

On a Family of Variational Time Discretization Methods

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt

dem Bereich Mathematik und Naturwissenschaften
der Technischen Universität Dresden

von

M.Sc. Simon Becher

geboren am 04.06.1991 in Lichtenstein

Gutachter: Prof. Dr. rer. nat. Gunar Matthies
Prof. Dr. Alexandre Ern

Eingereicht am: 19.05.2022
Tag der Disputation: 02.09.2022

Die Disseration wurde am Institut für Numerische Mathematik angefertigt.

Contents

Acknowledgments	v
List of Symbols and Abbreviations	vii

Introduction	1
--------------	---

I Variational Time Discretization Methods for Initial Value Problems	5
--	---

1 Formulation, Analysis for Non-Stiff Systems, and Further Properties	7
1.1 Formulation of the methods	8
1.1.1 Global formulation	9
1.1.2 Another formulation	10
1.2 Existence, uniqueness, and error estimates	11
1.2.1 Unique solvability	12
1.2.2 Pointwise error estimates	14
1.2.3 Superconvergence in time mesh points	18
1.2.4 Numerical results	19
1.3 Associated quadrature formulas and their advantages	23
1.3.1 Special quadrature formulas	23
1.3.2 Postprocessing	25
1.3.3 Connections to collocation methods	26
1.3.4 Shortcut to error estimates	27
1.3.5 Numerical results	29
1.4 Results for affine linear problems	30
1.4.1 A slight modification of the method	30
1.4.2 Postprocessing for the modified method	32
1.4.3 Interpolation cascade	37
1.4.4 Derivatives of solutions	39
1.4.5 Numerical results	41
2 Error Analysis for Stiff Systems	45
2.1 Runge–Kutta-like discretization framework	45
2.1.1 Connection between collocation and Runge–Kutta methods and its extension	46
2.1.2 A Runge–Kutta-like scheme	48
2.1.3 Existence and uniqueness	50

2.1.4	Stability properties	53
2.2	VTD methods as Runge–Kutta-like discretizations	54
2.2.1	Block structure of A^{VTD}	59
2.2.2	Eigenvalue structure of A^{VTD}	61
2.2.3	Solvability and stability	64
2.3	(Stiff) Error analysis	67
2.3.1	Recursion scheme for the global error	67
2.3.2	Error estimates	70
2.3.3	Numerical results	74

II Variational Time Discretization Methods for Parabolic Problems 79

3	Introduction to Parabolic Problems	81
3.1	Regularity of solutions	83
3.2	Semi-discretization in space	84
3.2.1	Reformulation as ode system	85
3.2.2	Differentiability with respect to time	89
3.2.3	Error estimates for the semi-discrete approximation	90
3.3	Full discretization in space and time	92
3.3.1	Formulation of the methods	92
3.3.2	Reformulation and solvability	93
4	Error Analysis for VTD Methods	95
4.1	Error estimates for the ℓ th derivative	98
4.1.1	Projection operators	99
4.1.2	Global L^2 -error in the H -norm	101
4.1.3	Global L^2 -error in the V -norm	109
4.1.4	Global (locally weighted) L^2 -error of the time derivative in the H -norm	117
4.1.5	Pointwise error in the H -norm	122
4.1.6	Supercloseness and its consequences	125
4.2	Error estimates in the time (mesh) points	131
4.2.1	Exploiting the collocation conditions	131
4.2.2	What about superconvergence!?	133
4.2.3	Satisfactory order convergence avoiding superconvergence	134
4.3	Final error estimate	139
4.4	Numerical results	144

Summary and Outlook 153

Appendix	157
A Miscellaneous Results	157
A.1 Discrete Gronwall inequality	157
A.2 Something about Jacobi-polynomials	158
B Abstract Projection Operators for Banach Space-Valued Functions	159
B.1 Abstract definition and commutation properties	159
B.2 Projection error estimates	163
B.3 Literature references on basics of Banach space-valued functions	166
C Operators for Interpolation and Projection in Time	167
C.1 Interpolation operators	167
C.2 Projection operators	168
C.3 Some commutation properties	173
C.4 Some stability results	174
D Norm Equivalences for Hilbert Space-Valued Polynomials	175
D.1 Norm equivalence used for the cGP-like case	176
D.2 Norm equivalence used for final error estimate	178
Bibliography	181

Acknowledgments

This work would not have been possible without a number of people whom I would like to thank at this point.

