c \in C^0(\Omega)$, we thus first assemble the $\mathcal{Y}_{\mathrm{t}}$-inner product matrix $\mathbf{Y}_{\mathrm{t}} \in \mathbb{R}^{n_y \times n_y}$ of $\mathcal{Y}_{\mathrm{t}}^{\delta}$ and the load vector $\mathbf{f} \in \mathbb{R}^{n_y}$ corresponding to the right-hand side as in standard FE implementations for elliptic equations, by using e.g. Gauss quadratures for the approximation of the integrals. We can then solve (3.22) as above by $\mathbf{w} := \mathbf{Y}_{\mathrm{t}}^{-1}\mathbf{f}$, $w^{\delta} := \sum_{i=1}^{n_y} [\mathbf{w}]_i \phi_i \in \mathcal{Y}_{\mathrm{t}}^{\delta}$. To compute the solution $u_{\delta} \in \mathcal{X}_{\mathrm{t}}^{\delta}$, we use the fact that we still have $w^{\delta} \in \bar{\mathcal{X}}_{\mathrm{t}}^{\delta}$ and $\frac{\partial w^{\delta}}{\partial x_i} \in \bar{\mathcal{X}}_{\mathrm{t}}^{\delta}, i = 1, \dots, n$, and store the corresponding $\bar{\mathcal{X}}_{\mathrm{t}}^{\delta}$-coefficients of w^{δ} and its derivatives separately, as well as the data functions. We can then evaluate $u^{\delta} = B^{*} w^{\delta}$ for arbitrary $x \in \Omega$ by evaluating all w^{δ}-dependent functions and all data functions in x and using the definition of B^{*} to get $u^{\delta}(x) = -\sum_{i=1}^{n} b_i(x) \frac{\partial w^{\delta}}{\partial x_i}(x) + (c - \nabla \cdot \mathbf{b})(x) w^{\delta}(x)$.

3.5 Numerical experiments

In this section, we report on results of our numerical experiments. We consider the parametric and the nonparametric case, starting with the latter. We are particularly interested in quantitative results concerning the rate of approximation for the discrete case as the discretization parameter δ (see above) approaches zero, quantitative comparisons of the inf-sup constant with existing methods from the literature, and the greedy convergence in the parametric case. We report on time-dependent and time-independent test cases. The source code to reproduce all results is provided in [24].

3.5.1 Non-parametric cases

Convergence rates for problems with different smoothness

As indicated in subsection 3.2.1, we can show the convergence of the proposed approximation for appropriate test spaces $\mathcal{Y}_{\mathrm{t}}^{\delta}$, but did not derive theoretical rates of convergence

Table 3.1: 1D: L^2-error and convergence rate as $h \to 0$ for linear and quadratic FE spaces.

	Linear FE		Quadratic FE	
$1/h$	L^2-error	rate	L^2-error	rate
4	0.03311	—	0.00247	—
8	0.01664	0.99274	0.00062	1.98932
16	0.00833	0.99817	0.00016	1.99729
32	0.00417	0.99954	3.896e-05	1.99932
64	0.00208	0.99989	9.741e-06	1.99983
128	0.00104	0.99997	2.435e-06	1.99996
256	0.00052	0.99999	6.088e-07	1.99999

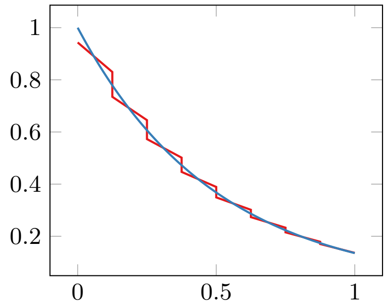

Figure 3.3: 1D: L^2-approximation vs. exact solution for linear FE space with $h = 1/8$.

in this work. Therefore, in this subsection we investigate the rate of convergence in numerical experiments. In all test cases we use as test space $\mathcal{Y}_{\mathrm{t}}^{\delta}$ a continuous FE space on a uniform hexahedral grid. Since we want to investigate here the best possible convergence rates, we choose test cases where the trial space restrictions due to tensor product spaces described in subsection 3.2.2 do *not* lead to additional errors. These cases will then afterwards be compared to cases where the restrictions indeed *do* lead to additional errors in subsection 3.5.1.

We start with the one-dimensional problem introduced in Example 3.2.1 and set $\Omega = (0,1), b_{\mathrm{t}}(x) \equiv 1, c(x) \equiv 2$ with boundary value $u(0) = 1$. We compute approximate solutions for linear FE spaces $\mathcal{Y}_{\mathrm{t}}^{h}$ (recall Figure 3.2 for the corresponding basis functions and see Figure 3.3 for an illustration of the solution) as well as quadratic FE spaces. We observe an (optimal) convergence rate of 1 for the linear case and 2 for the quadratic case (see Table 3.1).

Next, we consider $\Omega = (0,1)^2$, and choose $\mathbf{b} \equiv (\cos 30°, \sin 30°)^T$, $c \equiv 0, f \equiv 0$ and compare boundary values with different smoothness. In detail, we solve

$$\mathbf{b} \cdot \nabla u = 0 \quad \text{in } \Omega, \qquad u = g^i \quad \text{on } \Gamma_- = (\{0\} \times (0,1)) \cup ((0,1) \times \{0\}), \quad i = 1,2,3,$$

for the boundary values

$$g^1 \in C^1(\Gamma_-), \quad g^1(x,0) \equiv 1, \quad g^1(0,y) = \begin{cases} 31.25y^3 - 18.75y^2 + 1, & y \leq 0.4 \\ 0, & y > 0.4, \end{cases} \tag{3.43}$$

$$g^2 \in C^0(\Gamma_-), \quad g^2(x,0) \equiv 1, \quad g^2(0,y) = \begin{cases} 1, & y < 0.2 \\ 2 - 5y & 0.2 \leq y < 0.4 \\ 0, & 0.4 \leq y, \end{cases} \tag{3.44}$$

$$g^3 \in L^2(\Gamma_-), \quad g^3(x,0) \equiv 1, \quad g^3(0,y) = \begin{cases} 1, & y < 0.25 \\ 0, & 0.25 \leq y. \end{cases} \tag{3.45}$$

We use second order FEs on a uniform rectangular mesh with $n_h = h^{-1}$ cells in both dimensions, i.e., $\delta = (h,h)$. As already mentioned above, the data is chosen such that for all boundary conditions it holds that $u(1,1) = 0$ for the exact solution, so that we do not observe problems from the nonphysical restriction of the trial space. We observe a convergence of order about 1.65 for the differentiable case $g = g^1$, an order of 1 for

Table 3.2: L^2-errors and convergence rates for two-dimensional problem with boundary values (3.43), (3.44), and (3.45).

	$g = g^1 \in C^1(\Gamma_-)$		$g = g^2 \in C^0(\Gamma_-)$		$g = g^3 \in L^2(\Gamma_-)$	
$1/h$	L^2-error	rate	L^2-error	rate	L^2-error	rate
16	0.00768	—	0.01974	—	0.10630	—
32	0.00247	1.63387	0.00973	1.02096	0.08484	0.32533
64	0.00079	1.65196	0.00493	0.98128	0.06764	0.32683
128	0.00025	1.65937	0.00248	0.99302	0.05386	0.32862
256	7.872e-05	1.66280	0.00124	0.99476	0.04285	0.33009
512	2.483e-05	1.66452	0.00062	0.99636	0.03406	0.33120

Table 3.3: L^2-error and convergence rate for $\mathbf{b} = (1 - y, x)^T$ and $g = g^4$.

$1/h$	L^2-error	Rate
4	0.09317	—
8	0.03329	1.48458
16	0.01124	1.56702
32	0.00366	1.61950
64	0.00117	1.64276
128	0.00037	1.65386

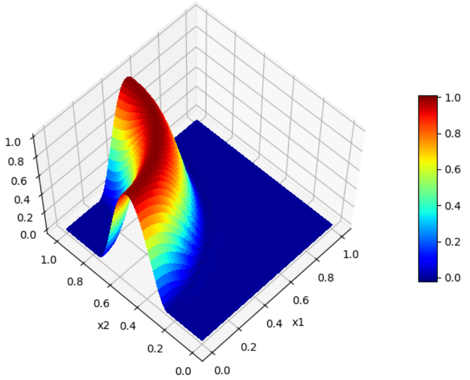

Figure 3.4: Approximate solution for $\mathbf{b} = (1 - y, x)$, $g = g^4$ and $h = 1/32$.

the continuous case $g = g^2$, and an order of about $1/3$ for the discontinuous boundary $g = g^3$ (see Table 3.2).

To assess the effect of a nonconstant transport direction on the convergence rate we use $\mathbf{b}(x, y) = (1 - y, x)^T$, which has an Ω-filling flow with $T = \frac{\pi}{2}$, $c \equiv 0$, $f \equiv 0$, and a C^1-boundary value $g^4 \in C^1(\Gamma_-)$ as

$$g^4(x,0) = 0, \quad g^4(0,y) = \begin{cases} 256y^4 - 512y^3 + 352y^2 - 96y + 9, & 0.25 \le x \le 0.75, \\ 0, & \text{else.} \end{cases}$$

We observe a convergence behavior even slightly better than that for the case of constant $\mathbf{b}$ with a C^1-boundary function; see Table 3.3. The curved transport is resolved without artifacts; see Figure 3.4.

Influence of restrictions due to tensor product spaces

So far we have investigated the convergence of discrete solutions for cases where the nonphysical boundary restrictions described in subsection 3.2.2 do *not* lead to problems. Here we want to compare these results to similar test cases where the restriction indeed *is* unphysical, i.e., for the exact solution we have $u \neq 0$ at the relevant outflow boundary part. We again choose $\Omega = (0,1)^2$, $\mathbf{b} \equiv (\cos 30°, \sin 30°)^T$, $c \equiv 0$, and $f \equiv 0$. We first consider a constant boundary value $\tilde{g} \equiv 1$, leading to $u \equiv 1$, where the impact of the unphysical restriction can be observed best, since the shifted version $g \equiv 0$ leading to $u \equiv 0$ would of course have no discretization error at all. We compare this to shifted

Table 3.4: L^2-errors and convergence rates for two-dimensional problem with different boundary conditions and unphysical restrictions of the trial space.

	$g \equiv 1$		$g = g^1 - 1 \in C^1(\Gamma_-)$		$g = g^2 - 1 \in C^0(\Gamma_-)$		$g = g^3 - 1 \in L^2(\Gamma_-)$	
$1/h$	L^2-error	Rate	L^2-error	Rate	L^2-error	Rate	L^2-error	Rate
16	0.01280	—	0.01479	—	0.02627	—	0.10618	—
32	0.00676	0.92191	0.00691	1.09798	0.01281	1.03615	0.08515	0.31838
64	0.00355	0.92883	0.00349	0.98507	0.00616	1.05729	0.06773	0.33028
128	0.00186	0.93469	0.00183	0.92944	0.00292	1.07500	0.05389	0.32963
256	0.00097	0.93973	0.00097	0.92081	0.00149	0.97073	0.04286	0.33058
512	0.00050	0.94411	0.00050	0.94099	0.00081	0.88878	0.03406	0.33141

versions of the boundary values considered in subsection 3.5.1, i.e., $\tilde{g}^i = g^i - 1, i = 1, 2, 3$ for g^1, g^2, and g^3 defined in (3.43)–(3.45).

In the constant case $g \equiv 1$ we have a convergence of order ≈ 1 (see Table 3.4). Comparing Tables 3.2 and 3.4, we see that indeed the restriction leads to an additional error that converges with order 1: While the problem for the C^1-boundary value g^1 converges with an order of about 1.65, the shifted problem for $\tilde{g}^1 = g^1 - 1$ converges only with an order of ≈ 1. For the less smooth boundaries $\tilde{g}^2 \in C^0(\Gamma_-)$ and $\tilde{g}^3 \in L^2(\Gamma_-)$ we see that the convergence order stays the same, and thus the full error is not dominated by the restriction artifacts. All in all, for the present test cases, the restriction due to the tensor product structure limits the convergence rate to 1, but does not deteriorate smaller convergence orders for less smooth problems, such that for these problems the additional error is negligible. Recall that we are primarily interested in such nonsmooth solutions in $L^2(\Omega)$.

Next, we investigate the approach proposed in subsection 3.2.2 to use an additional layer for the computational domain. In detail, we extend the data functions onto the larger domain $\Omega(\alpha)$ defined in (3.24), solve the problem for the discrete solution $u^\delta_{\Omega(\alpha)} \in L^2(\Omega(\alpha))$ on this extended problem, and then define the restriction $u^\delta_{\Omega(\alpha)}|_\Omega \in L^2(\Omega)$ as the discrete solution to the original problem.

We consider constant boundary values $g \equiv 1$. For each discrete space $\mathcal{Y}^\delta_{\mathrm{t}}, \delta = (h, h)$, we compare values of $\alpha = mh$, $m = 1, \ldots, 5$, i.e., we extend the domain by 1 to 5 layers of grid cells of the original size. The L^2- and L^∞-errors of these solutions and the respective solutions computed on the original domain Ω are shown in Figure 3.5. We see that using extended domains for the computation reduces the L^2-errors: A first layer of grid cells has the most significant effect, but also larger extensions further reduce the errors. Since the difference is larger for coarser meshes, the L^2-rates are slightly lower than those for the original solution, which improves, however, for finer mesh sizes. We obtained similar results for the boundary values $g = g^1 - 1$. Moreover, the extended domain approach has a positive impact on the L^∞-error of the solution and thus on the "optical quality": While for the computations on Ω we automatically have an L^∞-error of 1 for all mesh sizes, the error is reduced to values between about 0.16 for $\alpha = h$ and 0.05 for $\alpha = 5h$; also the L^∞-error on the extended domains seems to be relatively independent of the mesh size (see Figure 3.5). A comparison of the solution computed on Ω and $\Omega(h)$ is provided in Figure 3.6.

We conclude from these experiments that for the current test case the use of an extended domain slightly reduces the L^2-error while maintaining comparable convergence

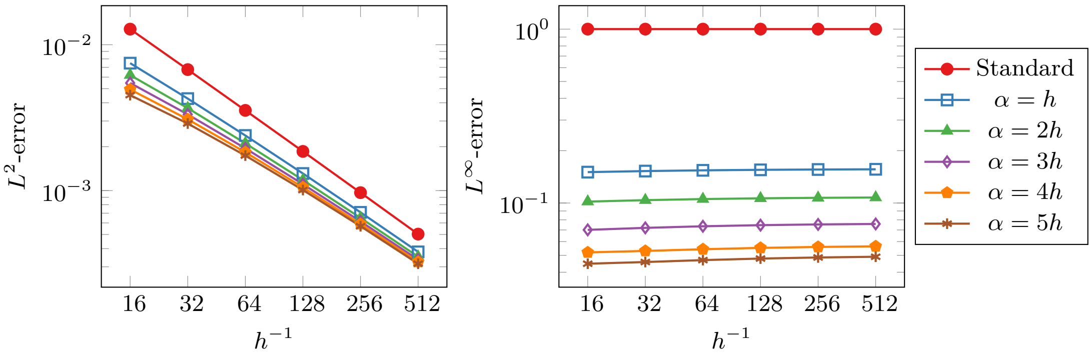

Figure 3.5: 2D, $g \equiv 1$, L^2-errors (left) and L^∞-errors (right) for solutions computed on the standard domain $\Omega = \Omega(0)$ and on extended domains $\Omega(\alpha)$, $\alpha = mh$, $m = 1, \dots, 5$ for different mesh sizes.

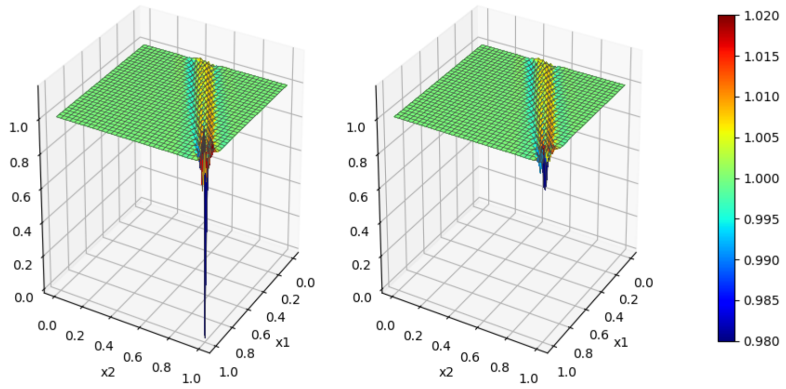

Figure 3.6: 2D, numerical approximation for $h = 1/32$, $g \equiv 1$. Left: Standard domain Ω. Right: $u^\delta|_\Omega$ solved on extended domain $\Omega(h)$

rates and considerably reduces the L^∞-error at the boundary. Hence, at the expense of (moderate) additional computational cost a better approximation of the solution on the outflow boundary can be achieved.

Assessment of postprocessing procedure

We next compare the approximation of discontinuities of a standard solution $u^\delta \in \mathcal{X}_t^\delta$ to the postprocessed solution $\tilde{u}^\delta$ described in subsection 3.2.3. To this end, we again consider the example in subsection 3.5.1 with boundary value $g^3 \in L^2(\Gamma_-)$ that is piecewise constant with a discontinuity. Note that the choice of a constant advection $\mathbf{b}$ and no reaction simplifies the postprocessing procedure, such that the postprocessed solution $\tilde{u}^\delta$ directly is the L^2-orthogonal projection of u^δ onto the discontinuous first order FE space. Comparing the errors of u^δ and $\tilde{u}^\delta$ (see Tables 3.2 and 3.5), we see that the errors for the postprocessed solutions are about 8% smaller than those for the standard solutions, while the order of convergence stays the same. Figure 3.7 shows that the postprocessing removes the severe overshoots of the standard solution at the jump discontinuity. We also note that the postprocessing is computationally inexpensive, since it is only based upon local multiplications of an element projection matrix for each grid cell. A comparison of the computational costs will be given in in the next paragraph.

Table 3.5: L^2-error and convergence rate for postprocessed solution $\tilde{u}^\delta$ for boundary g^3.

$1/h$	L^2-error	Rate
16	0.09769	—
32	0.07765	0.33128
64	0.06179	0.32946
128	0.04917	0.32965
256	0.03911	0.33042
512	0.03108	0.33123

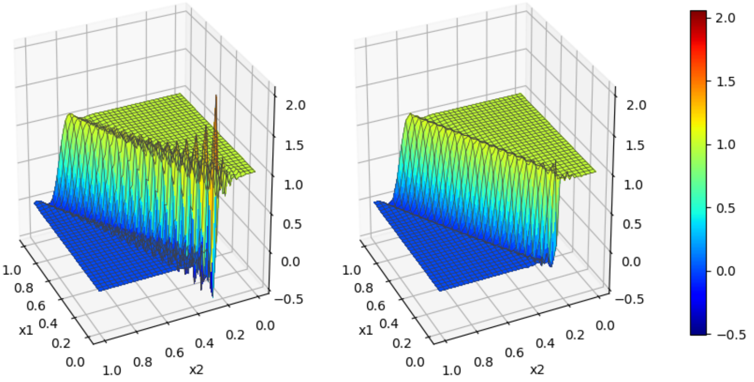

Figure 3.7: Standard solution u^δ (left) and postprocessed solution $\tilde{u}^\delta$ (right) for boundary g^3 and $h = 1/32$.

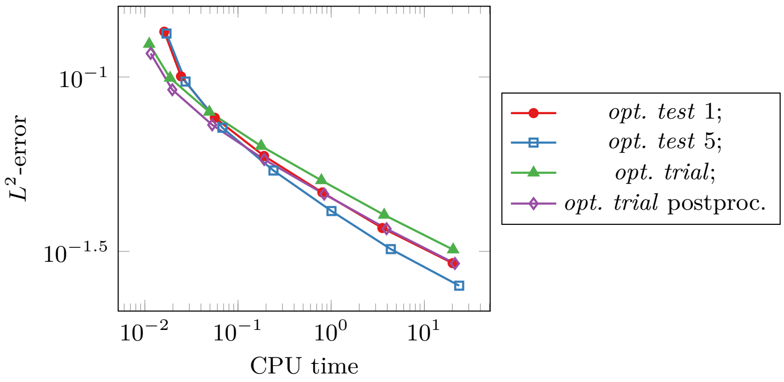

Figure 3.8: L^2-errors versus CPU-times for the *optimal test* method with 1 iteration (*opt. test* 1) and 5 iterations (*opt. test* 5) of the Uzawa algorithm, and for the *optimal trial* method in standard (*opt. trial*) and postprocessed (*opt. trial* postproc.) form.

Comparison of the *optimal trial* method and the *optimal test* method

As discussed in subsection 3.2.4, the approaches proposed in subsection 3.2.1 (*optimal trial* method) and in [43] (*optimal test* method) are closely related. To compare the results for both methods, we use the same test case as in [43]; i.e., we set $\Omega = (0,1)^2$, $\mathbf{b} \equiv (\cos 22.5°, \sin 22.5°)^T$, $c \equiv 0$, and $f \equiv 0$. For the boundary condition we again have the discontinuous boundary value $g = g^3$ defined in (3.45).

As the equations for $w^\delta \in \mathcal{Y}_t^\delta$ in (3.22) (*optimal trial*) and for $\hat{r}^{k,\delta} \in \mathcal{Z}_t^\delta$ in (3.28) (*optimal test*) are based on the same bilinear form, we choose the spaces such that $\mathcal{Y}_t^\delta = \mathcal{Z}_t^\delta$, which means that the same matrix has to be assembled for both methods. More precisely, we choose for the *optimal trial* method the same spaces as in the experiments above, i.e., $\mathcal{Y}_t^\delta$ is the space of continuous FEs of second order on a rectangular grid with mesh size $\delta = (h,h)$. Consistent with that, we choose – as proposed in [43] – for $\widehat{\mathcal{X}}_t^\delta$ the space of discontinuous bilinear FEs on a rectangular grid with mesh size $(2h, 2h)$, and $\mathcal{Z}_t^\delta = \mathcal{Y}_t^\delta$, such that here the grid for the test search space results from one uniform refinement of the grid of the trial space.

We first compare the relation of L^2-errors and CPU times for both methods. For the

Table 3.6: Inf-sup constants for the *optimal test* method and the 2D problem

1/(2h)	Inf-sup
4	0.74521
8	0.66426
16	0.55840
32	0.45422
64	0.36029
128	0.28273
256	0.21901

Table 3.7: Inf-sup constants for the *optimal test* method and the 3D problem

1/(2h)	Inf-sup
4	0.64800
8	0.60160
16	0.48294
32	0.38015

solution of the linear systems, we always use sparse LU factorization and subsequent forward and back substitution implemented in UMFPACK. Figure 3.8 shows the respective CPU-error plots for the *optimal test* method using 1 iteration and 5 iterations of the Uzawa algorithm (as proposed in [43]) and for the standard solution of the *optimal trial* method as well as the postprocessed solution described in subsection 3.2.3. We observe similar decay rates of the errors w.r.t. the CPU times for both methods. For the chosen linear solver, the *optimal test* methods with 5 iterations performs best, which is mainly due to the fact that assembly of the matrices and LU factorization dominate the computational costs. Therefore, the costs for 5 Uzawa iterations are only slightly higher than for e.g. only 1 Uzawa iteration, while the errors are reduced significantly. If we use iterative methods, e.g. the CG method , instead, the results depend on the used preconditioner: If the computation of the preconditioner dominates, the results are similar to the results using LU decomposition. In contrast, if the iterative solver takes as much time as or more time than the preconditioner, then the *optimal test* solutions using 5 Uzawa iterations would take considerably more time compared to the other solutions and we speculate that the postprocessed *optimal trial* solution might perform fairly equally to the *optimal test* solutions. However, a comparison of different preconditioners is outside the scope of this work.

Finally, we compare the inf-sup constants of both methods. While for the *optimal trial* method we automatically have an inf-sup constant of 1, this is not the case for the *optimal test* method. Since here not the truly optimal test space $(B^*)^{-1}\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta}$, but the projection onto the test search space $P_{\mathcal{Z}^\delta}((B^*)^{-1}\widehat{\mathcal{X}}_{\mathrm{t}}^{\delta})$ is used for the discrete test space, the inf-sup constant for the discrete problem as well as for the corresponding saddle-point problem on which the Uzawa iteration is based is suboptimal. Tables 3.6 and 3.7 show the inf-sup constants for the considered two-dimensional problem, i.e., $\Omega = (0,1)^2$, $\mathbf{b} = (\cos 22.5°, \sin 22.5°)^T$, $c \equiv 0$, and the corresponding time-dependent problem, i.e., a three-dimensional problem with $\Omega = (0,1)^3$ and $\mathbf{b} = (1, \cos 22.5°, \sin 22.5°)^T$, respectively. We clearly see that the inf-sup constants decrease with smaller mesh sizes; in both cases they decay roughly with an order of $h^{1/3}$.

3.5.2 Parametric cases: The reduced basis method

To examine our method in the parametric setting, we consider three different test cases. For all cases, we choose $\Omega = (0,1)^2$ and a parametrized constant transport direction $\mathbf{b}_\mu \in \mathbb{R}^2, \mu \in \mathcal{P}$, such that $\Gamma_- = (\{0\} \times (0,1)) \cup ((0,1) \times \{0\})$ for all $\mu \in \mathcal{P}$, as well as

Table 3.8: Data for parametric test cases.

	test case 1 (see [107])	test case 2 (cf. [45])	test case 3 (cf. [45])
$\mathbf{b}_\mu$	$(\mu, 1)^T$	$(\cos\mu, \sin\mu)^T$	$(\cos\mu, \sin\mu)^T$
$\mathcal{P}$	$[0.01, 1]$	$[0.2, \frac{\pi}{2} - 0.2]$	$[0.2, \frac{\pi}{2} - 0.2]$
c	$\equiv 0$	$\equiv 1$	$\equiv 1$
f	$\equiv 0$	$\equiv 1$	$\begin{cases} 0.5, & x < y \\ 1, & x \geq y \end{cases}$
g	$\begin{cases} 1, & x = 0 \\ 0, & y = 0 \end{cases}$	$\equiv 0$	$\begin{cases} 1 - y, & x \leq 0.5 \\ 0, & x \geq 0.5 \end{cases}$

parameter-independent reaction, source, and boundary data; see Table 3.8. Again, we want to solve for all $\mu \in \mathcal{P}$

$$\mathbf{b}_\mu \cdot \nabla u + cu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \Gamma_-.$$

For all test cases, we choose a training set of 500 equidistant parameter values distributed over $\mathcal{P}$ and set $\varepsilon = 10^{-4}$. We then generate reduced models with Algorithm 1 for different mesh sizes. The maximum model errors $\|u^N(\mu) - u^\delta(\mu)\|_{L^2(\Omega)}$ on an additional test set of 500 uniformly distributed random parameter values are shown in Figure 3.9.