I am particularly grateful to my supervisor Prof. Dr. Gunar Matthies who first sparked my interest in variational time discretizations. He was always available to advise me on questions and ambiguities, took a lot of time for joint mathematical discussions and research, as well as willingly read a wide variety of preliminary versions of my research results. His patience and encouragement made it possible for me to find a way to the hoped-for results after many unsatisfactory approaches. And last but not least, he provided me with most of the implementation of the methods under consideration.

I would also like to thank Prof. Dr. Hans-Görg Roos for motivating me to stay at the university for a dissertation after my master's studies and for following my path in research with great interest until today. His lectures and the associated exercises of Dr. Martin Schopf first got me passionate about “my” mathematical field – numerical analysis.

I really appreciate the pleasant and friendly working atmosphere and the positive community spirit at the Institute of Numerical Mathematics, TU Dresden. I am grateful to all my current and former colleagues for this. Special mention goes to Prof. Dr. Sebastian Franz whose door was always open for me with any questions and queries. He took away most of my self-doubts about publishing in English and agreed to proofread many of my papers and also large parts of this thesis. I would also like to thank Dr. Reiner Vanselow for motivating me during the lunches we had together and for sharing his experiences from many years at the university. Moreover, I wish to add my thanks to Dr. Hanne Hardering for the great conversations and valuable exchanges, especially during the years when we shared an office.

Of course, there is also a world beyond work and university. I would like to express my heartfelt gratitude to my family and especially to my parents. They have always left me the freedom to find and go my own way. And I could always be sure of their support.

But the most important person by my side for many years has been my beloved girlfriend. Her enormous energy for others, her forbearance with my frequent melancholy, and her patience with me I deeply admire and appreciate. Thank you, Marie, for choosing me, for standing by me, and for being my home.

List of Symbols and Abbreviations

Abbreviations

VTD	variational time discretization [method]
dG	discontinuous Galerkin [method]
cGP	continuous Galerkin–Petrov [method]
ode	ordinary differential equation
RK	Runge–Kutta [method]
RK1	Runge–Kutta-like [method]

Sets and spaces

\mathbb{C}	set of complex numbers
\mathbb{R}	set of real numbers
\mathbb{Z}	set of integers
\mathbb{N}	set of positive integers $\{1, 2, 3, \dots\}$
\mathbb{N}_0	set of non-negative integers $\{0, 1, 2, \dots\}$
I	(temporal) interval $I = (t_0, t_0 + T]$ with $t_0, T \in \mathbb{R}$, $T > 0$
J	arbitrary interval in \mathbb{R}
Ω	bounded (spatial) domain in \mathbb{R}^{d_Ω} , $d_\Omega \in \mathbb{N}$
$\partial\Omega$	boundary of the (spatial) domain Ω
X	Banach space over \mathbb{R}
V, W, H	Hilbert spaces
V_h	finite dimensional subspace of V , typically a finite element space
X', V', V'_h	(topological) dual spaces of X , V , and V_h , respectively
$L^p(J), L^p(\Omega)$	$1 \leq p < \infty$: Lebesgue space of p -power integrable functions over J and Ω , respectively $p = \infty$: Lebesgue space of measurable, essentially bounded functions over J and Ω , respectively
$L^p(J, X)$	Bochner(–Lebesgue) space that generalizes the concept of the L^p -space to X -valued functions over J

$W^{m,p}(J),$ $W^{m,p}(\Omega)$	standard Sobolev spaces with derivatives up to order m in $L^p(J)$ and $L^p(\Omega)$, respectively
$H^m(\Omega)$	Sobolev space $W^{m,2}(\Omega)$
$H_0^1(\Omega)$	subspace of $H^1(\Omega)$ of functions having zero boundary traces
$H^{-1}(\Omega)$	dual space of $H_0^1(\Omega)$, i.e., $H^{-1}(\Omega) = H_0^1(\Omega)'$
$W^{m,p}(J, X)$	Bochner–Sobolev space that generalizes the concept of the $W^{m,p}$ -space to X -valued functions over J
$H^m(J, X)$	Bochner–Sobolev space $W^{m,2}(J, X)$
$C^m(J, X)$	space of X -valued functions over J with continuous m th order derivatives; sometimes also $C^{-1}(J, X) = L^2(J, X)$
$P_m(J, X)$	space of X -valued polynomials of degree m over J ; $P_{-1}(J, X) = \{0\}$

Norms, inner products, bilinear forms

$\ \cdot\ _X$	norm of Banach space X
$\langle \cdot, \cdot \rangle_{X', X}$	duality pairing over $X' \times X$
$\ \cdot\ $	Euclidean norm, also norm of Hilbert space H
(\cdot, \cdot)	Euclidean inner product, also inner product of Hilbert space H
$\ \cdot\ _V$	norm of Hilbert space V
$(\cdot, \cdot)_V$	inner product of Hilbert space V
$a(\cdot, \cdot)$	bounded, V -elliptic bilinear form over $V \times V$
$B_n^{\mathcal{J}}(\cdot, \cdot)$	bilinear form used in the error analysis for the ℓ th derivative

Mesh quantities

N	number of time mesh (sub-)intervals in the decomposition of I
t_n	time (mesh) point, $n = 0, \dots, N$
I_n	local time mesh (sub-)interval $I_n = (t_{n-1}, t_n]$
T_n	affine (reference) transformation that maps $[-1, 1]$ to \bar{I}_n
τ_n	length of time mesh (sub-)interval I_n
τ	maximum among the lengths of time mesh (sub-)intervals
\mathcal{T}_h	triangulation of the (spatial) domain Ω
h	maximum among the diameters of mesh cells contained in \mathcal{T}_h

More notation

$\lfloor \cdot \rfloor$	floor function
$[\cdot]_n$	jump at t_n
$\mathcal{O}(\cdot)$	Landau symbol
$\delta_{i,j}$	Kronecker symbol: $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise
$S_1 \cup_{\text{cond.}} S_2$	$S_1 \cup S_2$ if “cond.” is fulfilled and S_1 otherwise
$P_m^{(\alpha,\beta)}$	m th Jacobi-polynomial on $(-1, 1)$ w.r.t. the weight $(1-t)^\alpha(1+t)^\beta$, $\alpha, \beta > -1$
$\hat{\cdot}$	quantity in a reference interval

Selected approximation operators

\mathcal{I}_n	(general) integrator on I_n
$\mathcal{I}_{[m,n]}$	compound (general) integrator on $(t_{m-1}, t_n]$, $\mathcal{I}_{[m,n]}[\cdot] = \sum_{\nu=m}^n \mathcal{I}_\nu[\cdot]$
Q_n	(general) quadrature formula on I_n
Q_k^r	quadrature formula (on I_n) assigned to the \mathbf{VTD}_k^r method
\mathcal{I}_n	(general) operator for approximating the right-hand side on I_n
\mathcal{I}_k^r	interpolation operator (on I_n) assigned to the Q_k^r - \mathbf{VTD}_k^r method, uses the quadrature points of Q_k^r for interpolation
\mathcal{C}_k^r	cascadic interpolation operator (on I_n), $\mathcal{C}_k^r = \mathcal{I}_k^r \circ \dots \circ \mathcal{I}_{2^{r-k}}^{2^{r-k}}$
$\mathcal{I}_{k-2,*}^r$	interpolation operator (on I_n) assigned to the postprocessed Q_{k-2}^{r-1} - \mathbf{VTD}_{k-2}^{r-1} method
Π_k^r	projection operator (on I_n) assigned to the exactly integrated \mathbf{VTD}_k^r method
Π_m	local L^2 -projection operator (on I_n)
$\tilde{\Pi}_l^{m,\mathcal{I}}$	generalized dG/cGP projection operator (on I_n), $l \in \{0, 1\}$
R_h	Ritz projection onto V_h
P_h	some generalization of the (global) L^2 -projection onto V_h
\tilde{P}_h^0	operator for (spatial) approximation of the initial value, $\tilde{P}_h^0 \in \{R_h, P_h\}$