Since we did not derive theoretical convergence results for the greedy algorithm, we investigate the convergence behavior numerically. To that end, we first consider a test case where the best possible convergence rate of linear approximations is known: In [107], it is shown that the Kolmogorov N-width of the solution set of test case 1 decays with an order of $N^{-1/2}$. In the corresponding results of our greedy algorithm, we indeed observe the same (and thus optimal) convergence behavior; see Figure 3.9.

In test case 2 we choose constant reaction and source terms that lead to more regular solutions. Here, the greedy algorithm shows a faster convergence of order about $N^{-3/2}$. With discontinuous source and boundary data in test case 3 we finally observe an order of roughly N^{-1}.

As described in subsection 3.3.6, our approach is closely related to the *double greedy* algorithm framework developed in [45]. To realize a fair comparison with our approach, we implemented a "strong" *double greedy* algorithm using the model error instead of a surrogate in [45, Algorithm 4] (*update-approximation*, see also page 49). For the full solutions we use the discretization of the *optimal test* method described in subsections 3.2.4 and 3.5.1. We then run the "strong" variant of the *double greedy* algorithm [45, Algorithm 5] for test case 3 on a training set of 500 equidistant parameter values distributed over $\mathcal{P}$ and with tolerance $\varepsilon = 0.01$ comparing different thresholds β_{min} for the inf-sup stability of the reduced spaces[12].

The resulting maximum model errors for 500 test parameter values are shown in Figure 3.10. For the smaller stability thresholds of 0.3 and 0.6 we observe slight in-

[12] In [45] it is proposed to use $\beta_{min} := \zeta\beta_\delta$, where $0 < \beta_\delta \leq 1$ is a lower bound of the discrete inf-sup constants of the full discretizations for all $\mu \in \mathcal{P}$ and some $0 < \zeta < 1$, such that the desired threshold is guaranteed to be achievable for all reduced spaces. Here, we simply compare different values of $\beta_{min} < 1$ without computing β_δ.

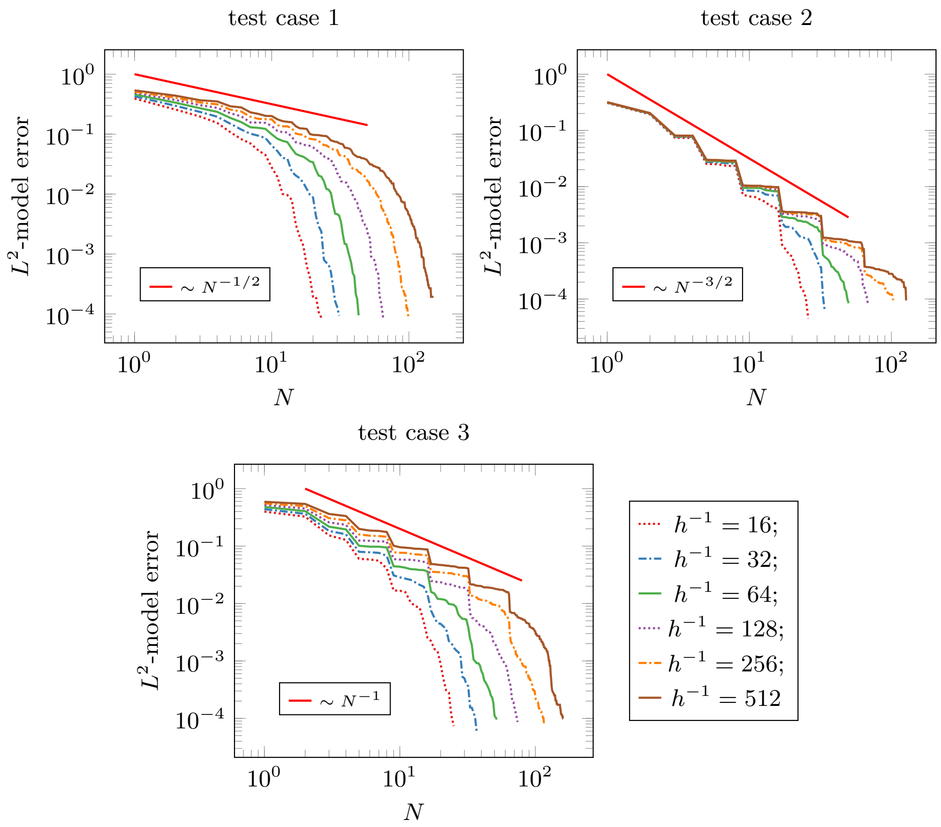

Figure 3.9: Maximum errors of 500 test parameter values for different model orders, mesh sizes, and test cases 1, 2, and 3.

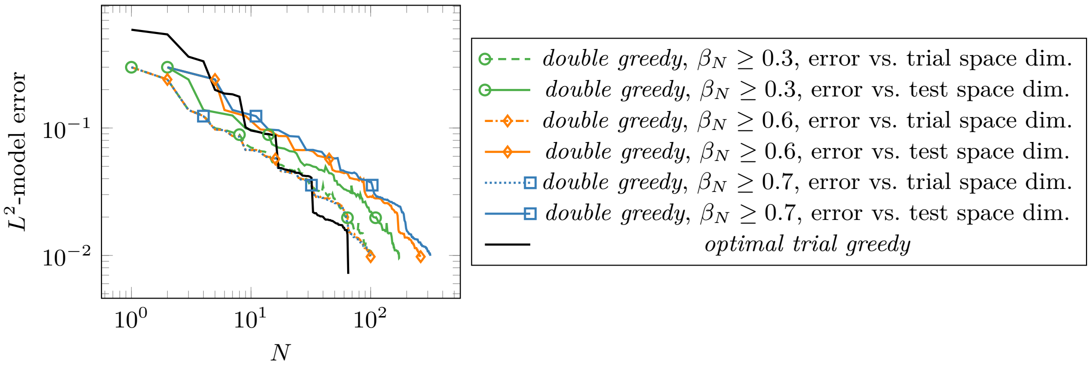

Figure 3.10: test case 3, $h^{-1} = 512$. Maximum errors of 500 test parameter values for reduced models from Algorithm 1 (*optimal trial greedy*) and the strong *double greedy* algorithm with different lower inf-sup bounds, plots of maximum error versus trial space dimension and test space dimension, respectively.

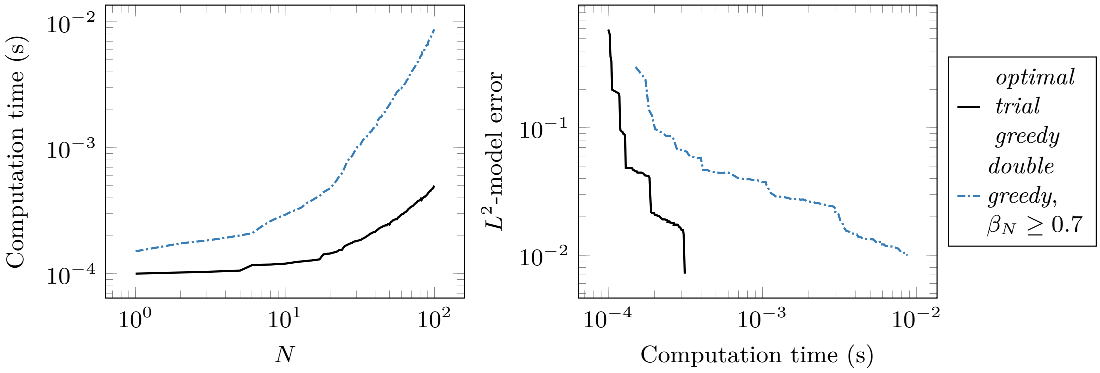

Figure 3.11: test case 3, $h^{-1} = 512$. Comparison of online computation times (median of 5000 runs) for reduced models from *optimal trial greedy* Algorithm and strong *double greedy* algorithm with $\beta_N \geq 0.7$. Left: Computation time versus trial space dimension, right: maximum model error versus computation time.

stabilities while for a threshold of 0.7 the maximum model errors are decreasing for increasing model orders. Comparing the approximation properties of the trial spaces of the *double greedy* and *optimal trial greedy* method, we see that for model orders up to 32 the *double greedy* trial spaces lead to smaller errors than the *optimal trial* spaces of same dimension, while for larger model orders the *optimal trial* reduced spaces perform better.

Since, unlike the new method, for the *double greedy* method the test spaces are significantly larger than the trial spaces (for test case 3, $\beta_N \geq 0.7$, approximately by a factor of 3), the test space dimensions are essential for the online complexity of the reduced saddle point problems. In Figure 3.11 online computation times for both methods are shown, where we use for the *double greedy* solutions a reformulation of the saddle point problem where the inversion of a test space sized matrix dominates the costs[13]. We clearly see that the *optimal trial* reduced models outperform the *double greedy* models both when comparing the same trial space dimensions and the same model errors[14].

These results show that for the rather challenging test case 3 the *optimal trial* method leads to comparable, and for larger model orders even better, approximation properties for the same dimension of the trial spaces and to faster online computation times than the *double greedy* method. We note that for smoother cases, e.g. test case 2, the *optimal trial* models show the same, but not better, convergence order as the *double greedy* models.

Finally, to test the hierarchical error estimator described in subsection 3.3.5, we use test case 2 with mesh size $\delta = (h, h)$, $h^{-1} = 512$. For the reduced space Y^N, we choose a greedy basis with tolerance $\varepsilon = 10^{-2}$, which here corresponds to $N = 13$. For the error estimator reference space $Y^M \supset Y^N$, we compare spaces with tolerances $\varepsilon = 10^{-2.5}, 10^{-3}, 10^{-3.5}$, and 10^{-4}, leading to $M = 31, 62, 91$, and 127, respectively. The

[13] Directly solving the larger linear system of size (trial space dim.)+(test space dim.) corresponding to the saddle point formulation leads to comparable results.

[14] Note, however, that as usual online computation times contain only the computation of the coefficients of the reduced solutions in the respective reduced basis. If an assembly of the full-dimensional solution vector is needed, this dominates the costs and is clearly faster for the *double greedy* models, since for the *optimal trial* method the separate parts of the affine decomposition of the trial space have to be assembled, and the trial space vector is usually larger.

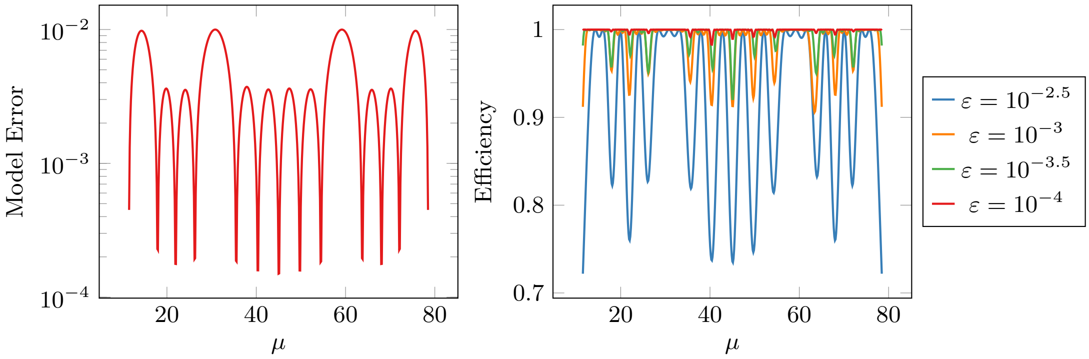

Figure 3.12: test case 2, $h^{-1} = 512$. Model errors $\|u^N - u^\delta\|_{L^2(\Omega)}$ for all test parameter values (left) and ratios of estimated and real model errors $\|u^N - u^M\|_{L^2(\Omega)}/\|u^N - u^\delta\|_{L^2(\Omega)}$ (right).

results in Figure 3.12 show the quantitative good performance. Note that the values of M are significantly larger than reported for the hierarchical error estimator in [79] which is due to the fact that M is determined differently and transport problems are not considered there.

4 The kinetic Fokker-Planck equation

In this chapter, we develop a stable and efficient Petrov-Galerkin approximation scheme for a kinetic Fokker-Planck equation of the type

$$\partial_t u((t,x),v) + v \cdot \nabla_x u((t,x),v) = \Delta_v \left(\tfrac{u((t,x),v)}{q(x,v)}\right) \quad \text{in } \Omega = I_t \times \Omega_x \times \Omega_v, \tag{4.1}$$

which is a prototype for the mesoscopic glioma tumor equation (2.3) introduced in section 2.1. Equation (4.1) describes the density of glioma tumor cells in phase space dependent on time $t \in I_t$, position $x \in \Omega_x \subset \mathbb{R}^d$, $d \in \{2,3\}$, and velocity $v \in \Omega_v = S^{d-1}$. As for the transport equation in chapter 3, we aim to develop a stable discretization based on a suitable variational formulation of the equation.

After a more detailed description of the considered Fokker-Planck equation in section 4.1, we establish the necessary function spaces for the variational formulation in section 4.2. We use Bochner-type spaces mapping the combined space-time domain $\Omega_{t,x} = I_t \times \Omega_x$ to a Sobolev space defined on the velocity domain Ω_v similar to spaces defined in [3, 34]. After introducing the spaces, we show necessary density and trace properties.

We then derive the variational formulation and prove the existence and uniqueness results in section 4.3. To that end, we take the viewpoint that the Fokker-Planck equation could be interpreted as a "generalization" of a parabolic equation with a $(d+1)$-dimensional kinetic transport operator $\partial_t + v \cdot \nabla_x$ instead of a one-dimensional time derivative ∂_t. Therefore, we analyze the well-posedness of the variational formulation for (4.1) by combining respective approaches developed for parabolic equations [65,119,129] and for transport equations [28,43,52]. We show existence of a weak solution by verifying the dual inf-sup condition. To that end, similarly to [65, 119] specific function pairs in the trial and test spaces are constructed: We associate a test space function p to a trial space function roughly defined as $w_p = p - (\Delta_v)^{-1}(\partial_t p + v \cdot \nabla_x p)$. Then the bilinear form evaluated in w_p and p can be bounded from below by the respective norms of w_p and p, which leads to a lower bound for the dual inf-sup constant. Under an additional assumption on the global traces of certain considered functions, we also show uniqueness of the solution and have a stability estimate dependent on the inf-sup constant which is similar to the respective estimates for parabolic equations.

In section 4.4, we then introduce the discrete scheme. We reuse the strategy from chapter 3 to design a Petrov-Galerkin discretization with problem-specific trial spaces ensuring stability: We first choose an arbitrary discrete test space $\mathcal{Y}^\delta_{\mathrm{fp}}$ and then define the discrete trial space roughly as $\mathcal{X}^\delta_{\mathrm{fp}} = \mathcal{Y}^\delta_{\mathrm{fp}} + (\Delta_v)^{-1}(\partial_t + v \cdot \nabla_x)\mathcal{Y}^\delta_{\mathrm{fp}}$. The spaces thus consist of pairs (w^δ_p, p^δ) that are the discrete counterparts of the pairs (w_p, p) used in the proof for the lower bound of the dual inf-sup constant. Therefore, a discrete inf-sup estimate follows analogously to the inf-sup estimate of the variational formulation. As the computation of the trial space involves $(\Delta_v)^{-1}$, i.e., the solution of elliptic problems in the velocity domain, we discuss how the trial space can be efficiently computed for certain discrete spaces and a separable form of the data functions.

We conclude the chapter in section 4.5 with a numerical example, where we investigate the sharpness of the inf-sup estimate and the efficiency of the scheme.

The contents of this chapter have been published as a preprint in [27].

4.1 Problem setting

We consider a simplified version of (2.3), the Fokker-Planck mesoscopic glioma model developed in [90, sect. 2.4.2], for details see section 2.1.

Let $\Omega_x \subset \mathbb{R}^d$, $d \in \{2,3\}$ be the spatial domain with piecewise C^1 boundary that is globally Lipschitz and let $I_t := (0,T)$ be the time interval. Moreover, let the velocity domain be the $(d-1)$-dimensional unit sphere $\Omega_v := S^{d-1}$, which corresponds to the assumption of particles with constant speed but varying direction. As we will often treat space and time variables simultaneously, we denote by $\Omega_{t,x} := I_t \times \Omega_x$ the space-time domain. The full domain is defined as $\Omega := \Omega_{t,x} \times \Omega_v$.

Boundary conditions have to be prescribed at the inflow part of $\partial\Omega$. To that end, we first define the spatial in- and outflow domains $\Gamma^x_\pm(v) := \{x \in \partial\Omega_x : \mathbf{n}(x) \cdot v \gtrless 0\} \subset \partial\Omega_x$, where $\mathbf{n}(x)$ is the unit outer normal to $\partial\Omega_x$ at x. The full in- and outflow domains Γ_- and Γ_+ are then defined as

$$\Gamma_\pm := \{((t,x),v) \in \partial\Omega_{t,x} \times \Omega_v \,:\, \left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \mathbf{n}(t,x) \gtrless 0\} \subset \partial\Omega,$$

where $\mathbf{n}(t,x)$ is the unit outer normal to $\partial\Omega_{t,x}$ at (t,x). $\Gamma_\pm$ thus contain both the temporal and the spatial boundaries, i.e., Γ_- contains the "initial boundary" and the (v-dependent) spatial inflow boundary whereas Γ_+ contains the final time boundary and the spatial outflow boundary.

The strong form of the Fokker-Planck equation then reads

$$\begin{aligned} \partial_t u((t,x),v) + v \cdot \nabla_x u((t,x),v) &= \Delta_v \left(\tfrac{u((t,x),v)}{q(x,v)}\right) && \text{in } \Omega, \\ u((t,x),v) &= g((t,x),v) && \text{on } \Gamma_-. \end{aligned} \tag{4.2}$$

Here, Δ_v is the Laplace-Beltrami operator on the unit sphere $\Omega_v = S^{d-1}$ (see Definition 2.3.13 and Example 2.3.14). The function $q : \Omega_x \times \Omega_v \to \mathbb{R}$ is the tissue fiber orientation distribution satisfying $q(x,v) \geq \alpha_q > 0$ for all $(x,v) \in \Omega_x \times \Omega_v$ and $\int_{\Omega_v} q(x,v)\,\mathrm{d}v = 1$ for all $x \in \Omega_x$ (see section 2.1), and $g : \Gamma_- \to \mathbb{R}$ is the inflow boundary condition that contains the initial condition $g|_{\{t=0\}}$ as well as the spatial inflow boundary condition $g|_{\Gamma^x_-(v)}, v \in \Omega_v$.

In section 4.3, we develop a variational formulation for this equation, where we allow for a more general differential operator on Ω_v, and give specific conditions on q and g leading to well-posedness.

4.2 Function spaces

To develop a variational formulation for (4.2) we first introduce the necessary function spaces. Since we aim for a full-dimensional (i.e., space-time-velocity) formulation, we use Bochner spaces mapping the space-time domain $\Omega_{t,x}$ to a space of functions on Ω_v.

We start with the function space for the velocity variable: Since the equation contains a Laplace-Beltrami operator on the velocity domain $\Omega_v = S^{d-1}$, we define $V :=$

$H^1(\Omega_v) \subset L^2(\Omega_v)$ as the Sobolev space of weakly differentiable functions on the surface $\Omega_v = S^{d-1}$ with norm $\|\phi\|_V^2 = \|\phi\|^2_{L^2(\Omega_v)} + \|\nabla_v \phi\|^2_{L^2(\Omega_v)}$, see Definition 2.3.16. We denote the dual space of V by $V' := H^{-1}(\Omega_v)$. V is a dense subspace of $L^2(\Omega_v)$ and we will make use of the Gelfand triple $V \hookrightarrow L^2(\Omega_v) \hookrightarrow V'$, where we denote the dual pairing by $\langle \cdot, \cdot \rangle_{V',V}$.

On the full domain, we define the Bochner space $L^2(\Omega_{t,x}; V)$ with norm

$$\|w\|^2_{L^2(\Omega_{t,x};V)} = \int_{\Omega_{t,x}} \|w(t,x)\|_V^2 \, \mathrm{d}(t,x) \tag{4.3}$$

for functions without space or time derivatives. To incorporate the kinetic space-time transport operator, we define (using from now on $\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p := \partial_t p + v \cdot \nabla_x p$)

$$H^1_{\mathrm{fp}}(\Omega) := \{p \in L^2(\Omega_{t,x}; V) \,:\, \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p \in L^2(\Omega_{t,x}; V')\}, \tag{4.4}$$

with norm

$$\|p\|^2_{H^1_{\mathrm{fp}}(\Omega)} := \|p\|^2_{L^2(\Omega_{t,x};V)} + \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p\|^2_{L^2(\Omega_{t,x};V')}. \tag{4.5}$$

This definition is similar to the spaces used for other variants of the kinetic Fokker-Planck equation e.g. in [3, 10, 34]. We use ideas from [3] to show the following:

Proposition 4.2.1. *The set $C^\infty(\bar{\Omega}_{t,x} \times \Omega_v) \cap H^1_{\mathrm{fp}}(\Omega)$ is dense in $H^1_{\mathrm{fp}}(\Omega)$.*

Proof. The claim is only a slight variant of [3, Prop. 7.1], where the respective density result is shown for the space

$$\tilde{H}^1_{\mathrm{fp}}(\Omega) := \{p \in L^2(\Omega_{t,x}; \tilde{V}) \,:\, \partial_t p - v \cdot \nabla_x p \in L^2(\Omega_{t,x}; \tilde{V}')\}$$

with $\tilde{V} = H^1_\gamma(\mathbb{R}^d)$ being the Sobolev space on $\mathbb{R}^d$ with standard Gaussian measure. The space $\tilde{H}^1_{\mathrm{fp}}(\Omega)$ is used to describe a Fokker-Planck equation similar to (4.2), but on $\tilde{\Omega}_v = \mathbb{R}^d$ and with a reverse sign for the transport term. We will therefore reuse the proofs of [3, Prop. 7.1] (and [3, Prop. 2.2], which treats the time-independent case) and modify only the parts dependent on V and Ω_v.

In step 1 of the proofs it is shown that we can assume without loss of generality that for every $z := (t,x) \in \Omega_{t,x} \subset \mathbb{R}^{d+1}$ and $\varepsilon \in (0,1]$ we have $B((1-\varepsilon)z, \varepsilon) \subset \Omega_{t,x}$, where $B(z,r)$ is the open ball with radius r around z.

Let then $f \in H^1_{\mathrm{fp}}(\Omega)$. As in step 2 of the proofs we take $\zeta \in C_0^\infty(\mathbb{R}^{d+1}, \mathbb{R})$ as a smooth function with compact support in $B(0,1)$ such that $\int_{\mathbb{R}^{d+1}} \zeta = 1$. For each $\varepsilon > 0$ and $z \in \mathbb{R}^{d+1}$ we write

$$\zeta_\varepsilon(z) := \varepsilon^{-(d+1)} \zeta(\varepsilon^{-1} z),$$

and define for $\varepsilon \in (0, \frac{1}{2}]$, $z \in \Omega_{t,x}$, and $v \in \Omega_v$ the mollification

$$f_\varepsilon(z, v) := \int_{\mathbb{R}^{d+1}} f((1-\varepsilon)z + z', v) \zeta_\varepsilon(z') \, \mathrm{d}z',$$

so that we have $f_\varepsilon \in C^\infty(\bar{\Omega}_{t,x}; V)$. We may then show completely analogous to step 2 of the proofs of [3, Prop. 2.2 and 7.1] that f belongs to the closed convex hull of the set $\{f_\varepsilon : \varepsilon \in (0, \frac{1}{2}]\}$ by just changing the spaces of all dual pairings and norms from $\tilde{V} = H^1_\gamma(\mathbb{R}^d)$ to $V = H^1(S^{d-1})$ and from $L^2_\gamma(\mathbb{R}^d)$ to $L^2(S^{d-1})$.

It then remains to be shown that for fixed $\varepsilon \in (0, \frac{1}{2}]$ the function f_ε belongs to $\operatorname{clos}_{\|\cdot\|_{H^1_{\mathrm{fp}}}} C^\infty(\bar{\Omega}_{t,x} \times \Omega_v)$ by approximating f_ε also in the v-variable.

We construct a basis of $V = H^1(\Omega_v)$ that is contained in $C^\infty(\Omega_v)$: Since V as a subspace of $L^2(\Omega_v)$ is separable and $C^\infty(\Omega_v) \subset V$, there exists a dense countable set in $(C^\infty(\Omega_v), \|\cdot\|_V)$, from which we can obtain an orthonormal basis $(\psi_i)_{i\in\mathbb{N}}$ by the Gram-Schmidt algorithm. Since $\operatorname{span}(\psi_i)_{i\in\mathbb{N}}$ is dense in $C^\infty(\Omega_v)$ which is again dense in V (see Theorem 2.3.19), $(\psi_i)_{i\in\mathbb{N}}$ is also an orthonormal basis of V.