Selected functions

f, F	functions of the “right-hand side”
g, G	approximations of f and F , respectively
u	exact solution of the initial value problem or the parabolic problem

U	discrete solution in the setting of initial value problems
u_h, U_h	solution of the semi-discretization in space and its basis representation, respectively
$u_{\tau h}, U_{\tau h}$	solution of the full discretization in space and time and its basis representation, respectively
$e_{\tau h, \ell}^{\mathcal{J}}$	fully discrete error of the ℓ th derivative, $e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}$

Further parameters and constants

r, k	parameters of the \mathbf{VTD}_k^r method, $0 \leq k \leq r$
ℓ	$\ell = \lfloor \frac{k}{2} \rfloor$ (in the context of \mathbf{VTD}_k^r methods)
$k_{\mathcal{J}}, k_{\mathcal{I}}$	derivative order needed for well-definedness of \mathcal{J}_n and \mathcal{I}_n , respectively
μ	one-sided Lipschitz constant
κ	approximation order of the spatial discretization w.r.t. the H^1 -norm
$\sigma, \tilde{\sigma}, \check{\sigma}$	auxiliary variables to handle presence/absence of H^2 -regularity or different operator choices for (spatial) approximation of the initial value
C	generic positive constant, independent of the mesh parameter(s), also independent of the function(s) under consideration, unless specified otherwise
$C_{...}, \mathfrak{C}_{...}$	constants associated to concrete quantities or inequalities

Introduction

Nowadays, in research as well as in industrial product development, costly experiments are more and more replaced by numerical simulations. For this purpose, many of the often time-dependent processes in science and engineering are first modeled by differential equations. Since these differential equations can rarely be solved exactly with reasonable effort, their solutions have to be approximated using appropriate numerical methods. Thereby, it is common to use different methods for approximation with respect to space variables and with respect to time. The reason for this is that the requirements on the schemes are usually quite different. Difficulties in spatial approximation often originate from complicated domains, the occurrence of layers, or the need to apply stabilization methods. In temporal approximation, however, stability or conservation properties of the methods are more relevant.

To illustrate the latter, we consider an example. Using the method of lines to treat a parabolic partial differential equation, semi-discretization in space results in a huge system of ordinary differential equations. This system becomes stiffer with finer spatial discretization. Hence, implicit methods are preferable in order to avoid upper bounds for the time step length. Moreover, the used time discretization should be at least *A*-stable to ensure suitable stability properties. So, implicit Runge–Kutta methods, as the first-order implicit Euler method or the second-order implicit trapezoidal rule, may first come to mind. However, if one is additionally interested in higher order temporal approximations, discontinuous Galerkin or continuous Galerkin–Petrov schemes are particularly popular.