For $k \in \mathbb{N}$, we define $f_{\varepsilon,k} : \Omega_{t,x} \times \Omega_v \to \mathbb{R}$ as

$$f_{\varepsilon,k}(z, v) := \sum_{i=1}^{k} (f_\varepsilon(z, \cdot), \psi_i)_V \psi_i(v).$$

Since we have $f_\varepsilon \in C^\infty(\bar{\Omega}_{t,x}; V)$, the map $z \mapsto (f_\varepsilon(z, \cdot), \psi_i)_V$ is in $C^\infty(\bar{\Omega}_{t,x})$. As $\psi_i \in C^\infty(\Omega_v)$ for all $i \in \mathbb{N}$, we have $f_{\varepsilon,k} \in C^\infty(\bar{\Omega}_{t,x} \times \Omega_v)$ for all $k \in \mathbb{N}$.

Next, we compute $\lim_{k\to\infty} \|f_\varepsilon - f_{\varepsilon,k}\|_{L^2(\Omega_{t,x};V)}$. First, fix $z \in \bar{\Omega}_{t,x}$. Since $(\psi_i)_{i\in\mathbb{N}}$ is an orthonormal basis of V we have $f_\varepsilon(z) = \sum_{i=1}^\infty (f_\varepsilon(z), \psi_i)_V \psi_i$ and thus

$$\|f_\varepsilon(z) - f_{\varepsilon,k}(z)\|_V = \left\| \sum_{i=k+1}^{\infty} (f_\varepsilon(z), \psi_i)_V \psi_i \right\|_V = \sum_{i=k+1}^{\infty} (f_\varepsilon(z), \psi_i)_V^2 \xrightarrow{k\to\infty} 0.$$

As this holds for all $z \in \bar{\Omega}_{t,x}$ and $\|f_\varepsilon(z) - f_{\varepsilon,k}(z)\|_V \le 2\|f_\varepsilon(z)\|_V$, we obtain by the dominated convergence theorem that $\lim_{k\to\infty} \|f_\varepsilon - f_{\varepsilon,k}\|_{L^2(\Omega_{t,x};V)} = 0$. To determine $\lim_{k\to\infty} \|\binom{1}{v}\cdot\nabla_z(f_\varepsilon - f_{\varepsilon,k})\|_{L^2(\Omega_{t,x};V')}$, we first consider the partial derivatives separately: Since $f_\varepsilon \in C^\infty(\bar{\Omega}_{t,x}; V)$, all first z-partial derivatives of f_ε lie in $L^2(\Omega_{t,x}; V)$ and we know that

$$\|\partial_{z_j} f_\varepsilon(z) - \partial_{z_j} f_{\varepsilon,k}(z)\|_V = \| \sum_{i=k+1}^{\infty} (\partial_{z_j} f_\varepsilon(z), \psi_i)_V \psi_i \|_V \xrightarrow{k\to\infty} 0$$

for $j = 1, \dots, d+1$, and all $z \in \bar{\Omega}_{t,x}$. Since $|\binom{1}{v}|$ is bounded on $\Omega_v = S^{d-1}$, we thus have

$$\begin{aligned}\|\tbinom{1}{v} \cdot \nabla_z (f_\varepsilon(z) - f_{\varepsilon,k}(z))\|_{L^2(\Omega_v)} &\le \sum_{j=1}^{d+1} \|\tbinom{1}{v}_j\|_{L^\infty(\Omega_v)} \|\partial_{z_j} f_\varepsilon(z) - \partial_{z_j} f_{\varepsilon,k}(z)\|_{L^2(\Omega_v)} \\ &\le \sum_{j=1}^{d+1} \|\tbinom{1}{v}_j\|_{L^\infty(\Omega_v)} \|\partial_{z_j} f_\varepsilon(z) - \partial_{z_j} f_{\varepsilon,k}(z)\|_V \xrightarrow{k\to\infty} 0,\end{aligned}$$

and again by the dominated convergence theorem that

$$\lim_{k\to\infty} \|\tbinom{1}{v} \cdot \nabla_z (f_\varepsilon - f_{\varepsilon,k})\|_{L^2(\Omega_{t,x};V')} \le \lim_{k\to\infty} \|\tbinom{1}{v} \cdot \nabla_z (f_\varepsilon - f_{\varepsilon,k})\|_{L^2(\Omega_{t,x};L^2(\Omega_v))} = 0.$$

Hence, $f_{\varepsilon,k}$ converges to f_ε in $H^1_{\mathrm{fp}}(\Omega)$, which completes the proof of Proposition 4.2.1. □

To discuss the boundary behavior of functions in $H^1_{\mathrm{fp}}(\Omega)$, we introduce weighted L^2-spaces as usually used for transport and kinetic equations (cf. subsections 2.3.1 and 2.3.2 and section 3.1) and also for different versions of the kinetic Fokker-Planck equation [3, 34]. For any $\Gamma \subseteq \partial\Omega$ we introduce $L^2(\Gamma, |(1, v)^T \cdot \mathbf{n}|)$ with norm

$$\|w\|_{L^2(\Gamma,|(1,v)^T\cdot\mathbf{n}|)} := \int_\Gamma w^2 \, |\tbinom{1}{v} \cdot \mathbf{n}| \, \mathrm{d}s. \tag{4.6}$$

Then, we can show first that functions in $H^1_{\mathrm{fp}}(\Omega)$ admit local traces in $\partial\Omega \setminus \Gamma_0$:

Proposition 4.2.2. *For every compact set $K \subset \Gamma_+$ (resp. $K \subset \Gamma_-$), the trace operator $w \mapsto w|_K$ from $C^\infty(\bar{\Omega})$ to $L^2(K, |(1,v)^T \cdot \mathbf{n}|)$ extends to a continuous linear operator on $H^1_{\mathrm{fp}}(\Omega)$.*

For the proof we need to estimate the product of $H^1_{\mathrm{fp}}(\Omega)$ functions with different test functions in the following way:

Lemma 4.2.3. *Let $\phi \in C^1(\bar{\Omega})$. Then, the mapping $f \mapsto \phi f$ is continuous in $H^1_{\mathrm{fp}}(\Omega)$ with the estimate*

$$\|\phi f\|_{H^1_{\mathrm{fp}}(\Omega)} \le C\|\phi\|_{C^1(\Omega)}\|f\|_{H^1_{\mathrm{fp}}(\Omega)}.$$

Proof. We estimate $\|\phi f\|_{H^1_{\mathrm{fp}}(\Omega)}$. Using the definition of the V-norm and the product rule we obtain for the first term[1]

$$\begin{aligned}\|\phi f\|^2_{\mathcal{X}_{\mathrm{fp}}} &= \|\phi f\|^2_{L^2(\Omega)} + \|(\nabla_v \phi) f + \phi \nabla_v f\|^2_{L^2(\Omega)} \\ &\le \|\phi^2\|_{L^\infty(\Omega)}\|f\|^2_{L^2(\Omega)} + 2\||\nabla_v\phi|^2\|_{L^\infty(\Omega)}\|f\|^2_{L^2(\Omega)} + 2\|\phi^2\|_{L^\infty(\Omega)}\|\nabla_v f\|^2_{L^2(\Omega)} \\ &\le 2\left(\|\phi\|^2_{L^\infty(\Omega)} + \|\nabla_v\phi\|^2_{L^\infty(\Omega)}\right)\|f\|^2_{\mathcal{X}_{\mathrm{fp}}}. \end{aligned} \tag{4.7}$$

By using the product rule, the characterization $\langle\cdot,\cdot\rangle_{\mathcal{X}'_{\mathrm{fp}},\mathcal{X}_{\mathrm{fp}}} = (\cdot,\cdot)_{L^2(\Omega)}$, and the continuity of $C^\infty(\Omega)$ in $H^1_{\mathrm{fp}}(\Omega)$ we see that for arbitrary $\psi \in \mathcal{X}_{\mathrm{fp}}$ it holds

$$\begin{aligned}\langle \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}(\phi f), \psi\rangle_{\mathcal{X}'_{\mathrm{fp}},\mathcal{X}_{\mathrm{fp}}} &= \langle \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} f, \phi\psi\rangle_{\mathcal{X}'_{\mathrm{fp}},\mathcal{X}_{\mathrm{fp}}} + (f(\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}\phi), \psi)_{L^2(\Omega)} \\ &\le \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} f\|_{\mathcal{X}'_{\mathrm{fp}}}\|\phi\psi\|_{\mathcal{X}_{\mathrm{fp}}} + \|f(\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}\phi)\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)}. \\ &\overset{(4.7)}{\le} \sqrt{2}\left(\|\phi\|^2_{L^\infty(\Omega)} + \|\nabla_v\phi\|^2_{L^\infty(\Omega)}\right)^{\frac{1}{2}} \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} f\|_{\mathcal{X}'_{\mathrm{fp}}}\|\psi\|_{\mathcal{X}_{\mathrm{fp}}} \\ &\quad + \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}\phi\|_{L^\infty(\Omega)}\|f\|_{L^2(\Omega)}\|\psi\|_{L^2(\Omega)} \\ &\le \sqrt{2}\left(\|\phi\|_{L^\infty(\Omega)} + \|\nabla_v\phi\|_{L^\infty(\Omega)} + \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}\phi\|_{L^\infty(\Omega)}\right)\|f\|_{H^1_{\mathrm{fp}}(\Omega)}\|\psi\|_{\mathcal{X}_{\mathrm{fp}}}.\end{aligned}$$

We thus have

$$\begin{aligned}\|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}(\phi f)\|_{\mathcal{X}'_{\mathrm{fp}}} \le 2\sqrt{2}\Big(&\|\phi\|_{L^\infty(\Omega)} + \|\nabla_v\phi\|_{L^\infty(\Omega)} \\ &+ \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}\phi\|_{L^\infty(\Omega)}\Big)\|f\|_{H^1_{\mathrm{fp}}(\Omega)}.\end{aligned} \tag{4.8}$$

Combining (4.7) and (4.8) and using that $|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right)|$ is bounded in Ω, we thus have

$$\|\phi f\|_{H^1_{\mathrm{fp}}(\Omega)} \le C\|\phi\|_{C^1(\Omega)}\|f\|_{H^1_{\mathrm{fp}}(\Omega)}.$$

□

Proof of Proposition 4.2.2. We use ideas of the proof of a similar result for transport equations e.g. in [47, Chap. XXI, Thm. 1, p. 220]. Analogous results for spaces similar to $H^1_{\mathrm{fp}}(\Omega)$ are also given in [3, Proofs of Lemmas 4.3, 7.6].

[1] As introduced in section 4.3, we write $\mathcal{X}_{\mathrm{fp}} = L^2(\Omega_{t,x}, V)$.

Given a compact set $K \subset \Gamma_+$, let $\eta_K \in C^1(\bar{\Omega})$ with $\eta_K = 1$ on K and $\operatorname{supp} \eta_K \cap \Gamma_- = \emptyset$. We then obtain by integrating by parts for $w \in C^\infty(\bar{\Omega})$

$$\begin{aligned}\int_K w^2 |\tbinom{1}{v} \cdot \mathbf{n}| \,\mathrm{d}s &= \int_K (\eta_K w)^2 |\tbinom{1}{v} \cdot \mathbf{n}| \,\mathrm{d}s \le \int_{\partial\Omega} (\eta_K w)^2 |\tbinom{1}{v} \cdot \mathbf{n}| \,\mathrm{d}s \\ &\overset{(*)}{=} \int_{\partial\Omega} (\eta_K w)^2 \tbinom{1}{v} \cdot \mathbf{n} \,\mathrm{d}s = 2 \int_\Omega \eta_K w \tbinom{1}{v} \cdot \nabla_{t,x} (\eta_K w) \,\mathrm{d}((t,x),v) \\ &\le 2 \|\eta_K w\|_{L^2(\Omega_{t,x},V)} \|\tbinom{1}{v} \cdot \nabla_{t,x}(\eta_K w)\|_{L^2(\Omega_{t,x},V')} \\ &\le 2 \|\eta_K w\|^2_{H^1_{\mathrm{fp}}(\Omega)} \overset{\text{Lemma 4.2.3}}{\le} C \|\eta_K\|^2_{C^1(\Omega)} \|w\|^2_{H^1_{\mathrm{fp}}(\Omega)}.\end{aligned}$$

We thus have continuity of the mapping $w \mapsto w|_K$ for all $w \in C^\infty(\bar{\Omega})$, and by density (Proposition 4.2.1) the map extends to a continuous operator $H^1_{\mathrm{fp}}(\Omega) \to L^2(K, |(1,v)^T \cdot \mathbf{n}|)$. For $K \subset \Gamma_-$ the claim can be shown analogously using $|\tbinom{1}{v} \cdot \mathbf{n}| = -\tbinom{1}{v} \cdot \mathbf{n}$ on $\operatorname{supp} \eta_K$ in $(*)$. □

This result ensures that $H^1_{\mathrm{fp}}(\Omega)$ functions have a trace on the non-characteristic boundary $\Gamma_+ \cup \Gamma_-$. However, from the local existence of traces we cannot directly deduce that these generally lie in global trace spaces as e.g. $L^2(\partial\Omega, |(1,v)^T \cdot \mathbf{n}|)$.

To include the boundary condition treatment in the function space, we define

$$H^1_{\mathrm{fp},\Gamma_+}(\Omega) := \operatorname{clos}_{\|\cdot\|_{H^1_{\mathrm{fp}}(\Omega)}} C^1_{\Gamma_+}(\bar{\Omega}), \tag{4.9}$$

which will be used as the test space for our variational formulation. With the restriction of functions in $H^1_{\mathrm{fp},\Gamma_+}(\Omega)$ on the outflow boundary and the definition through the closure, we can show that these functions have a trace in $L^2(\Gamma_-, |(1,v)^T \cdot \mathbf{n}|)$:

Proposition 4.2.4. *There exists a linear continuous trace operator* $\gamma_- : H^1_{\mathrm{fp},\Gamma_+}(\Omega) \to L^2(\Gamma_-, |(1,v)^T \cdot \mathbf{n}|)$ *such that*

$$\|\gamma_-(w)\|_{L^2(\Gamma_-,|(1,v)^T\cdot\mathbf{n}|)} \le C \|w\|_{H^1_{\mathrm{fp}}(\Omega)} \quad \forall w \in H^1_{\mathrm{fp},\Gamma_+}(\Omega).$$

Furthermore, the integration by parts formula

$$\int_{\Omega_{t,x}} \langle \tbinom{1}{v} \cdot \nabla_{t,x} w, w \rangle_{V',V} \,\mathrm{d}(t,x) = \tfrac{1}{2} \int_{\Gamma_-} w^2 \tbinom{1}{v} \cdot \mathbf{n} \,\mathrm{d}s$$

holds for all $w \in H^1_{\mathrm{fp},\Gamma_+}(\Omega)$.

Proof. The proof is similar to the respective result for transport equations in Proposition 3.1.6, see also [3, sect. 4]. Let $w \in C^\infty(\bar{\Omega})$ with $w \equiv 0$ on Γ_+. Performing integration by parts we obtain

$$\int_\Omega w \tbinom{1}{v} \cdot \nabla_{t,x} w \,\mathrm{d}((t,x),v) = -\int_\Omega \nabla_{t,x} w \cdot \tbinom{1}{v} w \,\mathrm{d}((t,x),v) + \int_{\Gamma_-} w^2 \underbrace{\tbinom{1}{v} \cdot \mathbf{n}}_{<0} \,\mathrm{d}s,$$

and thus

$$\begin{aligned}\|w\|^2_{L^2(\Gamma_-,|(1,v)^T\cdot\mathbf{n}|)} &= \int_{\Gamma_-} w^2 \, |\tbinom{1}{v} \cdot \mathbf{n}(t,x)| \,\mathrm{d}s = 2 \int_\Omega (-\tbinom{1}{v} \cdot \nabla_{t,x} w) w \,\mathrm{d}((t,x),v) \\ &\le 2 \| -\tbinom{1}{v} \cdot \nabla_{t,x} w\|_{L^2(\Omega_{t,x};V')} \|w\|_{L^2(\Omega_{t,x};V)} \le 2 \|w\|^2_{H^1_{\mathrm{fp}}(\Omega)}.\end{aligned}$$

By density (due to the definition of $H^1_{\mathrm{fp},\Gamma_+}(\Omega)$) the integration by parts formula and the bound for $\|w\|_{L^2(\Gamma_-,|(1,v)^T\cdot\mathbf{n}|)}$ hold for all $w \in H^1_{\mathrm{fp},\Gamma_+}(\Omega)$. □

Remark 4.2.5. Similarly, it can be shown that the space $H^1_{\mathrm{fp},\Gamma_-}(\Omega)$ defined analogously to (4.9) admits a continuous trace operator $\gamma_+ : H^1_{\mathrm{fp},\Gamma_-}(\Omega) \to L^2(\Gamma_+, |(1,v)^T \cdot \mathbf{n}|)$.

While the global existence of the trace and the integration by parts formula can be easily shown for functions in the closure of smooth functions vanishing on the outflow boundary in the same way as for the respective spaces for transport equations, we need a more general result, as well. To later show the uniqueness of the weak solution in section 4.3, we also need to verify the existence of a global trace and the integration by parts formula for certain functions in $H^1_{\mathrm{fp}}(\Omega)$ with vanishing trace on Γ_-, but not necessarily in $H^1_{\mathrm{fp},\Gamma_-}(\Omega)$.

This is established for spaces like $H_{\mathrm{nt}}(\Omega)$, where the kinetic term lies in $L^2(\Omega)$, see subsection 2.3.2. Similar or even stronger results for respective functions in $H^1_{\mathrm{fp}}(\Omega)$ are claimed to be proven in [3, 10, 34], however, we believe the arguments to be incomplete, for more details see Appendix A.

Since we were not able to prove the existence of a global trace for $H^1_{\mathrm{fp}}(\Omega)$ functions with vanishing trace on the inflow or the outflow boundary, we will formulate the exact result needed for uniqueness of the weak solution as an assumption in section 4.3.

4.3 Variational formulation

In this section, we develop a variational formulation for (4.2) and show its well-posedness. To that end, we first define a bilinear form on the velocity domain that can describe an arbitrary elliptic operator as a generalization of the specific diffusion term in (4.2). Then, we choose an ultraweak approach for the kinetic transport operator in the full-dimensional variational formulation: Parallel to the formulation for the transport equation developed in section 3.1, we define the bilinear form with the space and time derivatives on the test function. This is also done in similar formulations in [10, 34], while in [3] a formulation with the transport derivatives on the trial space is used. With our strategy, we can easily handle the inflow boundary conditions similarly to section 3.1 and can later apply the strategy to find stable discrete spaces from the transport equation to the Fokker-Planck equation.

Let $a_v : \Omega_{t,x} \times V \times V \to \mathbb{R}$ be a potentially (x,t)-dependent bilinear form defined on the velocity space V. Moreover, let a_v satisfy the following assumptions:

$$\text{the map } (t,x) \mapsto a_v((t,x);\phi,\psi) \text{ is measurable on } \Omega_{t,x} \text{ for all } \phi,\psi \in V, \tag{4.10}$$

$$a_v((t,x);\cdot,\cdot) \text{ is bilinear for a.e. } (t,x) \in \Omega_{t,x}, \tag{4.11}$$

$$a_v((t,x);\phi,\psi) \le \gamma_v \|\phi\|_V \|\psi\|_V \text{ with } \gamma_v < \infty \text{ for all } \phi,\psi \in V,\ \text{ a.e. } (x,t) \in \Omega_{t,x}, \tag{4.12}$$

$$a_v((t,x);\phi,\phi) + \lambda_v \|\phi\|^2_{L^2(\Omega_v)} \ge \alpha_v \|\phi\|^2_V \text{ with } \lambda_v \in \mathbb{R}, \alpha_v > 0 \tag{4.13}$$
$$\text{for all } \phi \in V,\ \text{ a.e. } (x,t) \in \Omega_{t,x}.$$

Note that γ_v, λ_v, and α_v are assumed to be independent of (x,t).

Example 4.3.1. For the strong form of the Fokker-Planck equation (4.2), a_v is given for all $\phi,\psi \in V$, a.e. $x \in \Omega_x$ by

$$\begin{aligned} a_v(x;\phi,\psi) &= \left(\nabla_v \left(q(x,v)^{-1}\phi(v)\right), \nabla_v \psi(v)\right)_{L^2(\Omega_v)} \\ &= \left(q(x,v)^{-1}\nabla_v \phi(v), \nabla_v \psi(v)\right)_{L^2(\Omega_v)} + \left(\nabla_v q(x,v)^{-1}\phi(v), \nabla_v \psi(v)\right)_{L^2(\Omega_v)}, \end{aligned}$$

where ∇_v is the tangential gradient on Ω_v, see Definition 2.3.13.

If $q^{-1} \in L^\infty(\Omega_x \times \Omega_v)$ with $\nabla_v q^{-1} \in L^\infty(\Omega_x \times \Omega_v)$ and $q^{-1}(x,v) \geq l_q > 0$ for a.e. (x,v), then a_v fulfills the conditions (4.10)–(4.13), for instance, with $\gamma_v = \|q^{-1}\|_{L^\infty} + \|\nabla_v q^{-1}\|_{L^\infty}$, $\alpha_v = \frac{1}{2} l_q$, and $\lambda_v = \|\nabla_v q^{-1}\|^2_{L^\infty}/(2l_q) + \frac{1}{2} l_q$. Depending on q, other estimates might be better, e.g. for $q = q(x)$ and thus $\nabla_v q = 0$ we can get $\alpha_v = \lambda_v = l_q$.

Recalling the function spaces introduced in (4.3) and (4.9), we define the full-dimensional trial and test spaces as

$$\mathcal{X}_{\text{fp}} := L^2(\Omega_{t,x}, V), \qquad \mathcal{Y}_{\text{fp}} := H^1_{\text{fp},\Gamma_+}(\Omega). \tag{4.14}$$

We then define the full bilinear form $b_{\text{fp}} : \mathcal{X}_{\text{fp}} \times \mathcal{Y}_{\text{fp}} \to \mathbb{R}$ for $w \in \mathcal{X}_{\text{fp}}, p \in \mathcal{Y}_{\text{fp}}$ by

$$b_{\text{fp}}(w,p) := \int_{\Omega_{t,x}} \langle w(t,x), -\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p(t,x) \rangle_{V,V'} + a_v((t,x); w(t,x), p(t,x)) \,\mathrm{d}(t,x). \tag{4.15}$$

By definition, b_{fp} is continuous on $\mathcal{X}_{\text{fp}} \times \mathcal{Y}_{\text{fp}}$, i.e.,

$$\begin{aligned} b_{\text{fp}}(w,p) &= \langle w, -\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p \rangle_{\mathcal{X},\mathcal{X}'} + \int_{\Omega_{t,x}} a_v((t,x); w(t,x), p(t,x)) \,\mathrm{d}(t,x) \\ &\leq \|w\|_{\mathcal{X}} \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p\|_{\mathcal{X}'} + \gamma_v \int_{\Omega_{t,x}} \|w(t,x)\|_V \|p(t,x)\|_V \,\mathrm{d}(t,x) \\ &\leq \|w\|_{\mathcal{X}} \|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p\|_{\mathcal{X}'} + \gamma_v \|w\|_{\mathcal{X}} \|p\|_{\mathcal{X}} \\ &\leq \max\{1, \gamma_v\} \|w\|_{\mathcal{X}} \left(\|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} p\|_{\mathcal{X}'} + \|p\|_{\mathcal{X}} \right) \\ &\leq \sqrt{2} \max\{1, \gamma_v\} \|w\|_{\mathcal{X}} \|p\|_{\mathcal{Y}}. \end{aligned}$$

The functional $f : \mathcal{Y}_{\text{fp}} \to \mathbb{R}$ containing the boundary condition $g \in L^2(\Gamma_-, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|)$ is given as

$$f(p) := \int_{\Gamma_-} gp \, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}| \,\mathrm{d}((t,x),v) \quad \forall\, p \in \mathcal{Y}_{\text{fp}},$$

which is well-defined due to Proposition 4.2.4, and we thus have $f \in \mathcal{Y}'_{\text{fp}}$.

We call $u \in \mathcal{X}_{\text{fp}}$ a weak solution of (4.2), if

$$b_{\text{fp}}(u,p) = f(p) \quad \forall\, p \in \mathcal{Y}_{\text{fp}}. \tag{4.16}$$

In the following, we examine the well-posedness of the variational formulation, using the Banach-Nečas-Babuška (or inf-sup) Theorem, see Theorem 2.2.4. We first prove existence of a weak solution in subsection 4.3.1. Then, in subsection 4.3.2 we also show uniqueness of the weak solution under an additional assumption on the trace of certain $H^1_{\text{fp}}(\Omega)$ functions.

4.3.1 Existence of a weak solution

We show the existence of a weak solution u to (4.16) by verifying a dual inf-sup condition. To that end, we construct stable pairs of trial and test space functions such that the application of the bilinear form to the function pairs can be estimated from below by the respective norms of the functions. In these pairs, the trial space functions are derived from the test space functions by the application of the kinetic transport operator and the inverse elliptic velocity operator. We thus generalize similar proofs for parabolic equations [65, 119] using a time derivative instead of the kinetic transport operator and for transport equations using only an application of the transport operator as in chapter 3 and e.g. [43, 52].

Theorem 4.3.2. *The bilinear form b_{fp} satisfies the dual inf-sup condition*

$$\inf_{p\in\mathcal{Y}_{\mathrm{fp}}}\sup_{w\in\mathcal{X}_{\mathrm{fp}}}\frac{b_{\mathrm{fp}}(w,p)}{\|w\|_{\mathcal{X}_{\mathrm{fp}}}\|p\|_{\mathcal{Y}_{\mathrm{fp}}}}\geq\beta_{\mathrm{fp}}$$

with an inf-sup constant

$$\begin{aligned}\beta_{\mathrm{fp}}&\geq\frac{\alpha_v}{\sqrt{2}\max\{1,\gamma_v\}}, && \text{if } a_v \text{ is coercive, i.e., } \lambda_v\leq 0,\\ \beta_{\mathrm{fp}}&\geq\frac{\alpha_v}{\sqrt{2}\max\{1,\gamma_v+\lambda_v\}}\frac{e^{-\lambda_v T}}{\sqrt{\max\{1+2\lambda_v^2,2\}}}, && \text{if } \lambda_v>0.\end{aligned}$$

Consequently, the variational formulation (4.16) has at least one weak solution $u\in\mathcal{X}_{\mathrm{fp}}$.