In this thesis, we consider a family of variational time discretizations that generalizes discontinuous Galerkin (dG) and continuous Galerkin–Petrov (cGP) methods. The origins of these discretizations lie in a preprint of Matthies and Schieweck [46] in which, after applying a postprocessing to dG and cGP schemes, new methods were found that in addition to variational conditions also contain collocation conditions in the time mesh points. Taking this idea, the considered family of methods was introduced by Matthies and the author. It was first published in a joint work with Wenzel [17] and then studied in more detail in [14, 16]. The methods of the family are characterized by two parameters that represent the local polynomial ansatz order and the number of non-variational conditions, which is also related to the global temporal regularity of the numerical solution. Moreover, with respect to Dahlquist’s stability problem the variational time discretization (**VTD**) methods either share their stability properties with the dG or the cGP method and, hence, are at least *A*-stable.

With every new method, however, the question naturally arises as to what advantages it has. Besides the potentially high convergence order and the stability properties, which both are also provided by dG and cGP methods, key feature of the new methods is that a high smoothness of the discrete solution with respect to time can be ensured. Moreover, we will see that under certain conditions superconvergence behavior in the time mesh points can be observed not only for the function values but also for the derivatives. Therefore, in appli-

cations where temporal smoothness is of interest or important target values are connected to derivatives, these new variational time discretizations could be quite advantageous.

But even from a purely theoretical point of view, it is worth looking at the whole family of **VTD** methods. The more general view on the variational time discretizations provides deeper insight into certain specifics of the well-known dG and cGP schemes. So, similarities and differences in the analysis of those methods become more apparent and, partly, even a unified analysis is possible. Furthermore, it reveals an approach to treat the break down of superconvergence for stiff problems that is observed for dG and cGP methods. This approach may also be used to avoid the order reduction phenomenon in the setting of initial-boundary value problems.

The overall goal of this thesis, addressed in Part II, is to investigate the family of **VTD** methods in combination with a finite element method for spatial approximation for problems in time and space. More specifically, for parabolic partial differential equations we want to prove optimal error estimates with respect to space and time under appropriate conditions. In preparation for this, extensive preliminary investigations are made. Especially, we first consider, in Part I, the methods in the context of initial value problems.

Therefore, this thesis may be seen as an overview of the state of knowledge about the considered family of variational time discretization methods. Here we mainly focus on theoretical studies and error analysis. In this sense, the numerical experiments included also are intended to highlight various properties of the methods using simple academic test examples, rather than presenting realistic application situations.

In Chapter 1 the **VTD** methods are formulated for systems of ordinary differential equations (odes). Moreover, under quite general, abstract assumptions an error analysis for non-stiff ode systems is presented. The obtained results especially clarify the influence of approximate integration and approximation of the right-hand side on the order of convergence. In addition, we discuss some key properties of the methods, which will often be of great importance later on. These include, in particular, the associated quadrature formulas, the postprocessing techniques, the connections to collocation methods with multiple nodes, the idea of cascadic interpolation, and the nestedness of conditions for the derivatives of the discrete solution. However, since most of the results have already been published in [14, 16], for brevity, we skip most of the proofs. New findings are given for methods with a modified right-hand side. In this context more general investigations of the postprocessing, the interpolation cascade, and the properties of derivatives of solutions are made. The different results are illustrated by numerical experiments.

Chapter 2 is devoted to the study of variational time discretizations for stiff systems of odes, where the considerations are restricted to affine linear problems with time-independent coefficients. To this end, we first introduce a new framework of Runge–Kutta-like schemes and study sufficient conditions for their unique solvability and some stability properties. Furthermore, we show that important representatives of the variational time discretizations can be written as Runge–Kutta-like methods and, in addition, provide solvability and stability under appropriate assumptions. This then allows us, by adapting and generalizing several techniques known from the (stiff) error analysis for Runge–Kutta methods, to derive error estimates for **VTD** methods also for the considered class of affine linear, stiff problems with time-independent coefficients. Computational results for a stiff example problem are

presented.