Remark 4.3.3. The estimates for β_{fp} are not worse than possible estimates for space-time variational formulations for parabolic equations. In fact, for the coercive case and assuming $\alpha_v\leq 1$ and $\gamma_v\geq 1$ the estimate in [119] roughly translates to $\beta_{\mathrm{parab}}\geq\alpha_v^2/(\sqrt{2}\gamma_v^2)$, while we here have $\beta_{\mathrm{fp}}\geq\alpha_v/(\sqrt{2}\gamma_v)$. The exponential dependence on the final time T for the non-coercive case is the same for both types of equations.

Proof of Theorem 4.3.2. We start with the case of a_v being coercive, i.e., $\lambda_v\leq 0$; the non-coercive case will be treated afterwards via a temporal transformation.

To show the inf-sup condition we combine ideas from well-posedness results for parabolic equations as e.g. in [65, 119] with the stable functions pairs defined for transport equations in chapter 3. To that end, we take $0\neq p\in\mathcal{Y}_{\mathrm{fp}}$ arbitrary, but fixed. We want to construct a suitable $w_p\in\mathcal{X}_{\mathrm{fp}}$ and show $b_{\mathrm{fp}}(w_p,p)\geq\beta_{\mathrm{fp}}\|w_p\|_{\mathcal{X}_{\mathrm{fp}}}\|p\|_{\mathcal{Y}_{\mathrm{fp}}}$ for a constant β_{fp} independent of p, which makes β_{fp} a lower bound for the inf-sup constant.

Since $p\in\mathcal{Y}_{\mathrm{fp}}$, we have $f_p:=-\binom{1}{v}\cdot\nabla_{t,x}p\in L^2(\Omega_{t,x};V')=\mathcal{X}'_{\mathrm{fp}}$. Similar to [100, pp. 235], we define the bilinear form $m:\mathcal{X}_{\mathrm{fp}}\times\mathcal{X}_{\mathrm{fp}}\to\mathbb{R}$ by

$$m(w_1,w_2):=\int_{\Omega_{t,x}}a_v((t,x);w_1(t,x),w_2(t,x))\,\mathrm{d}(t,x),\quad\forall w_1,w_2\in\mathcal{X}_{\mathrm{fp}}.$$

Since the function $(t,x)\mapsto a_v((t,x);\phi,\psi)$ is assumed to be measurable for all $\phi,\psi\in V$ (see (4.10)) and $a((t,x),\cdot,\cdot)$ is continuous and coercive with constants γ_v, α_v independent of (t,x) ((4.12) and (4.13) with $\lambda_v\leq 0$), m is well-defined and continuous and coercive over $\mathcal{X}_{\mathrm{fp}}\times\mathcal{X}_{\mathrm{fp}}$ with constants γ_v and α_v. Therefore, by the Lax-Milgram theorem there exists a unique $z_p\in\mathcal{X}_{\mathrm{fp}}$ with

$$m(z_p,w)=\langle f_p,w\rangle_{\mathcal{X}'_{\mathrm{fp}},\mathcal{X}_{\mathrm{fp}}}\quad\forall w\in\mathcal{X}_{\mathrm{fp}}.\tag{4.17}$$

Due to the definitions of z_p, f_p, and m, there holds

$$\int_{\Omega_{t,x}}a_v(z_p,w)\,\mathrm{d}(t,x)=\int_{\Omega_{t,x}}\langle-\binom{1}{v}\cdot\nabla_{t,x}p,w\rangle_{V',V}\,\mathrm{d}(t,x)\quad\forall w\in\mathcal{X}_{\mathrm{fp}}.\tag{4.18}$$

We now define $w_p := p + z_p \in \mathcal{X}_{\mathrm{fp}}$. To bound $b_{\mathrm{fp}}(w_p, p)$ from below we use (4.18) for $w = z_p$ and $w = p$, and the integration by parts formula from Proposition 4.2.4:

$$\begin{aligned}
b_{\mathrm{fp}}(w_p, p) &= \int_{\Omega_{t,x}} \langle p + z_p, -\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \nabla_{t,x} p\rangle_{V,V'} + a_v(p + z_p, p)\,\mathrm{d}(t,x) \\
&= \int_{\Omega_{t,x}} \langle p, -\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \nabla_{t,x} p\rangle_{V,V'} + a_v(z_p, z_p) \\
&\qquad + a_v(p,p) + \langle -\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \nabla_{t,x} p, p\rangle_{V',V}\,\mathrm{d}(t,x) \\
&\geq \alpha_v(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \|z_p\|^2_{\mathcal{X}_{\mathrm{fp}}}) + 2\int_{\Omega_{t,x}} \langle -\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \nabla_{t,x} p, p\rangle_{V',V}\,\mathrm{d}(t,x). \\
&= \alpha_v(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \|z_p\|^2_{\mathcal{X}_{\mathrm{fp}}}) + \int_{\Gamma_-} p^2\,|\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \mathbf{n}|\,\mathrm{d}s \\
&\geq \alpha_v(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \|z_p\|^2_{\mathcal{X}_{\mathrm{fp}}}).
\end{aligned} \tag{4.19}$$

Since we have $\langle f_p, w\rangle_{\mathcal{X}'_{\mathrm{fp}},\mathcal{X}_{\mathrm{fp}}} = m(z_p, w) \leq \gamma_v \|z_p\|_{\mathcal{X}_{\mathrm{fp}}} \|w\|_{\mathcal{X}_{\mathrm{fp}}}$ for all $w \in \mathcal{X}_{\mathrm{fp}}$, there holds

$$\|f_p\|_{\mathcal{X}'_{\mathrm{fp}}} \leq \gamma_v \|z_p\|_{\mathcal{X}_{\mathrm{fp}}}. \tag{4.20}$$

Using the definition of w_p, f_p, and the norm of $\mathcal{Y}_{\mathrm{fp}}$ as defined in (4.5), we can then estimate

$$\begin{aligned}
\|w_p\|_{\mathcal{X}_{\mathrm{fp}}} \|p\|_{\mathcal{Y}_{\mathrm{fp}}} &= \|p + z_p\|_{\mathcal{X}_{\mathrm{fp}}} \left(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \|f_p\|^2_{\mathcal{X}'_{\mathrm{fp}}}\right)^{1/2} \\
&\overset{(4.20)}{\leq} \left[\|p + z_p\|^2_{\mathcal{X}_{\mathrm{fp}}} \left(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \gamma_v^2 \|z_p\|^2_{\mathcal{X}_{\mathrm{fp}}}\right)\right]^{1/2} \\
&\leq \left[2\left(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \|z_p\|^2_{\mathcal{X}_{\mathrm{fp}}}\right)\left(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \gamma_v^2 \|z_p\|^2_{\mathcal{X}_{\mathrm{fp}}}\right)\right]^{1/2} \\
&\leq \sqrt{2}\max\{1, \gamma_v\}\left(\|p\|^2_{\mathcal{X}_{\mathrm{fp}}} + \|z_p\|^2_{\mathcal{X}_{\mathrm{fp}}}\right) \\
&\overset{(4.19)}{\leq} \frac{\sqrt{2}\max\{1, \gamma_v\}}{\alpha_v} b_{\mathrm{fp}}(w_p, p).
\end{aligned} \tag{4.21}$$

Since $p \in \mathcal{Y}_{\mathrm{fp}}$ was chosen arbitrarily, we thus have

$$\inf_{p \in \mathcal{Y}_{\mathrm{fp}}} \sup_{w \in \mathcal{X}_{\mathrm{fp}}} \frac{b_{\mathrm{fp}}(w, p)}{\|w\|_{\mathcal{X}_{\mathrm{fp}}} \|p\|_{\mathcal{Y}_{\mathrm{fp}}}} \geq \beta_{\mathrm{fp}} := \frac{\alpha_v}{\sqrt{2}\max\{1, \gamma_v\}}, \tag{4.22}$$

i.e., the claim for coercive a_v.

To address the case that a_v fulfills the Gårding inequality (4.13) with $\lambda_v > 0$, we use a standard temporal transformation of the full problem as proposed e.g. in [119, 129]. We set $\hat{w} := e^{-\lambda_v t} w$ for $w \in \mathcal{X}_{\mathrm{fp}}$, $\hat{p} = e^{\lambda_v t} p$ for $p \in \mathcal{Y}_{\mathrm{fp}}$, and define the bilinear form $\hat{b}_{\mathrm{fp}} : \mathcal{X}_{\mathrm{fp}} \times \mathcal{Y}_{\mathrm{fp}} \to \mathbb{R}$ by

$$\hat{b}_{\mathrm{fp}}(\hat{w}, \hat{p}) := \int_{\Omega_{t,x}} \langle \hat{w}, -\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \nabla_{t,x} \hat{p}\rangle_{V,V'} + a_v((t,x); \hat{w}, \hat{p}) + \lambda_v(\hat{w}, \hat{p})_{L^2(\Omega_v)}\,\mathrm{d}(t,x). \tag{4.23}$$

Then it holds $b_{\mathrm{fp}}(w, p) = \hat{b}_{\mathrm{fp}}(\hat{w}, \hat{p})$ for all $w \in \mathcal{X}_{\mathrm{fp}}, p \in \mathcal{Y}_{\mathrm{fp}}$. The transformed bilinear form $\hat{b}_{\mathrm{fp}}$ satisfies the definition of b with the transformed velocity bilinear form $\hat{a}_v$:

$V \times V \to \mathbb{R}$ defined by $\hat{a}_v((t,x);\phi,\psi) = a_v((t,x);\phi,\psi) + \lambda_v(\phi,\psi)_{L^2(\Omega_v)}$ for $\phi,\psi \in V$. Due to the Gårding inequality (4.13) and continuity (4.12) of a_v, $\hat{a}_v$ is coercive with constant $\hat{\alpha}_v = \alpha_v$ and continuous with constant $\hat{\gamma}_v = \gamma_v + \lambda_v$. As in [119], we can estimate the norms of $\hat{w} \in \mathcal{X}_{\text{fp}}$ and $\hat{p} \in \mathcal{Y}_{\text{fp}}$ by

$$\|\hat{w}\|_{\mathcal{X}_{\text{fp}}} \geq e^{-\lambda_v T}\|w\|_{\mathcal{X}_{\text{fp}}}, \qquad \|\hat{p}\|_{\mathcal{Y}_{\text{fp}}} \geq \left(\max\{1+2\lambda_v^2, 2\}\right)^{-\frac{1}{2}} \|p\|_{\mathcal{Y}_{\text{fp}}},$$

where we use $\|\psi\|_{V'} \leq \|\psi\|_{L^2(\Omega_v)} \leq \|\psi\|_V$ for the estimation of the $\mathcal{Y}_{\text{fp}}$-norm.

Then, the dual inf-sup constant of b can be bounded from below as follows

$$\begin{aligned}\inf_{p\in\mathcal{Y}_{\text{fp}}} \sup_{w\in\mathcal{X}_{\text{fp}}} \frac{b_{\text{fp}}(w,p)}{\|w\|_{\mathcal{X}_{\text{fp}}}\|p\|_{\mathcal{Y}_{\text{fp}}}} &= \inf_{\hat{p}\in\mathcal{Y}_{\text{fp}}} \sup_{\hat{w}\in\mathcal{X}_{\text{fp}}} \frac{\hat{b}_{\text{fp}}(\hat{w},\hat{p})}{\|\hat{w}\|_{\mathcal{X}_{\text{fp}}}\|\hat{p}\|_{\mathcal{Y}_{\text{fp}}}} \frac{\|\hat{w}\|_{\mathcal{X}_{\text{fp}}}}{\|w\|_{\mathcal{X}_{\text{fp}}}} \frac{\|\hat{p}\|_{\mathcal{Y}_{\text{fp}}}}{\|p\|_{\mathcal{Y}_{\text{fp}}}} \\ &\geq \frac{\alpha_v}{\sqrt{2}\max\{1,\gamma_v+\lambda_v\}} \frac{e^{-\lambda_v T}}{\sqrt{\max\{1+2\lambda_v^2,2\}}}.\end{aligned}$$

As b_{fp} is continuous, the associated operator $B_{\text{fp}} : \mathcal{X}_{\text{fp}} \to \mathcal{Y}'_{\text{fp}}$ defined by $\langle B_{\text{fp}}\cdot,\cdot\rangle_{\mathcal{Y}'_{\text{fp}},\mathcal{Y}_{\text{fp}}} = b_{\text{fp}}(\cdot,\cdot)$ is also bounded. By Proposition 2.2.1, the dual inf-sup condition implies surjectivity of B_{fp}. Therefore, there exists a weak solution $u \in \mathcal{X}_{\text{fp}}$ to (4.16), which concludes the proof. □

4.3.2 Uniqueness of the weak solution

As already mentioned in section 4.2, we were not able to prove all necessary trace results in our specific function space. To show uniqueness of the weak solution, we therefore assume the following:

Assumption 4.3.4. Let $w \in H^1_{\text{fp}}(\Omega)$ with $w = 0$ a.e. on Γ_-, i.e., $w|_K = 0$ for all compact $K \subset \Gamma_-$. Moreover, assume that $b_{\text{fp}}(w,p) = 0$ for all $p \in \mathcal{Y}_{\text{fp}}$. Then, we have $w \in L^2(\partial\Omega, |(1,v)^T \cdot \mathbf{n}|)$ and the integration by parts formula

$$\int_{\Omega_{t,x}} \langle \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w, w\rangle_{V',V} \,\mathrm{d}(t,x) = \frac{1}{2}\int_{\partial\Omega} w^2 \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n} \,\mathrm{d}s \tag{4.24}$$

holds.

As discussed in more detail in Appendix A, we do not know how to prove Assumption 4.3.4, since, for instance, ideas from existing approaches for the related space $H^1_{\text{nt}}(\Omega) = \{w \in L^2(\Omega) : \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w \in L^2(\Omega)\}$ cannot readily be transferred to the $H^1_{\text{fp}}(\Omega)$ case. We therefore leave it as an open problem. We emphasize that the respective trace and integration by parts result holds for all $H^1_{\text{nt}}(\Omega)$ functions with zero inflow or outflow trace (cf. [12,35,36], [47, Chap. XXI]), and also for all $H^1_{\text{fp}}(\Omega)$ functions that can be approximated by smooth functions vanishing on the inflow or outflow boundary (Proposition 4.2.4). Additionally, Assumption 4.3.4 is limited to $H^1_{\text{fp}}(\Omega)$ functions with vanishing trace on Γ_- and satisfying a weak form of the differential equation with zero boundary condition. This additional condition on the considered functions might make it possible to show and exploit a higher regularity of the considered functions to prove existence of suitable traces and (4.24).

We now show uniqueness of the weak solution in the form of surjectivity of the dual operator. To that end, we follow the general structure of respective proofs for parabolic equations [65, Thm 6.6, p. 283] and transport equations [6, Thm. 16], also used for a similar Fokker-Planck equation in [10]: We take a function $w \in \mathcal{X}_{\mathrm{fp}}$ solving (4.16) with zero right-hand side and prove that $w = 0$ step by step by showing that w possesses space- and time derivatives, that w has trace zero on the outflow boundary, and finally that w must therefore vanish on the whole domain.

Theorem 4.3.5. *If Assumption 4.3.4 holds, then for all $0 \neq w \in \mathcal{X}_{\mathrm{fp}}$ we have*

$$\sup_{p \in \mathcal{Y}_{\mathrm{fp}}} b_{\mathrm{fp}}(w, p) > 0.$$

Proof. Let $w \in \mathcal{X}_{\mathrm{fp}}$ such that

$$b_{\mathrm{fp}}(w, p) = 0 \quad \forall\, p \in \mathcal{Y}_{\mathrm{fp}}. \tag{4.25}$$

To prove the claim, we need to show that $w = 0$. First, we show that w has a weak derivative $-\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w \in \mathcal{X}_{\mathrm{fp}}' = L^2(\Omega_{t,x}; V')$. To that end, let $\psi \in C_0^\infty(\Omega_{t,x})$ and $\phi \in V$ be arbitrary. Then $\psi\phi = 0$ on $\partial\Omega$, and by approximating ϕ in $C^\infty(\Omega_v)$ we see that $\psi\phi \in \mathcal{Y}_{\mathrm{fp}}$. Using the definition of the weak (t, x)-derivative and testing (4.25) with $p = \psi\phi$ we obtain

$$\begin{aligned}
&\int_{\Omega_{t,x}} \langle \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w(t, x), \phi\rangle_{V',V}\, \psi(t, x) d(t, x) \\
&\quad = -\int_{\Omega_{t,x}} \langle w(t, x), \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} \psi(t, x)\phi\rangle_{V,V'}\, \mathrm{d}(t, x) \\
&\quad = -\int_{\Omega_{t,x}} a_v((t, x); w(t, x), \psi(t, x)\phi)\, \mathrm{d}(t, x) \\
&\quad = -\int_{\Omega_{t,x}} \langle A_v(t, x) w(t, x), \phi\rangle_{V',V} \psi(t, x)\, \mathrm{d}(t, x),
\end{aligned}$$

where the operator $A_v(t, x) \in \mathcal{L}(V, V')$ is defined as $\langle A_v(t, x)\phi, \rho\rangle_{V',V} = a_v((t, x); \phi, \rho)$ for all $\phi, \rho \in V$, a.e. $(t, x) \in \Omega_{t,x}$. Due to the density of $C_0^\infty(\Omega_{t,x})$ in $L^2(\Omega_{t,x})$ we have

$$-\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w = A_v w \in \mathcal{X}_{\mathrm{fp}}', \tag{4.26}$$

which especially means that $w \in H^1_{\mathrm{fp}}(\Omega)$.

Next, let $K \subset\subset \Gamma_-$ be an arbitrary but fixed compactly embedded subset of Γ_-. Moreover, let $z \in C^\infty(\bar{\Omega})$ with $z = 0$ on $\partial\Omega \setminus K$. We show $wz \in \mathcal{Y}_{\mathrm{fp}}$: Since $w \in H^1_{\mathrm{fp}}(\Omega)$, due to Proposition 4.2.1 there is a sequence $(w_n)_{n\in\mathbb{N}} \subset C^\infty(\bar{\Omega})$ with $\|w_n - w\|_{H^1_{\mathrm{fp}}(\Omega)} \overset{n\to\infty}{\to} 0$. Therefore, we have $w_n z \in C^\infty(\bar{\Omega})$ with $wz = 0$ on Γ_+. Due to Lemma 4.2.3, it holds

$$\|wz - w_n z\|_{H^1_{\mathrm{fp}}(\Omega)} \le C\|z\|_{C^1(\Omega)}\|w - w_n\|_{H^1_{\mathrm{fp}}(\Omega)}$$

and thus $w_n z \to wz$ in $H^1_{\mathrm{fp}}(\Omega)$ as $n \to \infty$. Invoking the definition of $\mathcal{Y}_{\mathrm{fp}}$ in (4.14),(4.9) we obtain $wz \in \mathcal{Y}_{\mathrm{fp}}$.

Since $K \subset \Gamma_-$ is compact, we may apply Proposition 4.2.2 to infer that w has a trace on K and $w|_K \in L^2(K, |(1, v)^T \cdot \mathbf{n}|)$. Thanks to $z|_{\partial\Omega} \in L^\infty(\partial\Omega)$ and $\operatorname{supp} z|_{\partial\Omega} \subset K$, we have

$$\left|\int_{\partial\Omega} w^2 z\, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|\, \mathrm{d}s\right| = \left|\int_K w^2 z\, |\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}|\, \mathrm{d}s\right| \le \|z\|_{L^\infty(K)}\|w\|^2_{L^2(K,|(1,v)^T\cdot\mathbf{n}|)} < \infty.$$

As a consequence we can apply the linear functional in (4.26) to $wz \in \mathcal{Y}_{\text{fp}} \subset \mathcal{X}_{\text{fp}}$, perform integration by parts, since the boundary integral exists, and use (4.25):

$$\begin{aligned}0 &= \int_{\Omega_{t,x}} \langle \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w + A_v w, wz\rangle_{V',V} d(t,x) \\ &= \int_{\Omega_{t,x}} \langle w, -\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x}(wz)\rangle_{V,V'} + a_v(w, wz) d(t,x) + \int_{\partial\Omega} w^2 z \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n} \, \mathrm{d}s \\ &= \underbrace{b_{\text{fp}}(w, wz)}_{=0} - \int_K w^2 z \left|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}\right| \, \mathrm{d}s = -\int_K w^2 z \left|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}\right| \, \mathrm{d}s.\end{aligned}$$

Since $z|_K \in C_0^\infty(K)$ can be chosen arbitrarily and $|\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}| > 0$ on K, the fundamental lemma of calculus of variations yields $w = 0$ a.e. on K. As also $K \subset \Gamma_-$ was chosen arbitrarily, we have $w = 0$ a.e. on Γ_-.

Thanks to Assumption 4.3.4, it therefore holds $w \in L^2(\partial\Omega, |(1,v)^T \cdot \mathbf{n}|)$. We can thus use integration by parts for (4.26) applied to w. Assuming first that a_v is coercive, i.e., $\lambda_v \leq 0$, we obtain

$$\begin{aligned}0 &= \int_{\Omega_{t,x}} \langle \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w + A_v w, w\rangle_{V',V} \, \mathrm{d}(t,x) \\ &= \int_{\Omega_{t,x}} \langle \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w, w\rangle_{V',V} \, \mathrm{d}(t,x) + \int_{\Omega_{t,x}} a_v(w,w) \, \mathrm{d}(t,x) \\ &\geq \tfrac{1}{2} \int_{\Gamma_+} w^2 \underbrace{\left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \mathbf{n}}_{>0} \, \mathrm{d}s + \alpha_v \|w\|^2_{\mathcal{X}_{\text{fp}}},\end{aligned}$$

which implies $w = 0$.

If a_v is not coercive, we use the temporal transformation described in the proof of Theorem 4.3.2. Setting $\hat{w} = e^{-\lambda_v t} w$ and using the definition of $\hat{b}_{\text{fp}}$ in (4.23), we see that (4.25) is equivalent to $\hat{b}_{\text{fp}}(\hat{w}, \hat{p}) = 0$ for all $\hat{p} \in \mathcal{Y}_{\text{fp}}$. Since $\hat{a}_v$ is coercive, we have proven that $\hat{w} = 0$ and thus also $w = 0$. □

We summarize our findings in the following theorem.

Theorem 4.3.6 (Well-posedness)**.** *There exists a solution $u \in \mathcal{X}_{\text{fp}}$ to the variational problem (4.16). If Assumption 4.3.4 holds, the solution is unique and satisfies the stability estimate*

$$\|u\|_{\mathcal{X}_{\text{fp}}} \leq \frac{1}{\beta_{\text{fp}}} \|f\|_{\mathcal{Y}'_{\text{fp}}}$$

for β_{fp} as defined in Theorem 4.3.2.

Proof. Theorem 4.3.2 asserts the dual inf-sup condition and the existence of a solution to (4.16). If additionally Assumption 4.3.4 holds, we also have the dual surjectivity Theorem 4.3.5. Therefore, by Theorem 2.2.4, the solution to (4.16) is unique, and the stability estimate holds. □

4.4 Discretization

We now design a stable and efficient discretization scheme for (4.16). To that end, we use a Petrov-Galerkin projection onto problem-dependent discrete spaces realizing the stable function pairs with test functions $p \in \mathcal{Y}_{\text{fp}}$ and trial functions $w_p \in \mathcal{X}_{\text{fp}}$ developed in the proof of Theorem 4.3.2. As a result, the discrete inf-sup stability and thus the well-posedness of the discrete problem follow analogously to the continuous results with the same stability constant. We then illustrate for a class of data functions how the trial space functions w_p^δ can be efficiently computed by solving low-dimensional elliptic problems in the velocity domain.

4.4.1 Stable Petrov-Galerkin schemes

To define an approximation of the solution $u \in \mathcal{X}_{\text{fp}}$ of (4.16), we use a Petrov-Galerkin projection onto suitable discrete spaces: Given discrete trial and test spaces $\mathcal{X}_{\text{fp}}^\delta \subset \mathcal{X}_{\text{fp}}$ and $\mathcal{Y}_{\text{fp}}^\delta \subset \mathcal{Y}_{\text{fp}}$, the Petrov-Galerkin approximation $u^\delta \in \mathcal{X}_{\text{fp}}^\delta$ is defined by

$$b_{\text{fp}}(u^\delta, v^\delta) = f(v^\delta) \quad \forall v^\delta \in \mathcal{Y}_{\text{fp}}^\delta. \tag{4.27}$$

Well-posedness then depends on the discrete inf-sup stability of the discrete problem. To find a pair of spaces leading to a stable scheme, we transfer the ideas from chapter 3 to the Fokker-Planck setting. For the transport equation, we built a stable discretization with a discrete inf-sup constant of one by fixing a discrete test space and defining a problem dependent trial space with optimal stability properties. For the Fokker-Planck equation we will use the same strategy: We start with a discrete test space and define the corresponding trial space based on the trial space functions used in the proof of Theorem 4.3.2.

To that end, we first define a discrete space $V^h \subset V$ for the discretization in the velocity direction. Since the $\mathcal{Y}_{\text{fp}}$-norm contains a term in the $\mathcal{X}'_{\text{fp}} = L^2(\Omega_{t,x}, V')$-norm which is not computable, we consider the norm

$$\|w\|^2_{L^2(\Omega_{t,x},(V^h)')} := \int_{\Omega_{t,x}} \|w(t,x)\|^2_{(V^h)'} \,\mathrm{d}(t,x), \qquad \|\psi\|_{(V^h)'} := \sup_{\phi^h \in V^h} \frac{\langle \psi, \phi^h \rangle_{V',V}}{\|\phi^h\|_V} \tag{4.28}$$

instead of $\|\cdot\|_{L^2(\Omega_{t,x},V')}$ where necessary.