In Chapter 3, which begins Part II of the thesis, we give a brief introduction to parabolic problems. Since most of the findings are standard results, the presentation is kept rather short. After discussing the weak formulation and introducing a model problem for our numerical analysis, we have a look at existence, uniqueness, and regularity of solutions. Further, following the method of lines, we first consider the semi-discretization in space. A reformulation of the semi-discrete problem as ode system shows the similarity to the stiff problems studied in Chapter 2. Moreover, stability estimates and the differentiability with respect to time are investigated for the semi-discrete solution, and abstract error estimates for the spatial semi-discretization are presented. Finally, we obtain full discretizations in space and time by applying the **VTD** schemes to the spatial semi-discrete problem.

An error analysis for the fully discrete method is developed in Chapter 4. To this end, results from all three previous chapters are reused and combined. First, estimates in various integral-based norms as well as pointwise estimates are proven for a certain time derivative of the error. Here, we take advantage of the nestedness of conditions for the derivatives of the discrete solution such that known approaches from the analysis of dG and cGP methods can be applied. Nevertheless, our way of presenting the error analysis is quite unusual since dG and cGP schemes are studied in parallel. This nicely reveals the great similarities but also the differences in the analysis of the two methods. Moreover, supercloseness results are obtained. Second, we address error estimates in the time (mesh) points also for lower derivatives. For this, we draw on the results from the (stiff) error analysis and the findings on the semi-discretization in space. In conclusion, combining all these observations, we obtain error estimates for the full discretization that are of optimal order with respect to space and time. Further, illustrating numerical results are given.

We close the main part of the thesis with a brief summary of the results. Moreover, we provide an outlook on how the findings could be used further and raise some open questions on variational time discretizations that may be answered in future work.

This thesis also contains an appendix. In it, some mathematical basics are compiled, but also several results are proven that are very important for our analysis in Chapter 4. Therefore, we also want to briefly outline its contents.

In Appendix A miscellaneous results are collected. A less common variant of the discrete Gronwall lemma is proven, which we need in our analysis, and some information on Jacobi-polynomials are given. Abstract projection operators for Banach space-valued functions are studied in Appendix B. We give an abstract definition for polynomial projection operators and investigate some commutation properties. Furthermore, some main results of standard finite element interpolation theory, in particular projection error estimates, are generalized to the univariate, Banach space-valued case. In Appendix C, we then give a compilation of the concrete temporal interpolation and projection operators that are used especially in Part II. We investigate their well-definedness and take a look at some of their properties. Finally, in Appendix D, we show for two examples how norm equivalences for real-valued polynomials can be generalized to norm equivalences for Hilbert space-valued polynomials.

Part I

Variational Time Discretization Methods for Initial Value Problems

1 Formulation, Analysis for Non-Stiff Systems, and Further Properties

We consider the initial value problem

$$Mu'(t) = F(t, u(t)), \quad u(t_0) = u_0 \in \mathbb{R}^d, \quad (1.1)$$

where $M \in \mathbb{R}^{d \times d}$ is a regular matrix and F , sufficiently smooth, satisfies a Lipschitz condition with respect to the second variable. Furthermore, let $I = (t_0, t_0 + T]$ be an arbitrary but fixed time interval with positive length T . The value u_0 at $t = t_0$ will be called the initial value in the following.

If the ode system (1.1) originates from a finite element semi-discretization in space of a parabolic partial differential equation, then M is the time-constant mass matrix. Since in this context the computation of M^{-1} is costly, usually a linear system with M is solved instead. By the explicit occurrence of M we can investigate where this is necessary.

To describe the vector-valued case ($d > 1$) in an easy way, let (\cdot, \cdot) be the standard inner product and $\|\cdot\|$ the Euclidean norm on \mathbb{R}^d , $d \in \mathbb{N}$. Besides, let e_j be the j th standard unit vector in \mathbb{R}^d , $1 \leq j \leq d$.