Let $\mathcal{Y}_{\text{fp}}^\delta \subset \mathcal{Y}_{\text{fp}}$ be a discrete space for which we assume $w^\delta(t,x) \in V^h$ for all $w^\delta \in \mathcal{Y}_{\text{fp}}^\delta$ and a.e. $(t,x) \in \Omega_{t,x}$. $\mathcal{Y}_{\text{fp}}^\delta$ will be used as test space for the Petrov-Galerkin approximation. We define the discrete version of the $\mathcal{Y}_{\text{fp}}$-norm by

$$\|w\|^2_{\mathcal{Y}_{\text{fp}}^\delta} := \|w\|^2_{L^2(\Omega_{t,x},V)} + \|\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \nabla_{t,x} w\|^2_{L^2(\Omega_{t,x},(V^h)')}. \tag{4.29}$$

Since we will make use of the function pairs developed in the proof of Theorem 4.3.2, we assume for the discretization that the velocity bilinear form a_v is coercive, i.e., $\lambda_v \leq 0$. For problems, where a_v only satisfies the Gårding inequality (4.13) with $\lambda_v > 0$, a temporal transform of the problem as described in section 4.3 can be performed, then the transformed problem with a coercive bilinear form $\hat{a}_v$ can be discretized.

We now define a problem-dependent discrete trial space. For each $p^\delta \in \mathcal{Y}_{\text{fp}}^\delta$, we denote $f_p^\delta := \left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right) \cdot \nabla_{t,x} p^\delta(t,x) \in \mathcal{X}'_{\text{fp}}$. We then define the function $z_p^\delta \in \mathcal{X}_{\text{fp}}$ as the solution of

$$a_v(z_p^\delta(t,x), \phi^h) = \langle f_p^\delta(t,x), \phi^h \rangle_{V',V}, \quad \forall \phi^h \in V^h, \text{ a.e. } (t,x) \in \Omega_{t,x}. \tag{4.30}$$

z_p^δ is the discrete counterpart of z_p defined in (4.17), but is here defined pointwise in $\Omega_{t,x}$ due to the discrete setting. Then, the discrete trial space $\mathcal{X}_{\text{fp}}^\delta \subset \mathcal{X}_{\text{fp}}$ is defined as

$$\mathcal{X}_{\text{fp}}^\delta := \{p^\delta + z_p^\delta : p^\delta \in \mathcal{Y}_{\text{fp}}^\delta\}. \tag{4.31}$$

Proposition 4.4.1. *If a_v is coercive, i.e., $\lambda_v \leq 0$ in (4.13), and if the discrete trial and test spaces $\mathcal{X}_{\mathrm{fp}}^\delta$ and $\mathcal{Y}_{\mathrm{fp}}^\delta$ are chosen according to (4.31), then there exists a unique solution $u^\delta \in \mathcal{X}_{\mathrm{fp}}^\delta$ to (4.27).*

Remark 4.4.2. For non-coercive a_v the respective result holds for the discretization of the transformed problem according to (4.23) with $\hat{a}_v$ being coercive.

Proof of Proposition 4.4.1. We can reuse all essential parts of the proof of the inf-sup constant for the continuous problem to also prove discrete inf-sup stability of (4.27).

Let $0 \neq w^\delta \in \mathcal{X}_{\mathrm{fp}}^\delta$ be fixed. Then, by definition of $\mathcal{X}_{\mathrm{fp}}^\delta$ there is $p^\delta \in \mathcal{Y}_{\mathrm{fp}}^\delta$ such that $w^\delta = p^\delta + z_p^\delta$ with z_p^δ defined as in (4.30). By using (4.30) and the same arguments as in (4.19) we obtain

$$b_{\mathrm{fp}}(w^\delta, p^\delta) = b_{\mathrm{fp}}(p^\delta + z_p^\delta, p^\delta) \geq \alpha_v \left(\|p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 + \|z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 \right). \tag{4.32}$$

As we have

$$\langle f_p^\delta(t,x), \phi^h \rangle_{V',V} = a_v(z_p^\delta(t,x), \phi^h) \leq \gamma_v \|z_p^\delta(t,x)\|_V \|\phi^h\|_V \quad \forall \phi^h \in V^h, \text{ a.e. } (t,x) \in \Omega_{t,x}$$

we can inflect that

$$\|f_p^\delta\|_{L^2(\Omega_{t,x},(V^h)')} \leq \gamma_v \|z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}. \tag{4.33}$$

Therefore, we obtain analogously to (4.21), but using the discrete $\mathcal{Y}_{\mathrm{fp}}^\delta$-norm,

$$\begin{aligned} \|w_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}} \|p^\delta\|_{\mathcal{Y}_{\mathrm{fp}}^\delta} &= \|p^\delta + z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}} \left(\|p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 + \|f_p^\delta\|_{L^2(\Omega_{t,x},(V^h)')}^2 \right)^{1/2} \\ &\overset{(4.33)}{\leq} \left[\|p^\delta + z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 \left(\|p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 + \gamma_v^2 \|z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 \right) \right]^{1/2} \\ &\leq \left[2 \left(\|p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 + \|z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 \right) \left(\|p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 + \gamma_v^2 \|z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 \right) \right]^{1/2} \\ &= \sqrt{2} \max\{1, \gamma_v\} \left(\|p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 + \|z_p^\delta\|_{\mathcal{X}_{\mathrm{fp}}}^2 \right) \\ &\overset{(4.32)}{\leq} \frac{\sqrt{2} \max\{1, \gamma_v\}}{\alpha_v} b_{\mathrm{fp}}(w_p^\delta, p^\delta). \end{aligned} \tag{4.34}$$

This means that b_{fp} is inf-sup stable on the spaces $(\mathcal{X}_{\mathrm{fp}}^\delta, \|\cdot\|_{\mathcal{X}_{\mathrm{fp}}}), (\mathcal{Y}_{\mathrm{fp}}^\delta, \|\cdot\|_{\mathcal{Y}_{\mathrm{fp}}^\delta})$ with constant $\beta_{\mathrm{fp}}^\delta \geq \alpha_v (\sqrt{2} \max\{1, \gamma_v\})^{-1}$. Since for all $0 \neq p^\delta$ it holds $b_{\mathrm{fp}}(w_p^\delta, p^\delta) > 0$ and thus $w_p^\delta \neq 0$, we have $\dim(\mathcal{X}_{\mathrm{fp}}^\delta) = \dim(\mathcal{Y}_{\mathrm{fp}}^\delta)$. Therefore, by Proposition 2.2.7, the discrete problem (4.27) is well-posed. □

Remark 4.4.3. Due to the finite-dimensional spaces, the Petrov-Galerkin approximation $u^\delta \in \mathcal{X}_{\mathrm{fp}}^\delta$ is unique even if Assumption 4.3.4 does not hold.

4.4.2 Efficient numerical scheme

Regarding the computational realization of the Petrov-Galerkin approximation, we have to take into account the specific choice of the discrete spaces according to (4.31). To assemble the linear system and to represent the discrete solution, the nonstandard parts of the $\mathcal{X}_{\mathrm{fp}}^\delta$-basis functions, i.e., the functions z_p^δ defined by (4.30), have to be computed for all basis functions of $\mathcal{Y}_{\text{fp}}^{\delta}$. We illustrate how this can be done very efficiently for the case where a_v is coercive and has the separable form

$$a_v((t,x),\phi,\psi) = d(t,x)\tilde{a}_v(\phi,\psi), \tag{4.35}$$

where $d \in L^\infty(\Omega_{t,x})$ satisfies $d(t,x) \geq \alpha^d > 0$ for a.e. $(t,x) \in \Omega_{t,x}$ and $\tilde{a}_v : V \times V \to \mathbb{R}$ is a coercive bilinear form.

To build the discrete test space, let first $\bar{\mathcal{Y}}_{\text{fp}}^{\delta,t,x} \subset H^1(\Omega_{t,x})$ be a discrete space in the space-time domain with basis $(p_i^{t,x,\delta}(t,x))_{i=1}^{n_{t,x}}$ and let $V^h \subset V$ be the already defined velocity discrete space with basis $(\psi_j^h(v))_{j=1}^{n_v}$. Denoting the tensor product of these spaces by $\bar{\mathcal{Y}}_{\text{fp}}^{\delta} := \bar{\mathcal{Y}}_{\text{fp}}^{\delta,t,x} \otimes V^h$, we then set

$$\mathcal{Y}_{\text{fp}}^{\delta} := \operatorname{span}\{p_{i,j}^{\delta} = p_i^{t,x,\delta}\psi_j^h : p_{i,j}^{\delta}|_{\Gamma_+} = 0\} \subset \bar{\mathcal{Y}}_{\text{fp}}^{\delta} \cap \mathcal{Y}_{\text{fp}}.$$

We may then use this tensor product structure to efficiently solve (4.30): Fixing a basis function $p_{i,j}^{\delta} = p_i^{t,x,\delta}\psi_j^h$ of $\mathcal{Y}_{\text{fp}}^{\delta}$, the right-hand side of (4.30) reads

$$\begin{aligned}\langle -\left(\begin{smallmatrix}1\\ v\end{smallmatrix}\right)\cdot\nabla_{t,x}p_{i,j}^{\delta}(t,x),\phi^h\rangle_{V',V} = &- \partial_t p_i^{t,x,\delta}(t,x)\int_{\Omega_v}\psi_j^h(v)\phi^h(v)\,\mathrm{d}v\\ &- \sum_{k=1}^{d}\partial_{x_k}p_i^{t,x,\delta}(t,x)\int_{\Omega_v} v_k\psi_j^h(v)\phi^h(v)\,\mathrm{d}v\end{aligned}$$

for all $\phi^h \in V^h$, a.e. $(t,x) \in \Omega_{t,x}$. Using the separable form of a_v (4.35), we can rewrite (4.30) as follows: Find $z_{i,j}^{\delta} := z_{p_{i,j}^{\delta}}^{\delta} \in \mathcal{X}_{\text{fp}}$, such that

$$\begin{aligned}d(t,x)\tilde{a}_v(z_{i,j}^{\delta}(t,x),\phi^h) = &-\partial_t p_i^{t,x,\delta}(t,x)\int_{\Omega_v}\psi_j^h(v)\phi^h(v)\,\mathrm{d}v\\ &- \sum_{k=1}^{d}\partial_{x_k}p_i^{t,x,\delta}(t,x)\int_{\Omega_v} v_k\psi_j^h(v)\phi^h(v)\,\mathrm{d}v\\ &\qquad\forall\phi^h \in V^h,\ \text{ a.e. } (t,x)\in\Omega_{t,x}.\end{aligned}$$

Hence, the computation of all $z_{i,j}^{\delta}$ can be separated in the following way: We first compute the solutions $\rho_j^1, \rho_j^{v_1}, \dots, \rho_j^{v_d} \in V^h$ to the problems

$$\begin{aligned}\tilde{a}_v(\rho_j^1,\phi^h) &= \int_{\Omega_v}\psi_j^h(v)\phi^h(v)\,\mathrm{d}v, \quad \forall\phi^h\in V^h,\\ \tilde{a}_v(\rho_j^{v_k},\phi^h) &= \int_{\Omega_v}v_k\psi_j^h(v)\phi^h(v)\,\mathrm{d}v, \quad \forall\phi^h\in V^h, k=1,\dots,d,\end{aligned} \tag{4.36}$$

for all basis functions $\psi_j^h \in V^h, j = 1,\dots,n_v$. Then, the $z_{i,j}^{\delta}$ are given by

$$z_{i,j}^{\delta}(t,x,v) = -d(t,x)^{-1}\left(\partial_t p_i^{t,x,\delta}(t,x)\rho_j^1(v) + \sum_{k=1}^{d}\partial_{x_k}p_i^{t,x,\delta}(t,x)\rho_j^{v_k}(v)\right). \tag{4.37}$$

The full solution process thus consists of the following steps:

1. Precompute $\rho_j^1, \rho_j^{v_k}$, i.e., solve $(d+1)\times n_v$ problems of size n_v, which can be done in parallel.

2. Assemble the stiffness matrix $[b_{\mathrm{fp}}(p^\delta_{i,j} + z^\delta_{i,j}, p^\delta_{k,l})]_{(k,l),(i,j)}$, using (4.37), and assemble the load vector $[f(p^\delta_{k,l})]_{(k,l)}$.

3. Solve the linear system of equations to obtain the coefficient vector $[u_{i,j}]_{(i,j)}$.

4. Compose the solution $u^\delta = \sum_{i,j} u_{i,j}(p^\delta_{i,j} + z^\delta_{i,j}) \in \mathcal{X}^\delta_{\mathrm{fp}}$ again using (4.37).

Compared to a naive approach using FE spaces without any stabilization, the additional costs thus only lie in the n_v-sized problems (step 1) and possibly more nonzero elements in the stiffness matrix. These effects only depend on the dimension n_v of V^h. Therefore, the proposed discretization strategy is especially well-suited for using specific spaces V^h of low dimension, which can be achieved for example by using polynomial bases or a hierarchical model reduction approach as proposed in [26].

In order to efficiently compute the problem-dependent basis functions, we heavily rely on the separable form of the bilinear form a_v given in (4.35). To consider more general bilinear forms, the method could for example be combined with low-rank approximations to efficiently compute a discrete trial space as done in a related setting in [17]. More generally, due to the high-dimensionality of the problem, it is especially desirable to combine the discretization with further approximations as the already mentioned hierarchical model reduction [26] or tensor-based methods that have already been used in similar Petrov-Galerkin settings [17, 82] and to discretize kinetic equations like the radiative transfer equation [74, 134] or the Vlasov equation [59, 60, 98].

4.5 Numerical experiments

We investigate the properties of the method developed in section 4.4 by implementing the discretization for a basic model problem. We are especially interested in analyzing how sharp the lower bound for the inf-sup constant is and examining the efficiency in light of the nonstandard discrete spaces. The source code to reproduce all results is provided in [25].

4.5.1 Test Case

We consider the time-independent model problem

$$v \cdot \nabla_x u(x,v) + c\,u(x,v) = d\,\Delta_v u(x,v) + f_0(x,v) \tag{4.38}$$

on the domain $\Omega = \Omega_x \times \Omega_v$ for $\Omega_x = (0,1)^2$ and $\Omega_v = S^1$ and with reaction and velocity diffusion constants $c, d \in \mathbb{R}$, $c, d > 0$. We assume zero inflow boundary conditions on $\Gamma_- \subset \partial\Omega$ and define a source function $f_0 \in L^2(\Omega)$ as a substitute for the initial condition of the time-dependent equation. We parametrize $\Omega_v = S^1$ by the angle $\varphi \in [0, 2\pi)$, leading to $v = \binom{\cos\varphi}{\sin\varphi}$ and $\Delta_v u = \frac{\partial^2}{\partial\varphi^2} u$.

Then, we have $V = H^1(\Omega_v)$ and the bilinear form $a_v : V \times V \to \mathbb{R}$ reads

$$a_v(\psi, \rho) = \int_0^{2\pi} d\,\psi'(\varphi)\rho'(\varphi) + c\,\psi(\varphi)\rho(\varphi)\,\mathrm{d}\varphi \quad \forall \psi, \rho \in V.$$

Thanks to the definition of the $H^1(\Omega_v)$-norm, the bilinear form a_v is coercive with constant $\alpha_v = \min(c,d) > 0$ and continuous with constant $\gamma_v = \max(c,d)$. We then

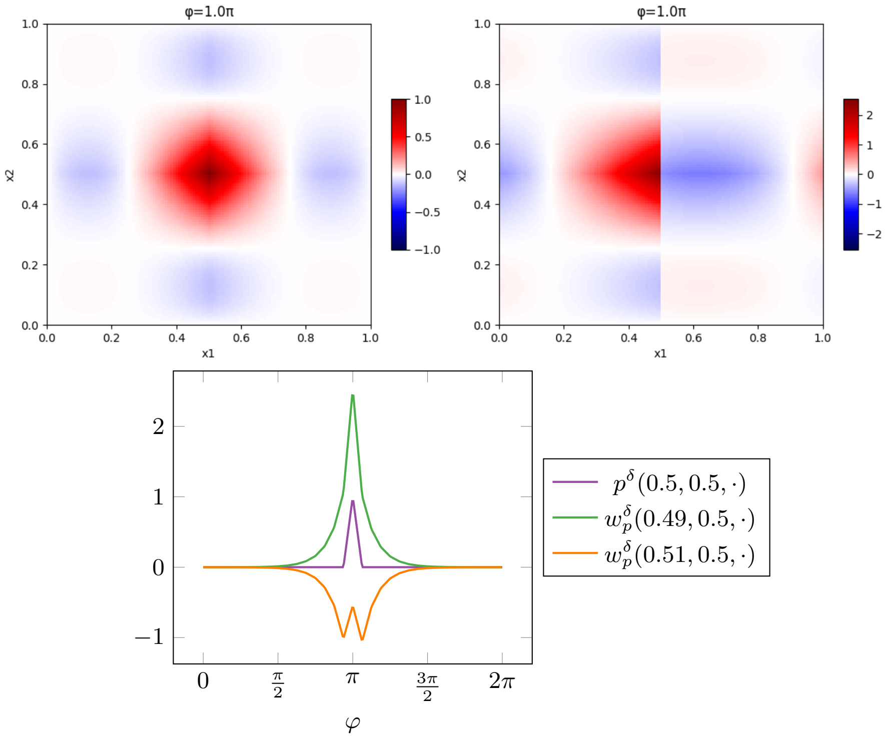

Figure 4.1: Plots of basis functions p^δ and w_p^δ on a grid with only one spatial grid cell ($n_x = 1$) and $n_v = 16$, for data $d = 0.1$, $c = 1$. Upper left: $p^\delta(\cdot, \cdot, \pi)$, upper right: $w_p^\delta(\cdot, \cdot, \pi)$. Lower plot: Velocity dependence of p^δ and w_p^δ for p^δ in the middle of the domain and w_p^δ left and right of the discontinuity.

have $\mathcal{X}_{\text{fp}} := L^2(\Omega_x; H^1(\Omega_v))$, and $\mathcal{Y}_{\text{fp}} = \text{clos}_{\|\cdot\|_{\mathcal{Y}_{\text{fp}}}}\{w \in C^1(\Omega) : w = 0 \text{ on } \Gamma_+\}$, where

$$\Gamma_+ = \{(x, v) \in \partial\Omega_x \times \Omega_v : \left(\begin{smallmatrix}\cos\varphi \\ \sin\varphi\end{smallmatrix}\right) \cdot \mathbf{n}_x > 0\} \subset \partial\Omega,$$
$$\|w\|^2_{\mathcal{Y}_{\text{fp}}} = \|w\|^2_{\mathcal{X}_{\text{fp}}} + \|\left(\begin{smallmatrix}\cos\varphi \\ \sin\varphi\end{smallmatrix}\right) \cdot \nabla_x w\|^2_{\mathcal{X}'_{\text{fp}}}.$$

The full bilinear form is

$$b_{\text{fp}}(w, p) := \int_{\Omega_x} \langle w(x), -\left(\begin{smallmatrix}\cos\varphi \\ \sin\varphi\end{smallmatrix}\right) \cdot \nabla_x p(x)\rangle_{V,V'} + a_v(w(x), p(x))\,\mathrm{d}x, \quad \forall w \in \mathcal{X}_{\text{fp}}, p \in \mathcal{Y}_{\text{fp}}$$

and the functional describing the source term is defined as

$$f(p) := \int_{\Omega_x} \int_0^{2\pi} f_0(x, \varphi) p(x, \varphi)\,\mathrm{d}\varphi\,\mathrm{d}x \quad \forall p \in \mathcal{Y}_{\text{fp}}.$$

Well-posedness of the weak formulation of (4.38) follows completely analogously to the time-dependent case, as a_v is coercive and $f \in \mathcal{Y}'_{\text{fp}}$.

For the discretization we choose $V^h \subset V = H^1(\Omega_v)$ as the continuous linear FE space on $[0, 2\pi)$ with periodic boundary condition and uniform mesh with size $h_v = 2\pi/n_v$. The space $\bar{\mathcal{Y}}^{\delta,x}_{\text{fp}} \subset H^1(\Omega_x)$ is chosen as the continuous $\mathbb{Q}_2$ FE space on a uniform rectangular mesh with size $h_x = \sqrt{2}/n_x$. The trial space $\mathcal{X}^\delta_{\text{fp}}$ is then chosen according to (4.31). For our test case, this amounts to the following form of the trial space functions:

Table 4.1: Computed discrete inf-sup constants for varying mesh sizes and values for the diffusion and reaction constants d and c.

h_x	h_φ	$d = 0.4$, $c = 1$	$d = 0.1$, $c = 1$	$d = 0.1$, $c = 0.1$
$\frac{\sqrt{2}}{4}$	$\frac{2\pi}{4}$	0.61855	0.41087	0.30579
$\frac{\sqrt{2}}{8}$	$\frac{2\pi}{8}$	0.44891	0.18628	0.14924
$\frac{\sqrt{2}}{16}$	$\frac{2\pi}{16}$	0.40915	0.11688	0.10585
$\frac{\sqrt{2}}{32}$	$\frac{2\pi}{32}$	0.40202	0.1033	0.10041
$\frac{\sqrt{2}}{48}$	$\frac{2\pi}{48}$	0.40088	0.10137	0.10008

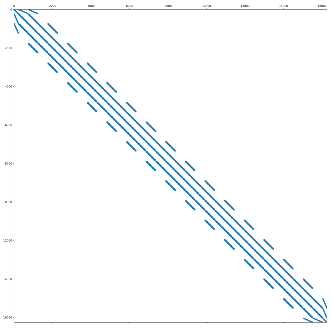

Figure 4.2: Sparsity pattern of the stiffness matrix for $h_{x_1} = h_{x_2} = \frac{1}{16}$, $h_\varphi = \frac{2\pi}{16}$, $\dim \mathcal{Y}_\delta = 16256$.

Given a test space function $p^\delta \in \mathcal{Y}_{\text{fp}}^\delta$, the corresponding stable trial space function defined in subsection 4.4.1 is given by

$$\begin{aligned} w_p^\delta &= p^\delta - (A_v^h)^{-1}\left(\left(\begin{smallmatrix}\cos\varphi\\ \sin\varphi\end{smallmatrix}\right)\cdot\nabla_x p^\delta\right) \\ &= -(-d\Delta_v + c\mathrm{Id})^{-1}\left(\left(\begin{smallmatrix}\cos\varphi\\ \sin\varphi\end{smallmatrix}\right)\cdot\nabla_x p^\delta\right), \end{aligned} \tag{4.39}$$

where we abbreviate the definition of z_p^δ in (4.30) by using the operator $A_v^h := -d\Delta_v + c\mathrm{Id}$ corresponding to the bilinear form a_v on $V^h \times V^h$.

In practice, the trial space $\mathcal{X}_{\text{fp}}^\delta$ is computed as described in subsection 4.4.2 by first solving $2n_v$ problems of dimension n_v. From the definition we see that $\mathcal{X}_{\text{fp}}^\delta \subset \bar{\mathcal{X}}_{\text{fp}}^{\delta,x} \otimes V^h$, with $\bar{\mathcal{X}}_{\text{fp}}^{\delta,x} \subset L^2(\Omega_x)$ being the discontinuous $\mathbb{Q}_2$ FE space.

4.5.2 Numerical results

We first visualize a pair of corresponding trial and test space basis functions. For a given FE basis function in the test space $p^\delta \in \mathcal{Y}_{\text{fp}}^\delta$, the corresponding trial space function $w_p^\delta \in \mathcal{X}_{\text{fp}}^\delta$ is given by (4.39). Since the definition of w_p involves the inverse velocity operator $(A_v^\delta)^{-1}$, w_p may have larger support in Ω_v than p^δ, while the support in Ω_x stays the same. Therefore, we visualize only one spatial grid cell, i.e., $n_x = 1$, and examine the velocity dependence of the functions. We take p^δ as FE nodal basis function with $p^\delta(0.5, 0.5, \pi) = 1$. In Figure 4.1, spatial plots of p^δ and w_p^δ as well as a plot in the velocity direction are given. Indeed, we see that w_p^δ has a non-local support in Ω_v. In the spatial plots, we observe that w_p^δ is positive in $(-1, 0)^T$ direction from the middle of the domain. Since p^δ is a nodal basis function in $\varphi = \pi$, which corresponds to $(\cos(\pi), \sin(\pi))^T = (-1, 0)^T$, we observe that w^δ in a way mimics the transport in $(-1, 0)^T$ direction directly in the spatial domain.

To investigate whether the estimate for the discrete inf-sup constant from section 4.4 is sharp, we compute the constants for different mesh sizes, and reaction and diffusion constants c and d; see Table 4.1. The estimate established in section 4.4 is given in our test case as $\beta_{\text{fp}}^\delta \geq \min\{c, d\}/(\sqrt{2}\max\{1, c, d\})$, which is $\min\{c, d\}/\sqrt{2}$ for all considered data values in Table 4.1. As can be seen in the table, the computed inf-sup constants tend to $\min\{c, d\}$ with increasing mesh sizes for all tested combinations of d and c. In these cases the estimate is thus sharp up to a factor of $\sqrt{2}$.

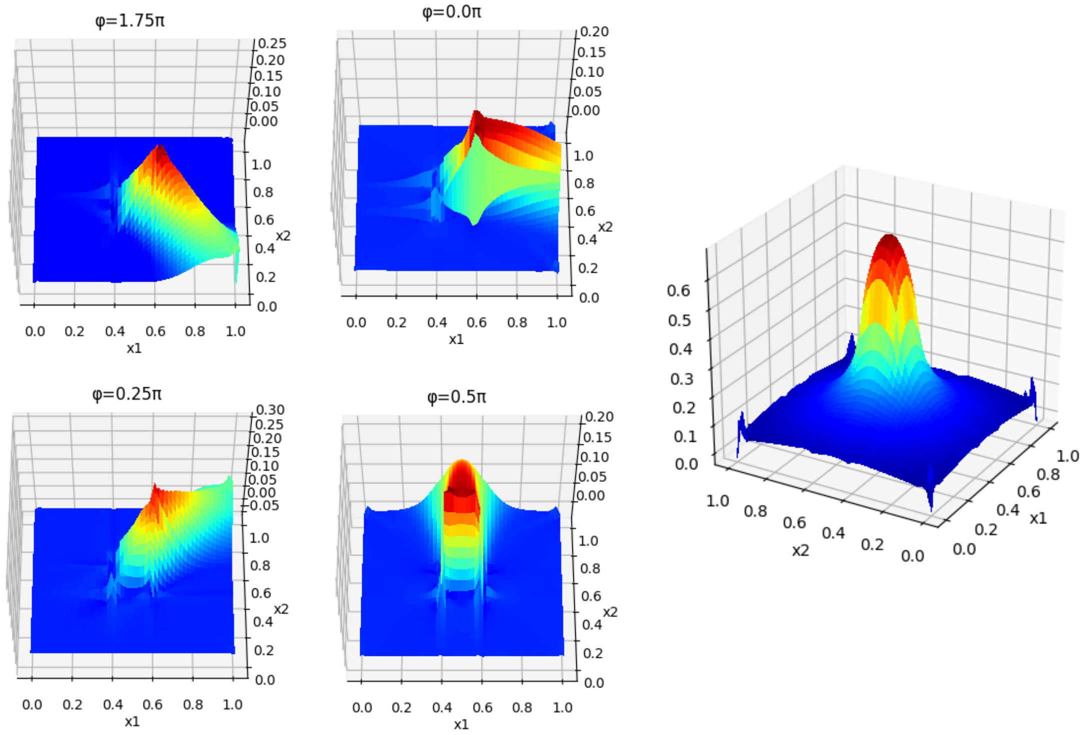

Figure 4.3: Plots of the solution u for $d = c = 0.1$, $f_0 = \chi_{[0.4,0.6]^2}$, $h_{x_1} = h_{x_2} = 1/48$, $h_\varphi = 2\pi/48$, $\dim \mathcal{Y}_\delta = 441984$. Left: $u(\cdot,\cdot,\varphi)$ for different angles φ. Right: moment $\int_0^{2\pi} u(\cdot,\cdot,\varphi)d\varphi$.

Since the basis functions of the discrete trial space $\mathcal{X}_{\mathrm{fp}}^\delta$ are not chosen as standard nodal basis functions but have larger support in Ω_v, one can ask if the choice of spaces still leads to an efficient numerical scheme. Therefore, in Figure 4.2 we plot the sparsity pattern of the stiffness matrix for a discrete problem of dimension 16256. We see that for our test case the choice of spaces indeed leads to a sparse matrix and thus does not induce efficiency problems.

To examine if the nonstandard trial space is able to capture the dynamics of the equation properly, we plot a discrete solution to (4.38) with a particle source f_0 in the middle of the domain, see Figure 4.3. The plots for different angles φ show the particle transport from the middle to the respective directions, the plot of the moment $\int_0^{2\pi} u^\delta(\cdot,\cdot,\varphi)d\varphi$, i.e., the spatial density then shows the overall picture of the particle dynamics. Indeed the nonstandard trial space leads to a realistic solution. We also see that there are no oscillations that would indicate instabilities of the method.

However, we observe small artifacts in the corners of the domain: Since we have chosen the discrete test space $\mathcal{Y}_{\mathrm{fp}}^\delta \subset \bar{\mathcal{Y}}_{\mathrm{fp}}^{\delta,x} \otimes V^h$ in Ω_x analogously to the test space $\mathcal{Y}_{\mathrm{t}}^\delta$ for the transport equation in chapter 3, similar nonphysical restrictions of the trial space as described in subsection 3.2.2 occur here, as well. More precisely, the space $\bar{\mathcal{Y}}_{\mathrm{fp}}^{\delta,x}$ has the same tensor product structure as $\mathcal{Y}_{\mathrm{t}}^\delta$. Moreover, functions in the full space $\mathcal{Y}_{\mathrm{fp}}^\delta$ vanish at the outflow boundary, which is here velocity dependent, but has, for fixed v, the same form as the outflow boundary of $\mathcal{Y}_{\mathrm{t}}^\delta$. The trial space $\mathcal{X}_{\mathrm{fp}}^\delta$ is then built from the test space $\mathcal{Y}_{\mathrm{fp}}^\delta$. This leads to trial space functions that vanish for each v on the respective “outflow corner” analogously to the transport case described in subsection 3.2.2. Therefore, we see in Figure 4.3 that the solution u vanishes e.g. for $\varphi = 1.75\pi$ on the corner $x = (1,0)^T$. The moment of u does not necessarily vanish on the corners. However, we observe artifacts on all corners, since the problem describes a transport from the middle of the domain to the exterior with zero inflow, so that we have only “outflow contributions” on all corners. To mitigate this effect, one could for example choose the computational domain larger than the domain of interest as proposed in subsection 3.2.2, or use other spaces that are not based on tensor product spaces.

5 Conclusion and outlook

5.1 Conclusion

In this thesis we developed stable Petrov-Galerkin discretizations for parametrized linear first-order transport equations and for kinetic Fokker-Planck equations.

The numerical scheme for the transport equation is based on an ultraweak variational formulation already used for related methods such as DPG formulations [51, 52] and in [43]. By putting all derivatives on the test function and choosing the trial space as $L^2(\Omega)$ and the test space norm including the whole adjoint operator B_{t}^*, the variational formulation is optimally conditioned on the infinite-dimensional level.

To retain this optimal stability also on the discrete level, we "reversed" the classical strategy to find a stable test space to a given trial space: Instead of fixing the trial space and trying to approximate the "optimal test space" consisting of the supremizer functions $B_{\mathrm{t}}^{-*}w \in \mathcal{Y}_{\mathrm{t}}$ for all $w \in \mathcal{X}_{\mathrm{t}}$, as has been done with different strategies in related works [29, 43, 51, 52], we fix a discrete test space $\mathcal{Y}^\delta$ and choose the nonstandard trial space $\mathcal{X}^\delta := B_{\mathrm{t}}^* \mathcal{Y}^\delta$. With this choice, we obtain a discrete inf-sup constant of one for a Petrov-Galerkin scheme with directly computable spaces. This especially means that the discrete solution is the L^2-best approximation in the (nonstandard) trial space. This best-approximation problem can be rewritten as a respective problem in the test space $\mathcal{Y}^\delta$ with the transport-operator-norm $\|\cdot\|_{\mathcal{Y}_{\mathrm{t}}}$, we can therefore ensure convergence of the scheme by choosing an appropriate FE space for $\mathcal{Y}^\delta$.

The numerical experiments show convergence of order about $\frac{1}{3}$ for non-smooth L^2-solutions. Despite the L^2-framework, higher convergence orders between 1 and 2 can be observed for smooth solutions, even though tensor product discrete spaces may limit the convergence order to 1 due to unphysical restrictions of the trial space at the outflow boundary. The proposed method shows similar ratios of errors and computational costs to [43], where fixed trial spaces are used. We thus conclude that our nonstandard problem-dependent trial spaces have satisfying approximation properties for the considered test cases.

We used this framework to develop an efficient realization and implementation of RB methods for parametrized transport equations. Unlike standard RB models, our reduced model consists of a reduced test space with fixed functions, but a parameter-dependent norm, while the reduced trial space has parameter-dependent basis functions but the common L^2-norm. A (strong) greedy algorithm generates the reduced test space consisting of "test space snapshots", which ensure that for the chosen parameter values the model error is zero, even if the reduced trial spaces are not solely consisting of common trial space snapshots. This unusual choice of spaces guarantees that the reduced model is automatically optimally stable for all parameter values, which means that we do not need additional stabilization in the basis generation process and have a reduced model with the same trial and test space dimension, unlike the related reduced models generated by the double greedy algorithm [45]. Since due to the nonstandard test space norm the standard residual-based RB error estimator is not offline-/online decomposable in our setting, we proposed as an alternative a hierarchical error estimator based on the comparison of reduced spaces of different model order.

In the numerical experiments, we saw that our reduced model realizes the convergence order of the Kolmogorov n-width for a non-smooth transport problem. A comparison with the algorithm in [45] showed comparable, or even better convergence rates and significantly lower online costs for the new framework. The results suggest that the new framework might be especially beneficial for problems where a stabilization is rather challenging.

We then presented a stable Petrov-Galerkin discretization of a kinetic Fokker-Planck equation. To that end, we first developed a new proof for the well-posedness of a variational formulation in all dimensions with the kinetic transport operator on the test space. Combining ideas from similar proofs for parabolic equations and transport equations we gave a lower bound for the dual inf-sup constant which is not worse than respective estimates for parabolic equations [119]. The proof is based on specific "stable function pairs" given by a test space function and a trial space function obtained by applying the kinetic transport and the inverse velocity Laplace-Beltrami operator. Under an additional assumption on the traces of certain functions in the Fokker-Planck function space $H^1_{\mathrm{fp}}(\Omega)$, we obtained well-posedness of the variational formulation.

To derive stable discrete spaces, we adapted the "optimal trial" strategy from our discretization of the transport equation to the Fokker-Planck case: By defining the discrete trial space dependent on the chosen discrete test space through the application of the kinetic transport and the inverse velocity Laplace-Beltrami operator, we built the "stable function pairs" introduced in the continuous inf-sup estimate into our discrete spaces. We hence obtained a well-posed numerical scheme with the same lower bound of the discrete inf-sup constant as for the continuous problem independently of the mesh size. We showed that under suitable conditions on the data functions these spaces can be computed efficiently, since as in the transport case the (high-dimensional) kinetic transport operator only has to be applied to the test space functions, while for the inverse Laplace-Beltrami operator low-dimensional elliptic problems have to be solved, which are, however, not dominant in the overall computational costs.

The numerical experiments showed that for the examined test case the estimate of the discrete inf-sup constant is sharp up to a factor of $\sqrt{2}$ and confirmed that our choice of spaces leads to an efficient scheme.

5.2 Outlook

The proposed methods provide several starting points for future research.

On the one hand, the proposed "optimal trial" framework for the parametrized transport equation can be explored further by evaluating other possibilities for discrete test spaces beyond the simple tensor product FE spaces we used so far. Depending on the choice of the space, the observed nonphysical corner restriction and overshoots might be mitigated and possibly specific convergence estimates for the scheme might be derived. On the other hand, it would be especially interesting to apply the "optimal trial" strategy to other equations. For the kinetic Fokker-Planck equation we used a formulation with the transport operator on the test space, but velocity derivatives still on both spaces as usual. Therefore, the computation of the optimal function pair still included the inversion of the (in this case low-dimensional) elliptic operator. The “optimal trial” framework is however especially favorable for true “ultraweak” approaches with all derivatives on the test space, since then the trial space is obtained by only applying differential operators. An application to other transport problems as, for instance, hyperbolic systems or to a true “ultraweak” formulation of second-order equations might be interesting.

In the same way, the model reduction framework might be beneficial in the reduction of other parametrized problems where the stability of reduced models is an issue. However, the success of our linear approach is limited by the decay of the Kolmogorov-n-width, which is very slow for transport problems. It might therefore be interesting to also look at different proposed strategies to circumvent the Kolmogorov-n-width by nonlinear transformations or adaptations and combine our framework with these.

For the kinetic Fokker-Planck equation we could show the existence of the weak solution, but the uniqueness only under an additional assumption on the global traces of certain $H^1_{\mathrm{fp}}(\Omega)$ functions, which we left as an open problem, see Assumption 4.3.4 and Appendix A. The further investigation of the trace properties of $H^1_{\mathrm{fp}}(\Omega)$ remains thus subject of our future work.

Regarding the discretization, we here developed a stable numerical scheme for the kinetic Fokker-Planck equation. However, the largest problem for the numerical solution of the equation in practice beyond low-dimensional test problems is the dimension of the underlying phase space which makes schemes based on standard FE spaces prohibitively expensive. Therefore, dimensional reduction approaches or tensorized methods are necessary. The hierarchical model reduction approach developed in [26] used an expansion in problem-dependent basis functions in the velocity variable, which fits into our scheme where a small dimension in the velocity space is favorable for an efficient computation of the trial space functions. A combination with our scheme (that would mitigate the stability problems encountered in [26]) might therefore be interesting. Alternatively, a combination with low-rank or tensor-based decompositions might be interesting.

A Discussion: Global traces in $\mathbf{H}^1_{\mathbf{fp}}(\boldsymbol{\Omega})$

As already mentioned in section 4.2, we believe that the statement in Assumption 4.3.4 is still an open problem, despite the fact that more general results implying the respective version of Assumption 4.3.4 hold true for L^2-based spaces, and that similar results for Fokker-Planck equations are given in other works.

More precisely, on the one hand, [3, Lemma 4.5] states that the space $C_0^\infty(\bar{\Omega} \setminus \Gamma_0)$ of smooth functions vanishing in a neighborhood of Γ_0 is dense in $H^1_{\text{fp}}(\Omega)$, and that $H^1_{\text{fp}}(\Omega)$ functions lying in $L^2(\Gamma_+, |(1,v)^T \cdot \mathbf{n}|)$ or $L^2(\Gamma_-, |(1,v)^T \cdot \mathbf{n}|)$ already have a full global trace in $L^2(\partial\Omega, |(1,v)^T \cdot \mathbf{n}|)$.

On the other hand, in [34], and based on that also in [10] the following is stated[1]:

Claim A.1 (cf. [34, Lemma 2.3], [10, p. 3493])**.** Let $w \in H^1_{\text{fp}}(\Omega)$. Then, w has traces $w|_{\Gamma_\pm} \in L^2(\Gamma_\pm, |(1,v)^T \cdot \mathbf{n}|)$ and the integration by parts formula (4.24) holds.

Note that [3, Lemma 4.5] would already imply Assumption 4.3.4 (where functions that have zero trace on Γ_- are considered), while Claim A.1 is an even stronger claim. However, we believe that the arguments both for [3, Lemma 4.5] and for Claim A.1 given in [3, 10, 34] are incomplete.

While the function spaces considered for the different versions of the Fokker-Planck equation are typically of the form

$$H^1_{\text{fp}}(\Omega) = \{w \in \mathcal{X} : \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w \in \mathcal{X}'\} \text{ with } \mathcal{X} = L^2(\Omega_{t,x}, H^1(\Omega_v)),$$

(cf. (4.4)), function spaces for other kinetic equations like neutron transport require considering

$$H_{\text{nt}}(\Omega) = \{w \in L^2(\Omega) : \left(\begin{smallmatrix}1\\v\end{smallmatrix}\right) \cdot \nabla_{t,x} w \in L^2(\Omega)\},$$

(cf. (2.16)). In subsection 2.3.2, we summarize some trace properties of $H_{\text{nt}}(\Omega)$ that were shown in [12, 35, 36], see also [47, Chap. XXI]. On the one hand one can show similarly for $H_{\text{nt}}(\Omega)$ and $H^1_{\text{fp}}(\Omega)$ that the spaces both admit *local* L^2-traces on Γ_- and Γ_+, see Proposition 2.3.8 and Proposition 4.2.2.

We note that while the statement of [3, Lemma 4.5] and thus of Assumption 4.3.4 holds for $H_{\text{nt}}(\Omega)$ instead of $H^1_{\text{fp}}(\Omega)$ (see Proposition 2.3.10), Claim A.1 is not true for $H_{\text{nt}}(\Omega)$ functions, see Example 2.3.9.

The argument to show Claim A.1 in [34] is based on two steps:

Step 1: Show that the space $C_0^\infty(\bar{\Omega} \setminus \Gamma_0)$ is dense in $H^1_{\text{fp}}(\Omega)$ (this is also included in [3, Lemma 4.5])

Step 2: Decompose $w \in C_0^\infty(\bar{\Omega} \setminus \Gamma_0)$ into $w = w_+ + w_-$ with $w_\pm$ vanishing on $\Gamma_\pm$ and use the density from *Step 1* to show the claim.

[1] In describing the estimates in different cited works, we substitute the notation and the concrete spaces to the respective equivalent in this work to simplify the discussion. This sometimes slightly changes the spaces, but has no effect on the used arguments.

We believe the arguments in both steps to be incomplete.

For *Step 1*, the author of [34] refers to [49], where a time-dependent Fokker-Planck equation in one space dimension is considered and where it is stated that using an argument of Bardos ([12, p. 203]), it can be seen that $C_0^\infty(\bar{\Omega} \setminus \Gamma_0)$ is dense in $H^1_{\mathrm{fp}}(\Omega)$.

The work of Bardos [12] considers the L^2-based function spaces $H(\Omega, \mathbf{b})$ and $H_{\mathrm{nt}}(\Omega)$, see subsections 2.3.1 and 2.3.2. The mentioned argument corresponds to Lemma 2.3.4.

In the respective proof, a family of functions $(\phi_\varepsilon)_{\varepsilon>0} \subset C^\infty(\bar{\Omega})$ is constructed such that ϕ_ε vanishes in the $\frac{\varepsilon}{2}$-neighborhood of $\partial\Gamma_-$, and is equal to unity outside of the ε-neighborhood, see (2.15). Given $u \in H_{\mathrm{nt}}(\Omega) \cap L^\infty(\Omega)$, it is then shown that $u\phi_\varepsilon \to u$ in $L^2(\Omega)$ and $\binom{1}{v} \cdot \nabla_{t,x}(u\phi_\varepsilon) \to \binom{1}{v} \cdot \nabla_{t,x}u$ in $L^2(\Omega)$ as $\varepsilon \to 0$. The estimate of $\|u\binom{1}{v} \cdot \nabla_{t,x}\phi_\varepsilon\|_{L^2(\Omega)}$ uses the fact that $|\nabla\phi_\varepsilon| < C\varepsilon^{-1}$ while $\operatorname{supp}\phi_\varepsilon$ has a measure bounded by $C\varepsilon^2$ and $u \in L^\infty(\Omega)$.

Subsequently, in [12] this density result is used to show that functions with vanishing trace on Γ_- can be approximated by smooth functions that vanish on Γ_-, see also Proposition 2.3.5.

To use the same approach for $H^1_{\mathrm{fp}}(\Omega)$, one needs to show that $\|u - u\phi_\varepsilon\|_{H^1_{\mathrm{fp}}(\Omega)} \to 0$ as $\varepsilon \to 0$, which has not been addressed in [34, 49]. The convergence of $\|u - u\phi_\varepsilon\|_{\mathcal{X}}$ can indeed be shown analogously to the proof of Lemma 2.3.4, since the additional term $\|\nabla_v(u - u\phi_\varepsilon)\|_{L^2(\Omega)}$ can be treated exactly as the L^2-derivative term in [12], see also [3].

However, it is unclear to us how to show convergence (or even boundedness independently of ε) for $\|\binom{1}{v} \cdot \nabla_{t,x}(u - u\phi_\varepsilon)\|_{\mathcal{X}'}$. Instead of L^2-norms, here it is required to have an estimate of the form

$$\langle \phi_\varepsilon \tbinom{1}{v} \cdot \nabla_{t,x}u, \psi\rangle_{\mathcal{X}',\mathcal{X}} = \langle \tbinom{1}{v} \cdot \nabla_{t,x}u, \phi_\varepsilon\psi\rangle_{\mathcal{X}',\mathcal{X}} \leq C(u)\|\psi\|_{\mathcal{X}} \quad \forall \psi \in \mathcal{X}.$$

Unfortunately, we do not know how to obtain such an estimate. Note, that $\|\phi_\varepsilon\psi\|_{\mathcal{X}}$ cannot be bounded analogously to the proof of Lemma 2.3.4, since generally $\psi \notin L^\infty(\Omega)$. We therefore do not see how to obtain a bound of $\|\phi_\varepsilon\psi\|_{\mathcal{X}}$ independently of ε as claimed in [3]. Therefore, it is unclear to us if and how the approach for $H_{\mathrm{nt}}(\Omega)$ can be transferred to $H^1_{\mathrm{fp}}(\Omega)$ to show Assumption 4.3.4, [3, Lemma 4.5], and *Step 1* in the proof of Claim A.1 by [34].

In *Step 2* to prove Claim A.1, the authors of [10, 34] decompose a function $\psi \in C_0^\infty(\bar{\Omega} \setminus \Gamma_0)$ into a sum $\psi = \psi_+ + \psi_-$, where $\psi_\pm \in C^\infty(\bar{\Omega})$ vanish on $\Gamma_\pm$. Using integration by parts separately for $\psi_\pm$, one can show that

$$\|\psi_+\|_{L^2(\Gamma_-, |(1,v)^T \cdot \mathbf{n}|)} \leq C\|\psi_+\|_{H^1_{\mathrm{fp}}(\Omega)} \quad \text{and} \quad \|\psi_-\|_{L^2(\Gamma_+, |(1,v)^T \cdot \mathbf{n}|)} \leq C\|\psi_-\|_{H^1_{\mathrm{fp}}(\Omega)}.$$

The authors conclude from this that $\|\psi\|_{L^2(\Gamma_+ \cup \Gamma_-, |(1,v)^T \cdot \mathbf{n}|)} \leq C\|\psi\|_{H^1_{\mathrm{fp}}(\Omega)}$. However, as already noted in [3, Appendix], it is not clear if this conclusion holds for a constant C independently of ψ, since the decomposition actually leads to

$$\|\psi\|_{L^2(\Gamma_+ \cup \Gamma_-, |(1,v)^T \cdot \mathbf{n}|)} \leq C(\|\psi_+\|_{H^1_{\mathrm{fp}}(\Omega)} + \|\psi_-\|_{H^1_{\mathrm{fp}}(\Omega)}). \tag{A.1}$$

It is unclear to us whether ψ_+ and ψ_- can be chosen in such a way that their single norms can be bounded from above by $\|\psi\|_{H^1_{\mathrm{fp}}(\Omega)}$, see also the discussion in [3].

We emphasize again that Claim A.1 does not hold for $H_{\mathrm{nt}}(\Omega)$, as demonstrated in Example 2.3.9. Since $H^1_{\mathrm{fp}}(\Omega)$ functions have additional regularity in the velocity variable, one cannot conclude from the $H_{\mathrm{nt}}(\Omega)$ case that Claim A.1 must be false. However, we conjecture that a proof of Claim A.1 has to rely on this additional regularity to exploit the difference between the spaces.

An indicator that the additional regularity may indeed make a difference can be seen when returning to Example 2.3.9 and evaluating the proposed function in the $H^1_{\mathrm{fp}}(\Omega)$-norm instead of the $H_{\mathrm{nt}}(\Omega)$-norm:

Example A.2. Consider Example 2.3.9 (which stems from [102, pp. 562-563]) in two space dimensions. Let $\Omega_x = B_1(0) \subset \mathbb{R}^2$ and $\Omega_v = S^1$. For $q \geq 0$, we define $w_q : \Omega_x \times \Omega_v \to \mathbb{R}$ as[2]

$$w_q(x_1, x_2, \phi) = (1 - |x_2 \cos\phi - x_1 \sin\phi|)^{-q},$$

which is in polar coordinates $(x_1, x_2) = (r\cos(\varphi), r\sin(\varphi))$:

$$w_q(r, \varphi, \phi) = (1 - r|\sin(\varphi - \phi)|)^{-q}.$$

For the L^2-norm, we obtain by using the 2π-periodicity of the sin function

$$\begin{aligned}
\|w_q\|^2_{L^2(\Omega)} &= \int_0^1 \int_0^{2\pi} \int_0^{2\pi} (1 - r|\sin(\varphi - \phi)|)^{-2q} r \,\mathrm{d}\phi \,\mathrm{d}\varphi \,\mathrm{d}r \\
&= \int_0^1 \int_0^{2\pi} \int_{-\phi}^{2\pi-\phi} (1 - r|\sin(\hat\varphi)|)^{-2q} r \,\mathrm{d}\hat\varphi \,\mathrm{d}\phi \,\mathrm{d}r \\
&= 2\pi \int_0^1 \int_0^{2\pi} (1 - r|\sin(\hat\varphi)|)^{-2q} r \,\mathrm{d}\hat\varphi \,\mathrm{d}r \\
&= 2\pi \int_{-1}^1 \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} (1 - |y|)^{-2q} \,\mathrm{d}x \,\mathrm{d}y \\
&= 8\pi \int_0^1 \sqrt{1 - y^2}(1 - y)^{-2q} \,\mathrm{d}y \\
&= 8\pi \int_0^1 \sqrt{1 + y}(1 - y)^{\frac{1}{2} - 2q} \,\mathrm{d}y.
\end{aligned}$$

With $1 \leq \sqrt{1+y} \leq 2$, we see that $w_q \in L^2(\Omega)$ if and only of

$$\int_0^1 (1 - y)^{\frac{1}{2} - 2q} \,\mathrm{d}y < \infty \quad \Longleftrightarrow \quad q < \tfrac{3}{4}.$$

For $\|w_q\|_{L^2(\Omega_x, V)} = (\|w_q\|^2_{L^2(\Omega)} + \|\partial_\phi w_q\|^2_{L^2(\Omega)})^{\frac{1}{2}}$, we additionally need the velocity derivative. We have

$$\partial_\phi (1 - r|\sin(\phi - \varphi)|)^{-q} = -q(1 - r|\sin(\phi - \varphi)|)^{-q-1}(-r\,\mathrm{sgn}(\sin(\phi - \varphi))\cos(\phi - \varphi),$$

and thus

$$\begin{aligned}
\|\partial_\phi w_q\|^2_{L_2(\Omega)} &= \int_0^1 \int_0^{2\pi} \int_0^{2\pi} q^2 r^2 \cos^2(\phi - \varphi)(1 - r|\sin(\phi - \varphi)|)^{-2q-2} r \,\mathrm{d}\phi \,\mathrm{d}\varphi \,\mathrm{d}r \\
&= 2\pi q^2 \int_0^1 \int_0^{2\pi} r^2 \cos^2(\hat\varphi)(1 - r|\sin(\hat\varphi)|)^{-2q-2} r \,\mathrm{d}\hat\varphi \,\mathrm{d}r
\end{aligned}$$

[2] Note, that w_q is the respective 2D version of the function w_q from Example 2.3.9. Here, we give an explicit definition of the velocity dependence that was described as the appropriate rotation of a function for fixed $v = v_0$ in Example 2.3.9.

$$\begin{aligned}
&= 2\pi q^2 \int_{-1}^{1} \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} x^2 (1-|y|)^{-2q-2} \,\mathrm{d}x\,\mathrm{d}y \\
&= 4\pi q^2 \int_0^1 \tfrac{2}{3}(1-y^2)^{\frac{3}{2}} (1-y)^{-2q-2} \,\mathrm{d}y \\
&= \frac{8\pi q^2}{3} \int_0^1 (1+y)^{\frac{3}{2}} (1-y)^{-2q-2+\frac{3}{2}} \,\mathrm{d}y.
\end{aligned}$$

Hence, it holds $\partial_\phi w_q \in L^2(\Omega)$ if and only if

$$\int_0^1 (1-y)^{-2q-\frac{1}{2}} \,\mathrm{d}y < \infty \quad \Longleftrightarrow \quad q < \tfrac{1}{4}.$$

The function is chosen such that $v \cdot \nabla_x w_q = 0$. Hence, with these computations we see that

$$\begin{aligned}
\|w_q\|^2_{H_{\mathrm{nt}}(\Omega)} &= \|w_q\|^2_{L^2(\Omega)} + \|v \cdot \nabla_x w_q\|^2_{L^2(\Omega)} < \infty \quad \Longleftrightarrow \quad q < \tfrac{3}{4}, \\
\|w_q\|^2_{H^1_{\mathrm{fp}}(\Omega)} &= \|w_q\|^2_{L^2(\Omega_x,V)} + \|v \cdot \nabla_x w_q\|^2_{L^2(\Omega,V')} < \infty \quad \Longleftrightarrow \quad q < \tfrac{1}{4}.
\end{aligned}$$

From [102], we also see

$$\begin{aligned}
\|w_q\|^2_{L^2(\partial\Omega,|v\cdot\mathbf{n}|)} &= \int_{\partial\Omega} w^2 |v \cdot \mathbf{n}| \,\mathrm{d}(x,v) < \infty \Longleftrightarrow q < \tfrac{1}{2}, \\
\|w_q\|^2_{L^2(\partial\Omega)} &= \int_{\partial\Omega} w_q^2 \,\mathrm{d}(x,v) < \infty \Longleftrightarrow q < \tfrac{1}{4}.
\end{aligned}$$

Therefore, choosing q such that $w_q \in H^1_{\mathrm{fp}}(\Omega)$ implies $w_q \in L^2(\partial\Omega, |v \cdot \mathbf{n}|)$ and even $w_q \in L^2(\partial\Omega)$.

We see that this particular counterexample for the $H_{\mathrm{nt}}(\Omega)$-version of Claim A.1 is *not* a counterexample for Claim A.1. This shows that the additional velocity regularity of $H^1_{\mathrm{fp}}(\Omega)$ may indeed be helpful (and crucial) to show global trace results for $H^1_{\mathrm{fp}}(\Omega)$. We note, however, that here the “problematic” term $\|v \cdot \nabla_x w_q\|_{L^2(\Omega_x,V')}$ vanished.

To summarize the discussion, we believe the arguments in [3,10,34] to be incomplete. Moreover, we do not know how to use ideas from the existing approaches for $H^1_{\mathrm{nt}}(\Omega)$ to show Assumption 4.3.4, since we are unsure how to compensate for the missing L^2 regularity of the transport term even with a higher regularity in the velocity direction. Therefore, we leave Assumption 4.3.4 as an open problem.

List of Symbols

$\mathcal{X}$	Trial space for abstract variational formulation	12
$\mathcal{Y}$	Test space for abstract variational formulation	12
b	abstract bilinear form on $\mathcal{X} \times \mathcal{Y}$	12
B	Operator associated to the bilinear form b	12
B^*	Adjoint operator associated to the bilinear form b	12
γ	continuity constant of a bilinear form	12
β	inf-sup constant of a bilinear form	14
$\mathcal{X}^\delta$	Discrete subspace of $\mathcal{X}$ as trial space for Petrov-Galerkin projection	15
$\mathcal{Y}^\delta$	Discrete subspace of $\mathcal{Y}$ as test space for Petrov-Galerkin projection	15
Ω	(Full) domain	18, 64
$\partial\Omega$	Boundary of Ω	18
$\mathbf{n}$	Unit normal on $\partial\Omega$	18, 28
$\mathbf{b}$	Transport vector field	18, 28
c	Reaction coefficient, $c \in C^0(\bar{\Omega})$	28
Γ_-	Inflow boundary, part of $\partial\Omega$	18, 28, 64
Γ_+	Outflow boundary, part of $\partial\Omega$	18, 28, 64
Γ_0	Characteristic/no-flow boundary, part of $\partial\Omega$	18
$H(\Omega, \mathbf{b})$	Sobolev space of functions with directional derivative $\mathbf{b} \cdot \nabla u \in L^2$	18
$H_{\Gamma_\pm}(\Omega, \mathbf{b})$	Subspace of $H(\Omega, \mathbf{b})$ with zero trace on $\Gamma_\pm$	19
$L^2(K, \lvert\mathbf{b} \cdot \mathbf{n}\rvert)$	Weighted L^2 space on $K \subset \partial\Omega$	18, 33
$C^1_{\Gamma_\pm}(\Omega)$	C^1 functions that are zero on $\Gamma_\pm$	28
$\mathcal{Y}_\mathrm{t}$	Transport test space, closure of $C^1_{\Gamma_+}(\Omega)$ in $\lVert \cdot \rVert_{\mathcal{Y}_\mathrm{t}}$	33
b_t	Transport bilinear form on $\mathcal{X}_\mathrm{t} \times \mathcal{Y}_\mathrm{t}$, $b_\mathrm{t}(v, w) := (v, B^*_\mathrm{t} w)_{L^2(\Omega)}$	33
B^*_t	Adjoint transport operator $B^*_\mathrm{t} : \mathcal{Y}_\mathrm{t} \to L^2(\Omega), B^*_\mathrm{t} w = -\mathbf{b} \cdot \nabla w + w(c - \nabla \cdot \mathbf{b})$	33
γ_t	continuity constant of $b_\mathrm{t} : \mathcal{X}_\mathrm{t} \times \mathcal{Y}_\mathrm{t} \to \mathbb{R}$	34
β_t	inf-sup constant of $b_\mathrm{t} : \mathcal{X}_\mathrm{t} \times \mathcal{Y}_\mathrm{t} \to \mathbb{R}$	34
$\mathcal{X}^\delta_\mathrm{t}$	Optimal discrete trial space for transport, $\mathcal{X}^\delta_\mathrm{t} := B^*_\mathrm{t}(\mathcal{Y}^\delta_\mathrm{t}) \subset L^2(\Omega)$	36
$\mathcal{Y}^\delta_\mathrm{t}$	Discrete test space for transport $\mathcal{Y}^\delta_\mathrm{t} \subset \mathcal{Y}_\mathrm{t}$	36, 44
β^δ_t	discrete inf-sup constant of $b_\mathrm{t} : \mathcal{X}^\delta_\mathrm{t} \times \mathcal{Y}^\delta_\mathrm{t} \to \mathbb{R}$	35
e^δ	error of the discrete solution, $e^\delta := u - u^\delta$	35

r^δ	residual of the discrete solution, $r^\delta := f - b_\mathrm{t}(u^\delta, \cdot) = b_\mathrm{t}(e^\delta, \cdot) \in \mathcal{Y}_\mathrm{t}'$	35
$\sigma_{L^2(\Omega)}(u; \mathcal{X}_\mathrm{t}^\delta)$	L^2-best approximation error of u in $\mathcal{X}_\mathrm{t}^\delta$	36
$\mathcal{P}$	Set of parameters	42
$\mathcal{Y}_{\mathrm{t},\mu}$	Parametrized transport test space	42
b_μ	Parametrized transport bilinear form	42
B_μ^*	Parametrized adjoint transport operator	42
$\bar{\mathcal{Y}}_\mathrm{t}$	Intersection of all $\mathcal{Y}_{\mathrm{t},\mu}$, $\mu \in \mathcal{P}$	42
$\mathcal{Y}_{\mathrm{t},\mu}^\delta$	Discrete test space with parameter-dependent norm $\mathcal{Y}_\mathrm{t}^\delta = (\mathcal{Y}_\mathrm{t}^\delta, \|\cdot\|_{\mathcal{Y}_{\mathrm{t},\mu}}) \subset \mathcal{Y}_{\mathrm{t},\mu}$	44
$\mathcal{X}_{\mathrm{t},\mu}^\delta$	Optimal discrete trial space for parameter-dependent problem, $\mathcal{X}_{\mathrm{t},\mu}^\delta := B_{\mathrm{t},\mu}^*(\mathcal{Y}_{\mathrm{t},\mu}^\delta) \subset L^2(\Omega)$	44
Y^N	Reduced test space in $\mathcal{Y}_\mathrm{t}^\delta$	45
Y_μ^N	Y^N with norm $\|\cdot\|_{\mathcal{Y}_{\mathrm{t},\mu}}$	45
X_μ^N	Reduced trial space $X_\mu^N := B_\mu^*(Y_\mu^N) \subset \mathcal{X}_{\mathrm{t},\mu}^\delta$	45
$\Omega_{t,x}$	Space-time domain	64
Ω_v	Velocity domain $\Omega_v := S^{d-1}$	64
$W_\mathrm{nt}^p(\Omega)$	Banach space for neutron transport, $w \in L^p(\Omega)$ with $\binom{1}{v} \cdot \nabla_{t,x} w \in L^p(\Omega)$	21
$H_\mathrm{nt}(\Omega)$	Hilbert space for neutron transport, $H_\mathrm{nt}(\Omega) = W_\mathrm{nt}^2(\Omega)$	21
∇_Γ	Tangential gradient on the surface Γ	23
Δ_v	Laplace-Beltrami operator on Ω_v	23, 64
V	Velocity Sobolev space $V := H^1(\Omega_v) \subset L^2(\Omega_v)$	65
$H_\mathrm{fp}^1(\Omega)$	Fokker-Planck function space, $p \in L^2(\Omega_{t,x}; V)$ with $\binom{1}{v} \cdot \nabla_{t,x} p \in L^2(\Omega_{t,x}; V')$	65
$H_{\mathrm{fp},\Gamma_+}^1(\Omega)$	Closure of $C_{\Gamma_+}^1(\bar{\Omega})$ in $\|\cdot\|_{H_\mathrm{fp}^1(\Omega)}$	68
$L^2(\Gamma, \lvert(1,v)^T \cdot \mathbf{n}\rvert)$	Weighted L^2 space on $\Gamma \subset \partial\Omega$	66
a_v	Velocity bilinear form on $V \times V$	69
A_v	Operator corresponding to a_v	74
α_v	Coercivity constant of a_v	69
γ_v	Continuity constant of a_v	69
λ_v	Constant in Gårding-inequality of a_v	69
$\mathcal{X}_\mathrm{fp}$	Fokker-Planck trial space $\mathcal{X}_\mathrm{fp} := L^2(\Omega_{t,x}, V)$	70
$\mathcal{Y}_\mathrm{fp}$	Fokker-Planck test space $\mathcal{Y}_\mathrm{fp} := H_{\mathrm{fp},\Gamma_+}^1(\Omega)$	70
b_fp	Fokker-Planck bilinear form on $\mathcal{X}_\mathrm{fp} \times \mathcal{Y}_\mathrm{fp}$	70
β_fp	Inf-sup constant of $b_\mathrm{fp} : \mathcal{X}_\mathrm{fp}^\delta \times \mathcal{Y}_\mathrm{fp}^\delta \to \mathbb{R}$	71
w_p	Stable trial function in $\mathcal{X}_\mathrm{fp}$ for $p \in \mathcal{Y}_\mathrm{fp}$	72
$\hat{b}_\mathrm{fp}$	Transformed Fokker-Planck bilinear form	72
$\hat{a}_v$	Transformed velocity bilinear form	72
$\mathcal{Y}_\mathrm{fp}^\delta$	Fokker-Planck discrete test space in $\mathcal{Y}_\mathrm{fp}$	76
$\mathcal{X}_\mathrm{fp}^\delta$	Fokker-Planck discrete trial space $\mathcal{X}_\mathrm{fp}^\delta := \{p^\delta + z_p^\delta : p^\delta \in \mathcal{Y}_\mathrm{fp}^\delta\} \subset \mathcal{X}_\mathrm{fp}$	76
$\|\cdot\|_{L^2(\Omega_{t,x},(V^h)')}$	Discrete variant of $\|\cdot\|_{\mathcal{X}_\mathrm{fp}'}$	76
β_fp^δ	discrete inf-sup constant of $b_\mathrm{fp} : \mathcal{X}_\mathrm{fp}^\delta \times \mathcal{Y}_\mathrm{fp}^\delta \to \mathbb{R}$	77

List of Acronyms

CG	conjugate gradients	58
CPU	central processing unit	57
DDMRes	Discrete-Dual Minimal-Residual	5
DG	discontinuous Galerkin	5
DPG	Discontinuous Petrov-Galerkin	5
DTI	diffusion tensor imaging	11
ECM	extracellular matrix	9
FE	Finite Element	1
GNAT	Gauss-Newton with approximated tensors	6
LSFEM	Least-squares Finite Element method	4
LU	lower–upper decomposition	58
PDE	partial differential equation	1
PGD	proper generalized decomposition	8
RB	Reduced Basis	1
SDFEM	streamline diffusion Finite Element method	4
SPLS	saddle point least squares	5
SUPG	streamline upwind Petrov-Galerkin	4
UMFPACK	Unsymmetric MultiFrontal PACKage	58

Bibliography

[1] R. Abgrall, D. Amsallem, and R. Crisovan. Robust model reduction by L^1-norm minimization and approximation via dictionaries: application to nonlinear hyperbolic problems. *Advanced Modeling and Simulation in Engineering Sciences*, 3(1), 2016. https://doi.org/10.1186/s40323-015-0055-3. (Cited on page 6.)

[2] R. A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65. (Cited on pages 17 and 18.)

[3] S. Armstrong and J.-C. Mourrat. Variational methods for the kinetic Fokker-Planck equation, 2019. https://arxiv.org/abs/1902.04037. (Cited on pages 3, 7, 63, 65, 66, 67, 68, 69, 89, 90, and 92.)

[4] M. Asadzadeh and P. Kowalczyk. Convergence analysis of the streamline diffusion and discontinuous Galerkin methods for the Vlasov-Fokker-Planck system. *Numer. Methods Partial Differential Equations*, 21(3):472–495, 2005. https://doi.org/10.1002/num.20044. (Cited on page 8.)

[5] M. Asadzadeh and A. Sopasakis. Convergence of a *hp*-streamline diffusion scheme for Vlasov-Fokker-Planck system. *Math. Models Methods Appl. Sci.*, 17(8):1159–1182, 2007. https://doi.org/10.1142/S0218202507002236. (Cited on page 8.)

[6] P. Azérad. *Analyse des équations de Navier-Stokes en bassin peu profond et de l'équation de transport*. Ph.D. thesis, Université de Neuchâtel, Neuchâtel, Switzerland, 1996. (Cited on pages 28, 30, 31, 32, 35, and 74.)

[7] P. Azérad and J. Pousin. Inégalité de Poincaré courbe pour le traitement variationnel de l'équation de transport. *C. R. Acad. Sci. Paris Sér. I Math.*, 322(8):721–727, 1996. (Cited on pages 28, 30, and 31.)

[8] I. Babuška. Error-bounds for finite element method. *Numer. Math.*, 16:322–333, 1970/1971. https://doi.org/10.1007/BF02165003. (Cited on pages 12, 15, and 16.)

[9] C. Bacuta and K. Qirko. A saddle point least squares approach to mixed methods. *Comput. Math. Appl.*, 70(12):2920–2932, 2015. https://doi.org/10.1016/j.camwa.2015.10.001. (Cited on pages 5 and 41.)

[10] G. Bal and B. Palacios. Pencil-beam approximation of stationary Fokker-Planck. *SIAM J. Math. Anal.*, 52(4):3487–3519, 2020. https://doi.org/10.1137/19M1295775. (Cited on pages 7, 65, 69, 74, 89, 90, and 92.)

[11] F. Ballarin, A. Manzoni, A. Quarteroni, and G. Rozza. Supremizer stabilization of POD-Galerkin approximation of parametrized steady incompressible Navier-Stokes equations. *Internat. J. Numer. Methods Engrg.*, 102(5):1136–1161, 2015. https://doi.org/10.1002/nme.4772. (Cited on page 6.)

[12] C. Bardos. Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport. *Ann. Sci. École Norm. Sup. (4)*, 3:185–233, 1970. (Cited on pages 17, 19, 20, 73, 89, and 90.)

[13] M. Barrault, Y. Maday, N. Nguyen, and A. Patera. An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. *C. R. Math. Acad. Sci. Paris Series I*, 339:667–672, 2004. https://doi.org/10.1016/j.crma.2004.08.006. (Cited on page 48.)

[14] P. Benner, S. Gugercin, and K. Willcox. A survey of projection-based model reduction methods for parametric dynamical systems. *SIAM Rev.*, 57(4):483–531, 2015. https://doi.org/10.1137/130932715. (Cited on page 6.)

[15] P. Benner, M. Ohlberger, A. Cohen, and K. Willcox, editors. *Model Reduction and Approximation*. Society for Industrial and Applied Mathematics, 2017. https://doi.org/10.1137/1.9781611974829. (Cited on page 6.)

[16] 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. https://doi.org/10.1016/0045-7825(95)00944-2. (Cited on page 5.)

[17] M. Billaud-Friess, A. Nouy, and O. Zahm. A tensor approximation method based on ideal minimal residual formulations for the solution of high-dimensional problems. *ESAIM Math. Model. Numer. Anal.*, 48(6):1777–1806, 2014. https://doi.org/10.1051/m2an/2014019. (Cited on page 79.)

[18] P. Bochev and M. Gunzburger. Least-squares methods for hyperbolic problems. In *Handbook of numerical methods for hyperbolic problems*, *Handb. Numer. Anal.*, volume 17, pp. 289–317. Elsevier/North-Holland, Amsterdam, 2016. https://doi.org/10.1016/bs.hna.2016.07.002. (Cited on page 4.)

[19] P. B. Bochev and J. Choi. Improved least-squares error estimates for scalar hyperbolic problems. *Comput. Methods Appl. Math.*, 1(2):115–124, 2001. https://doi.org/10.2478/cmam-2001-0008. (Cited on pages 4 and 40.)

[20] P. B. Bochev and M. D. Gunzburger. *Least-squares finite element methods*, *Applied Mathematical Sciences*, volume 166. Springer, New York, 2009. https://doi.org/10.1007/b13382. (Cited on pages 4 and 40.)

[21] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011. https://doi.org/10.1007/978-0-387-70914-7. (Cited on pages 13 and 14.)

[22] D. Broersen, W. Dahmen, and R. P. Stevenson. On the stability of DPG formulations of transport equations. *Math. Comp.*, 87(311):1051–1082, 2018. https://doi.org/10.1090/mcom/3242. (Cited on page 5.)

[23] A. N. Brooks and T. J. R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32(1-3):199–259, 1982. https://doi.org/10.1016/0045-7825(82)90071-8. FENOMECH ”81, Part I (Stuttgart, 1981). (Cited on page 4.)

[24] J. Brunken. Source code to “(Parametrized) first order transport equations: Realization of optimally stable Petrov-Galerkin methods”, 2018. https://doi.org/10.5281/zenodo.1413553. (Cited on page 52.)

[25] J. Brunken. Source code to “Stable and efficient Petrov-Galerkin methods for a kinetic Fokker-Planck equation”, 2020. https://doi.org/10.5281/zenodo.4106757. (Cited on page 79.)

[26] J. Brunken, T. Leibner, M. Ohlberger, and K. Smetana. Problem adapted hierachical model reduction for the Fokker-Planck equation. In A. Handlovičova and D. Sevčovič, editors, *Proceedings of ALGORITMY 2016, the 20th Conference on Scientific Computing (Vysoke Tatry, Podbanske, Slovakia, 2016)*, pp. 13–22. Publishing House of Slovak University of Technology in Bratislava, 2016. (Cited on pages 7, 79, and 87.)

[27] J. Brunken and K. Smetana. Stable and efficient Petrov-Galerkin methods for a kinetic Fokker-Planck equation, 2020. https://arxiv.org/abs/2010.15784. (Cited on page 64.)

[28] J. Brunken, K. Smetana, and K. Urban. (Parametrized) first order transport equations: realization of optimally stable Petrov-Galerkin methods. *SIAM J. Sci. Comput.*, 41(1):A592–A621, 2019. https://doi.org/10.1137/18M1176269. (Cited on pages 28, 35, and 63.)

[29] T. Bui-Thanh, L. Demkowicz, and O. Ghattas. Constructively well-posed approximation methods with unity inf-sup and continuity constants for partial differential equations. *Math. Comp.*, 82(284):1923–1952, 2013. https://doi.org/10.1090/S0025-5718-2013-02697-X. (Cited on pages 2, 5, 27, 40, and 85.)

[30] M. J. Cáceres, J. A. Carrillo, and L. Tao. A numerical solver for a nonlinear Fokker-Planck equation representation of neuronal network dynamics. *J. Comput. Phys.*, 230(4):1084–1099, 2011. https://doi.org/10.1016/j.jcp.2010.10.027. (Cited on page 7.)

[31] K. Carlberg. Adaptive h-refinement for reduced-order models. *Internat. J. Numer. Methods Engrg.*, 102(5):1192–1210, 2015. https://doi.org/10.1002/nme.4800. (Cited on page 7.)

[32] K. Carlberg, C. Bou-Mosleh, and C. Farhat. Efficient non-linear model reduction via a least-squares Petrov-Galerkin projection and compressive tensor approximations. *Internat. J. Numer. Methods Engrg.*, 86(2):155–181, 2011. https://doi.org/10.1002/nme.3050. (Cited on page 6.)

[33] K. Carlberg, C. Farhat, J. Cortial, and D. Amsallem. The GNAT method for nonlinear model reduction: effective implementation and application to computational fluid dynamics and turbulent flows. *J. Comput. Phys.*, 242:623–647, 2013. https://doi.org/10.1016/j.jcp.2013.02.028. (Cited on page 6.)

[34] J. A. Carrillo. Global weak solutions for the initial-boundary-value problems to the Vlasov-Poisson-Fokker-Planck system. *Math. Methods Appl. Sci.*, 21(10):907–938, 1998. https://doi.org/10.1002/(SICI)1099-1476(19980710)21:10<907::AID-MMA977>3.3.CO;2-N. (Cited on pages 3, 7, 63, 65, 66, 69, 89, 90, and 92.)

[35] M. Cessenat. Théorèmes de trace L^p pour des espaces de fonctions de la neutronique. *C. R. Acad. Sci. Paris Sér. I Math.*, 299(16):831–834, 1984. (Cited on pages 21, 73, and 89.)

[36] M. Cessenat. Théorèmes de trace pour des espaces de fonctions de la neutronique. *C. R. Acad. Sci. Paris Sér. I Math.*, 300(3):89–92, 1985. (Cited on pages 21, 22, 73, and 89.)

[37] 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. https://doi.org/10.1051/m2an/2012003. (Cited on pages 5 and 6.)

[38] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015. https://doi.org/10.1017/S0962492915000033. (Cited on page 6.)

[39] G. Corbin. *Numerical methods for multi-scale cell migration models*. Ph.D. thesis, Technische Universität Kaiserslautern, 2020. http://nbn-resolving.de/urn:nbn:de:hbz:386-kluedo-61258. (Cited on page 11.)

[40] G. Corbin, A. Hunt, A. Klar, F. Schneider, and C. Surulescu. Higher-order models for glioma invasion: from a two-scale description to effective equations for mass density and momentum. *Math. Models Methods Appl. Sci.*, 28(9):1771–1800, 2018. https://doi.org/10.1142/S0218202518400055. (Cited on pages 9 and 11.)

[41] G. Corbin, A. Klar, C. Surulescu, C. Engwer, M. Wenske, J. Nieto, and J. Soler. Modeling glioma invasion with anisotropy- and hypoxia-triggered motility enhancement: From subcellular dynamics to macroscopic PDEs with multiple taxis. *Math. Models Methods Appl. Sci.*, 31(1):177–222, 2021. https://doi.org/10.1142/S0218202521500056. (Cited on pages 1, 10, and 11.)

[42] W. Dahmen, F. Gruber, and O. Mula. An adaptive nested source term iteration for radiative transfer equations. *Math. Comp.*, 89(324):1605–1646, 2020. https://doi.org/10.1090/mcom/3505. (Cited on pages 5 and 8.)

[43] 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. https://doi.org/10.1137/110823158. (Cited on pages 2, 3, 5, 6, 27, 28, 29, 30, 31, 33, 34, 40, 41, 48, 57, 58, 63, 71, and 85.)

[44] W. Dahmen, G. Kutyniok, W.-Q. Lim, C. Schwab, and G. Welper. Adaptive anisotropic Petrov-Galerkin methods for first order transport equations. *J. Comput. Appl. Math.*, 340:191–220, 2018. https://doi.org/10.1016/j.cam.2018.02.023. (Cited on page 5.)

[45] W. Dahmen, C. Plesken, and G. Welper. Double greedy algorithms: reduced basis methods for transport dominated problems. *ESAIM Math. Model. Numer. Anal.*, 48(3):623–663, 2014. https://doi.org/10.1051/m2an/2013103. (Cited on pages 3, 5, 6, 27, 28, 42, 43, 45, 46, 48, 49, 59, and 86.)

[46] W. Dahmen and R. P. Stevenson. Adaptive strategies for transport equations. *Comput. Methods Appl. Math.*, 19(3):431–464, 2019. https://doi.org/10.1515/cmam-2018-0230. (Cited on page 5.)

[47] R. Dautray and J.-L. Lions. *Mathematical analysis and numerical methods for science and technology. Vol. 6.* Springer-Verlag, Berlin, 1993. https://doi.org/10.1007/978-3-642-58004-8. Evolution problems. II, With the collaboration of Claude Bardos, Michel Cessenat, Alain Kavenoky, Patrick Lascaux, Bertrand Mercier, Olivier Pironneau, Bruno Scheurer and Rémi Sentis, Translated from the French by Alan Craig. (Cited on pages 21, 67, 73, and 89.)

[48] H. De Sterck, T. A. Manteuffel, S. F. McCormick, and L. Olson. Least-squares finite element methods and algebraic multigrid solvers for linear hyperbolic PDEs. *SIAM J. Sci. Comput.*, 26(1):31–54, 2004. https://doi.org/10.1137/s106482750240858x. (Cited on pages 4 and 40.)

[49] P. Degond and S. Mas-Gallic. Existence of solutions and diffusion approximation for a model Fokker-Planck equation. *Transport Theory Statist. Phys.*, 16(4-6):589–636, 1987. https://doi.org/10.1080/00411458708204307. (Cited on pages 7 and 90.)

[50] P. Degond, L. Pareschi, and G. Russo, editors. *Modeling and Computational Methods for Kinetic Equations.* Birkhäuser Boston, 2004. https://doi.org/10.1007/978-0-8176-8200-2. (Cited on page 7.)

[51] 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. https://doi.org/10.1016/j.cma.2010.01.003. (Cited on pages 2, 5, 27, 40, and 85.)

[52] L. 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. https://doi.org/10.1002/num.20640. (Cited on pages 2, 5, 27, 40, 63, 71, and 85.)

[53] A. Dietrich, N. Kolbe, N. Sfakianakis, and C. Surulescu. Multiscale modeling of glioma invasion: from receptor binding to flux-limited macroscopic PDEs, 2020. https://arxiv.org/abs/2010.03277. (Cited on pages 10 and 11.)

[54] G. Dimarco and L. Pareschi. Numerical methods for kinetic equations. *Acta Numer.*, 23:369–520, 2014. https://doi.org/10.1017/S0962492914000063. (Cited on page 7.)

[55] W. Dörfler, S. Findeisen, and C. Wieners. Space-time discontinuous Galerkin discretizations for linear first-order hyperbolic evolution systems. *Comput. Meth. in Appl. Math.*, 16(3):409–428, 2016. https://doi.org/10.1515/cmam-2016-0015. (Cited on page 5.)

[56] G. Dziuk and C. M. Elliott. Finite element methods for surface PDEs. *Acta Numerica*, 22:289–396, 2013. https://doi.org/10.1017/S0962492913000056. (Cited on pages 22, 23, 24, and 25.)

[57] H. Egger and M. Schlottbom. A mixed variational framework for the radiative transfer equation. *Math. Models Methods Appl. Sci.*, 22(3):1150014, 30, 2012. https://doi.org/10.1142/S021820251150014X. (Cited on pages 7 and 8.)

[58] H. Egger and M. Schlottbom. Stationary radiative transfer with vanishing absorption. *Math. Models Methods Appl. Sci.*, 24(5):973–990, 2014. https://doi.org/10.1142/S0218202513500735. (Cited on pages 7 and 8.)

[59] V. Ehrlacher and D. Lombardi. A dynamical adaptive tensor method for the Vlasov-Poisson system. *J. Comput. Phys.*, 339:285–306, 2017. https://doi.org/10.1016/j.jcp.2017.03.015. (Cited on pages 8 and 79.)

[60] L. Einkemmer and C. Lubich. A low-rank projector-splitting integrator for the Vlasov-Poisson equation. *SIAM J. Sci. Comput.*, 40(5):B1330–B1360, 2018. https://doi.org/10.1137/18M116383X. (Cited on pages 8 and 79.)

[61] C. Engwer, T. Hillen, M. Knappitsch, and C. Surulescu. Glioma follow white matter tracts: a multiscale DTI-based model. *J. Math. Biol.*, 71(3):551–582, 2015. https://doi.org/10.1007/s00285-014-0822-7. (Cited on pages 1, 9, 10, and 11.)

[62] C. Engwer, A. Hunt, and C. Surulescu. Effective equations for anisotropic glioma spread with proliferation: a multiscale approach and comparisons with previous settings. *Mathematical Medicine and Biology: A Journal of the IMA*, 33(4):435–459, 2015. https://doi.org/10.1093/imammb/dqv030. (Cited on pages 1, 9, 10, and 11.)

[63] C. Engwer, M. Knappitsch, and C. Surulescu. A multiscale model for glioma spread including cell-tissue interactions and proliferation. *Math. Biosci. Eng.*, 13(2):443–460, 2016. https://doi.org/10.3934/mbe.2015011. (Cited on pages 10 and 11.)

[64] C. Engwer and M. Wenske. Estimating the extent of glioblastoma invasion. *J. Math. Biol.*, 82(1-2):10, 2021. https://doi.org/10.1007/s00285-021-01563-9. (Cited on page 1.)

[65] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements.* Applied Mathematical Sciences. Springer New York, 2004. https://doi.org/10.1007/978-1-4757-4355-5. (Cited on pages 3, 12, 15, 29, 63, 71, and 74.)

[66] L. C. Evans. *Partial differential equations*, *Graduate Studies in Mathematics*, volume 19. American Mathematical Society, Providence, RI, 1998. (Cited on pages 17 and 20.)

[67] M. Frank, H. Hensel, and A. Klar. A fast and accurate moment method for the Fokker-Planck equation and applications to electron radiotherapy. *SIAM J. Appl. Math.*, 67(2):582–603, 2006/07. https://doi.org/10.1137/06065547X. (Cited on page 7.)

[68] J.-F. Gerbeau and D. Lombardi. Approximated Lax pairs for the reduced order integration of nonlinear evolution equations. *J. Comput. Phys.*, 265:246–269, 2014. https://doi.org/10.1016/j.jcp.2014.01.047. (Cited on page 7.)

[69] T. A. Germogenova. Generalized solutions of boundary value problems for the transport equation. *Ž. Vyčisl. Mat i Mat. Fiz.*, 9:605–625, 1969. (Cited on page 21.)

[70] A.-L. Gerner and K. Veroy. Certified reduced basis methods for parametrized saddle point problems. *SIAM J. Sci. Comput.*, 34(5):A2812–A2836, 2012. https://doi.org/10.1137/110854084. (Cited on page 6.)

[71] G. Geymonat and P. Leyland. Transport and propagation of a perturbation of a flow of a compressible fluid in a bounded region. *Arch. Rational Mech. Anal.*, 100(1):53–81, 1987. https://doi.org/10.1007/BF00281247. (Cited on pages 17, 18, 19, and 20.)

[72] M. B. Giles and E. Süli. Adjoint methods for PDEs: a posteriori error analysis and postprocessing by duality. *Acta Numerica*, 11:145–236, 2002. https://doi.org/10.1017/s096249290200003x. (Cited on page 29.)

[73] S. Glas, A. T. Patera, and K. Urban. A reduced basis method for the wave equation. *Int. J. Comput. Fluid Dyn.*, 34(2):139–146, 2020. https://doi.org/10.1080/10618562.2019.1686486. (Cited on page 6.)

[74] K. Grella and C. Schwab. Sparse tensor spherical harmonics approximation in radiative transfer. *J. Comput. Phys.*, 230(23):8452–8473, 2011. https://doi.org/10.1016/j.jcp.2011.07.028. (Cited on pages 8 and 79.)

[75] M. A. Grepl and A. T. Patera. A posteriori error bounds for reduced-bias approximations of parametrized parabolic partial differential equations. *ESAIM Math. Model. Numer. Anal.*, 39(1):157–181, 2005. https://doi.org/10.1051/m2an:2005006. (Cited on page 6.)

[76] B. Haasdonk. Reduced basis methods for parametrized PDEs—a tutorial introduction for stationary and instationary problems. In *Model reduction and approximation*, *Comput. Sci. Eng.*, volume 15, pp. 65–136. SIAM, Philadelphia, PA, 2017. https://doi.org/10.1137/1.9781611974829.ch2. (Cited on pages 3, 6, and 42.)

[77] B. Haasdonk and M. Ohlberger. Reduced basis method for finite volume approximations of parametrized linear evolution equations. *ESAIM Math. Model. Numer. Anal.*, 42(2):277–302, 2008. https://doi.org/10.1051/m2an:2008001. (Cited on page 6.)

[78] B. Haasdonk and M. Ohlberger. Reduced basis method for explicit finite volume approximations of nonlinear conservation laws. In *Hyperbolic problems: theory,*

numerics and applications, *Proc. Sympos. Appl. Math.*, volume 67, pp. 605–614. Amer. Math. Soc., Providence, RI, 2009. https://doi.org/10.1090/psapm/067.2/2605256. (Cited on page 6.)

[79] S. Hain, M. Ohlberger, M. Radic, and K. Urban. A hierarchical a posteriori error estimator for the reduced basis method. *Adv. Comput. Math.*, 45(5-6):2191–2214, 2019. https://doi.org/10.1007/s10444-019-09675-z. (Cited on pages 48 and 62.)

[80] W. Han, Y. Li, Q. Sheng, and J. Tang. A numerical method for generalized Fokker-Planck equations. In *Recent advances in scientific computing and applications*, *Contemp. Math.*, volume 586, pp. 171–179. Amer. Math. Soc., Providence, RI, 2013. https://doi.org/10.1090/conm/586/11649. (Cited on page 8.)

[81] E. Hebey. *Nonlinear Analysis on Manifolds: Sobolev Spaces and Inequalities.* Courant lecture notes in mathematics. Courant Institute of Mathematical Sciences, 2000. (Cited on pages 22 and 25.)

[82] J. Henning, D. Palitta, V. Simoncini, and K. Urban. Matrix oriented reduction of space-time Petrov-Galerkin variational problems, 2019. https://arxiv.org/abs/1912.10082. (Cited on pages 6 and 79.)

[83] J. S. Hesthaven, G. Rozza, and B. Stamm. *Certified reduced basis methods for parametrized partial differential equations.* SpringerBriefs in Mathematics. Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao, 2016. https://doi.org/10.1007/978-3-319-22470-1. BCAM SpringerBriefs. (Cited on pages 3, 6, and 42.)

[84] P. Houston, C. Schwab, and E. Süli. Stabilized *hp*-finite element methods for first-order hyperbolic problems. *SIAM J. Numer. Anal.*, 37(5):1618–1643, 2000. https://doi.org/10.1137/S0036142998348777. (Cited on page 5.)

[85] P. Houston, C. Schwab, and E. Süli. Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 39(6):2133–2163, 2002. https://doi.org/10.1137/S0036142900374111. (Cited on page 8.)

[86] 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. https://doi.org/10.1137/S1064827500378799. (Cited on page 5.)

[87] P. Houston and E. Süli. Stabilised *hp*-finite element approximation of partial differential equations with nonnegative characteristic form. *Computing*, 66(2):99–119, 2001. https://doi.org/10.1007/s006070170030. (Cited on page 8.)

[88] 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)*, *AMD*, volume 34, pp. 19–35. Amer. Soc. Mech. Engrs. (ASME), New York, 1979. (Cited on page 4.)

[89] T. J. R. Hughes, L. P. Franca, and G. M. Hulbert. A new finite element formulation for computational fluid dynamics. VIII. The Galerkin/least-squares method for

advective-diffusive equations. *Comput. Methods Appl. Mech. Engrg.*, 73(2):173–189, 1989. https://doi.org/10.1016/0045-7825(89)90111-4. (Cited on page 4.)

[90] A. Hunt. *DTI-Based Multiscale Models for Glioma Invasion.* Ph.D. thesis, TU Kaiserslautern, 2017. https://nbn-resolving.org/urn:nbn:de:hbz:386-kluedo-53575. (Cited on pages 1, 9, 10, 11, and 64.)

[91] A. Hunt and C. Surulescu. A multiscale modeling approach to glioma invasion with therapy. *Vietnam J. Math.*, 45(1-2):221–240, 2017. https://doi.org/10.1007/s10013-016-0223-x. (Cited on page 11.)

[92] H. J. Hwang, J. Jang, and J. Jung. The Fokker-Planck equation with absorbing boundary conditions in bounded domains. *SIAM J. Math. Anal.*, 50(2):2194–2232, 2018. https://doi.org/10.1137/16M1109928. (Cited on page 7.)

[93] A. Iollo and D. Lombardi. Advection modes by optimal mass transfer. *Physical Review E*, 89(022923), 2014. https://doi.org/10.1103/physreve.89.022923. (Cited on page 7.)

[94] S. Karimghasemi. *Convergence of Approximate Solutions to the Transport Equation.* Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, 2020. (Cited on page 5.)

[95] J. Kelkel and C. Surulescu. A multiscale approach to cell migration in tissue networks. *Math. Models Methods Appl. Sci.*, 22(3):1150017, 25, 2012. https://doi.org/10.1142/S0218202511500175. (Cited on pages 1, 9, and 10.)

[96] A. D. Kim and P. Tranquilli. Numerical solution of the Fokker-Planck equation with variable coefficients. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 109(5):727–740, 2008. https://doi.org/10.1016/j.jqsrt.2007.09.011. (Cited on page 7.)

[97] A. Klar, F. Schneider, and O. Tse. Approximate models for stochastic dynamic systems with velocities on the sphere and associated Fokker-Planck equations. *Kinet. Relat. Models*, 7(3):509–529, 2014. https://doi.org/10.3934/krm.2014.7.509. (Cited on page 7.)

[98] K. Kormann. A semi-Lagrangian Vlasov solver in tensor train format. *SIAM J. Sci. Comput.*, 37(4):B613–B632, 2015. https://doi.org/10.1137/140971270. (Cited on pages 8 and 79.)

[99] O. Lehtikangas, T. Tarvainen, V. Kolehmainen, A. Pulkkinen, S. Arridge, and J. Kaipio. Finite element approximation of the Fokker-Planck equation for diffuse optical tomography. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 111(10):1406 – 1417, 2010. https://doi.org/10.1016/j.jqsrt.2010.03.003. (Cited on page 7.)

[100] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I.* Springer-Verlag, New York-Heidelberg, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181. (Cited on page 71.)

[101] Q. Liu and S. Zhang. Adaptive least-squares finite element methods for linear transport equations based on an $H(div)$ flux reformulation. *Comput. Methods Appl. Mech. Engrg.*, 366:113041, 25, 2020. https://doi.org/10.1016/j.cma.2020.113041. (Cited on page 4.)

[102] T. A. Manteuffel, K. J. Ressel, and G. Starke. A boundary functional for the least-squares finite-element solution of neutron transport problems. *SIAM J. Numer. Anal.*, 37(2):556–586, 2000. https://doi.org/10.1137/S0036142998344706. (Cited on pages 4, 8, 21, 22, 91, and 92.)

[103] I. Muga, M. J. W. Tyler, and K. G. van der Zee. The discrete-dual minimal-residual method (DDMRes) for weak advection-reaction problems in Banach spaces. *Comput. Methods Appl. Math.*, 19(3):557–579, 2019. https://doi.org/10.1515/cmam-2018-0199. (Cited on page 5.)

[104] J. Nečas. Sur une méthode pour résoudre les équations aux dérivées partielles du type elliptique, voisine de la variationnelle. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)*, 16:305–326, 1962. http://www.numdam.org/item/ASNSP_1962_3_16_4_305_0. (Cited on pages 12 and 15.)

[105] R. Nochetto, K. Siebert, and A. Veeser. Theory of adaptive finite element methods: An introduction. In R. DeVore and A. Kunoth, editors, *Multiscale, nonlinear and adaptive approximation*, pp. 409–542. Springer, Berlin, 2009. https://doi.org/10.1007/978-3-642-03413-8_12. (Cited on page 12.)

[106] M. Ohlberger and S. Rave. Nonlinear reduced basis approximation of parameterized evolution equations via the method of freezing. *C. R. Math. Acad. Sci. Paris*, 351(23-24):901–906, 2013. https://doi.org/10.1016/j.crma.2013.10.028. (Cited on page 7.)

[107] M. Ohlberger and S. Rave. Reduced basis methods: Success, limitations and future challenges. In Handlovičova, A. and Sevčovič, D., editor, *Proceedings of ALGORITMY 2016, the 20th Conference on Scientific Computing (Vysoke Tatry, Podbanske, Slovakia, 2016)*, pp. 1–12. Publishing House of Slovak University of Technology in Bratislava, 2016. (Cited on pages 7 and 59.)

[108] M. Ohlberger and K. Smetana. Approximation of skewed interfaces with tensor-based model reduction procedures: application to the reduced basis hierarchical model reduction approach. *J. Comput. Phys.*, 321:1185–1205, 2016. https://doi.org/10.1016/j.jcp.2016.06.021. (Cited on page 7.)

[109] P. Pacciarini and G. Rozza. Stabilized reduced basis method for parametrized advection-diffusion PDEs. *Comput. Methods Appl. Mech. Engrg.*, 274:1–18, 2014. https://doi.org/10.1016/j.cma.2014.02.005. (Cited on page 6.)

[110] P. Perrochet and P. Azérad. Space-time integrated least-squares: solving a pure advection equation with a pure diffusion operator. *J. Comput. Phys.*, 117(2):183–193, 1995. https://doi.org/10.1006/jcph.1995.1057. (Cited on page 4.)

[111] J. Qiu and C.-W. Shu. A comparison of troubled-cell indicators for Runge–Kutta discontinuous Galerkin methods using weighted essentially nonoscillatory limiters. *SIAM J. Sci. Comput.*, 27(3):995–1013, 2005. https://doi.org/10.1137/04061372x. (Cited on page 39.)

[112] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced basis methods for partial differential equations*, *Unitext*, volume 92. Springer, Cham, 2016. https://doi.org/10.1007/978-3-319-15431-2. An introduction, La Matematica per il 3+2. (Cited on pages 3, 6, and 42.)

[113] J. Reiss, P. Schulze, J. Sesterhenn, and V. Mehrmann. The shifted proper orthogonal decomposition: a mode decomposition for multiple transport phenomena. *SIAM J. Sci. Comput.*, 40(3):A1322–A1344, 2018. https://doi.org/10.1137/17M1140571. (Cited on page 7.)

[114] G. Rozza, D. B. P. Huynh, and A. Manzoni. Reduced basis approximation and a posteriori error estimation for Stokes flows in parametrized geometries: roles of the inf-sup stability constants. *Numer. Math.*, 125(1):115–152, 2013. https://doi.org/10.1007/s00211-013-0534-8. (Cited on page 6.)

[115] G. Rozza and K. Veroy. On the stability of the reduced basis method for stokes equations in parametrized domains. *Comput. Methods in Appl. Mech. and Engrg.*, 196(7):1244–1260, 2007. https://doi.org/10.1016/j.cma.2006.09.005. (Cited on page 6.)

[116] J. Schaeffer. Convergence of a difference scheme for the Vlasov-Poisson-Fokker-Planck system in one dimension. *SIAM J. Numer. Anal.*, 35(3):1149–1175, 1998. https://doi.org/10.1137/S0036142996302554. (Cited on page 8.)

[117] F. Schneider, G. Alldredge, M. Frank, and A. Klar. Higher order mixed-moment approximations for the Fokker-Planck equation in one space dimension. *SIAM J. Appl. Math.*, 74(4):1087–1114, 2014. https://doi.org/10.1137/130934210. (Cited on page 7.)

[118] F. Schneider, A. Roth, and J. Kall. First-order quarter- and mixed-moment realizability theory and Kershaw closures for a Fokker-Planck equation in two space dimensions. *Kinet. Relat. Models*, 10(4):1127–1161, 2017. https://doi.org/10.3934/krm.2017044. (Cited on page 7.)

[119] C. Schwab and R. Stevenson. Space-time adaptive wavelet methods for parabolic evolution problems. *Math. Comp.*, 78(267):1293–1318, 2009. https://doi.org/10.1090/s0025-5718-08-02205-9. (Cited on pages 3, 6, 63, 71, 72, 73, and 86.)

[120] C. Schwab, E. Süli, and R. A. Todor. Sparse finite element approximation of high-dimensional transport-dominated diffusion problems. *ESAIM Math. Model. Numer. Anal.*, 42(5):777–819, 2008. https://doi.org/10.1051/m2an:2008027. (Cited on page 8.)

[121] Q. Sheng and W. Han. Well-posedness of the Fokker-Planck equation in a scattering process. *J. Math. Anal. Appl.*, 406(2):531–536, 2013. https://doi.org/10.1016/j.jmaa.2013.04.063. (Cited on page 7.)

[122] C.-W. Shu. Discontinuous Galerkin method for time-dependent problems: Survey and recent developments. In X. Feng, O. Karakashian, and Y. Xing, editors, *Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations: 2012 John H Barrett Memorial Lectures*, pp. 25–62. Springer, Cham, 2014. https://doi.org/10.1007/978-3-319-01818-8_2. (Cited on pages 5 and 39.)

[123] K. Smetana and M. Ohlberger. Hierarchical model reduction of nonlinear partial differential equations based on the adaptive empirical projection method and reduced basis techniques. *ESAIM Math. Model. Numer. Anal.*, 51(2):641–677, 2017. https://doi.org/10.1051/m2an/2016031. (Cited on page 47.)

[124] E. Süli, C. Schwab, and P. Houston. *hp*-DGFEM for partial differential equations with nonnegative characteristic form. In *Discontinuous Galerkin methods (Newport, RI, 1999)*, *Lect. Notes Comput. Sci. Eng.*, volume 11, pp. 221–230. Springer, Berlin, 2000. https://doi.org/10.1007/978-3-642-59721-3_16. (Cited on page 8.)

[125] T. Taddei. A registration method for model order reduction: data compression and geometry reduction. *SIAM J. Sci. Comput.*, 42(2):A997–A1027, 2020. https://doi.org/10.1137/19M1271270. (Cited on page 7.)

[126] T. Taddei, S. Perotto, and A. Quarteroni. Reduced basis techniques for nonlinear conservation laws. *ESAIM Math. Model. Numer. Anal.*, 49(3):787–814, 2015. https://doi.org/10.1051/m2an/2014054. (Cited on page 7.)

[127] T. Taddei and L. Zhang. Space-time registration-based model reduction of parameterized one-dimensional hyperbolic PDEs. *ESAIM Math. Model. Numer. Anal.*, 55(1):99–130, 2021. https://doi.org/10.1051/m2an/2020073. (Cited on page 7.)

[128] K. Urban and A. Patera. A new error bound for reduced basis approximation of parabolic partial differential equations. *C. R. Math. Acad. Sci. Paris*, 350(3-4):203–207, 2012. https://doi.org/10.21236/ada557547. (Cited on page 6.)

[129] K. Urban and A. Patera. An improved error bound for reduced basis approximation of linear parabolic problems. *Math. Comp.*, 83(288):1599–1615, 2014. https://doi.org/10.1090/s0025-5718-2013-02782-2. (Cited on pages 3, 6, 63, and 72.)

[130] G. Welper. *Infinite dimensional stabilization of convection-dominated problems.* Ph.D. thesis, Rheinisch-Westfälische Technische Hochschule Aachen, Aachen, Germany, 2013. https://nbn-resolving.org/urn:nbn:de:hbz:82-opus-45352. (Cited on page 32.)

[131] G. Welper. *h* and *hp*-adaptive interpolation by transformed snapshots for parametric and stochastic hyperbolic PDEs, 2017. https://arxiv.org/abs/1710.11481. (Cited on page 7.)

[132] G. Welper. Interpolation of functions with parameter dependent jumps by transformed snapshots. *SIAM J. Sci. Comput.*, 39(4):A1225–A1250, 2017. https://doi.org/10.1137/16m1059904. (Cited on page 7.)

[133] G. Welper. Transformed snapshot interpolation with high resolution transforms. *SIAM J. Sci. Comput.*, 42(4):A2037–A2061, 2020. https://doi.org/10.1137/19M126356X. (Cited on page 7.)

[134] G. Widmer, R. Hiptmair, and C. Schwab. Sparse adaptive finite elements for radiative transfer. *J. Comput. Phys.*, 227(12):6071–6105, 2008. https://doi.org/10.1016/j.jcp.2008.02.025. (Cited on pages 8 and 79.)

[135] S. Wollman and E. Ozizmir. Numerical approximation of the Vlasov–Poisson–Fokker–Planck system in two dimensions. *J. Comput. Phys.*, 228(18):6629 – 6669, 2009. https://doi.org/10.1016/j.jcp.2009.05.027. (Cited on page 8.)

[136] J. Xu and L. Zikatanov. Some observations on Babuška and Brezzi theories. *Numer. Math.*, 94(1):195–202, 2003. https://doi.org/10.1007/s002110100308. (Cited on pages 16 and 17.)

[137] M. Yano. A space-time Petrov-Galerkin certified reduced basis method: application to the Boussinesq equations. *SIAM J. Sci. Comput.*, 36(1):A232–A266, 2014. https://doi.org/10.1137/120903300. (Cited on page 6.)

[138] M. Yano, A. Patera, and K. Urban. A space-time *hp*-interpolation-based certified reduced basis method for Burgers' equation. *Math. Models Methods Appl. Sci.*, 24(9):1903–1935, 2014. https://doi.org/10.1142/s0218202514500110. (Cited on page 6.)

[139] O. Zahm and A. Nouy. Interpolation of inverse operators for preconditioning parameter-dependent equations. *SIAM J. Sci. Comput.*, 38(2):A1044–A1074, 2016. https://doi.org/10.1137/15m1019210. (Cited on pages 3 and 6.)

[140] 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. https://doi.org/10.1016/j.jcp.2010.12.001. (Cited on page 5.)