For an arbitrary interval J and $q \in \mathbb{N}$, the spaces of continuous and m -times continuously differentiable \mathbb{R}^q -valued functions on J are written as $C(J, \mathbb{R}^q)$ and $C^m(J, \mathbb{R}^q)$, respectively. Furthermore, the space of square-integrable \mathbb{R}^q -valued functions shall be denoted by $L^2(J, \mathbb{R}^q)$ or, for convenience, sometimes also by $C^{-1}(J, \mathbb{R}^q)$. For non-negative integers s , we write $P_s(J, \mathbb{R}^q)$ for the space of \mathbb{R}^q -valued polynomials on J of degree less than or equal to s . Moreover, $P_{-1}(J, \mathbb{R}^q) := \{0\}$. For $q = 1$, we sometimes omit \mathbb{R}^q . Further notation is introduced later at the beginning of the sections where it is needed.

In order to describe the methods, we need a time mesh. Therefore, the interval I is decomposed by

$$t_0 < t_1 < \cdots < t_{N-1} < t_N = t_0 + T$$

into N disjoint subintervals $I_n := (t_{n-1}, t_n]$, $n = 1, \dots, N$. Furthermore, we set

$$\tau_n := t_n - t_{n-1}, \quad \tau := \max_{1 \leq n \leq N} \tau_n.$$

For convenience and to simplify the notation, we assume $\tau \leq 1$, which is not really a restriction since we are interested in the asymptotic error behavior for $\tau \rightarrow 0$. For any piecewise continuous function v , we define by

$$v(t_n^+) := \lim_{t \rightarrow t_n^+} v(t), \quad v(t_n^-) := \lim_{t \rightarrow t_n^-} v(t), \quad [v]_n := v(t_n^+) - v(t_n^-)$$

the one-sided limits and the jump of v at t_n . Moreover, with $\lfloor \cdot \rfloor$ the standard notation for the floor function is used.

Hereinafter C denotes a generic positive constant independent of the mesh parameter(s), especially τ , and the function(s) under consideration, unless specified otherwise.

1.1 Formulation of the methods

We now present some general formulation of the variational time discretization methods \mathbf{VTD}_k^r investigated in [14, 16]. Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Then, the local version of the numerical method (I_n -problem) reads as follows

Given $U(t_{n-1}^-) \in \mathbb{R}^d$, find $U \in P_r(I_n, \mathbb{R}^d)$ such that

$$U(t_{n-1}^+) = U(t_{n-1}^-), \quad \text{if } k \geq 1, \quad (1.2a)$$

$$MU^{(i+1)}(t_n^-) = \frac{d^i}{dt^i} \left(F(t, U(t)) \right) \Big|_{t=t_n^-}, \quad \text{if } k \geq 2, i = 0, \dots, \left\lfloor \frac{k}{2} \right\rfloor - 1, \quad (1.2b)$$

$$MU^{(i+1)}(t_{n-1}^+) = \frac{d^i}{dt^i} \left(F(t, U(t)) \right) \Big|_{t=t_{n-1}^+}, \quad \text{if } k \geq 3, i = 0, \dots, \left\lfloor \frac{k-1}{2} \right\rfloor - 1, \quad (1.2c)$$

and

$$\mathcal{J}_n \left[(MU', \varphi) \right] + \delta_{0,k} \left(M[U]_{n-1}, \varphi(t_{n-1}^+) \right) = \mathcal{J}_n \left[(\mathcal{I}_n F(\cdot, U(\cdot)), \varphi) \right] \quad \forall \varphi \in P_{r-k}(I_n, \mathbb{R}^d), \quad (1.2d)$$

where $U(t_0^-) = u_0$ and $\delta_{i,j}$ is the Kronecker symbol.

Here, \mathcal{J}_n denotes an integrator that typically represents either the integral over I_n or the application of a quadrature formula for approximate integration. Details will be described later on. Moreover, \mathcal{I}_n could be used to model some projection/interpolation of f or the usage of some special quadrature rules even if \mathcal{J}_n is just the integral.

We agree that both \mathcal{J}_n and \mathcal{I}_n are local versions (obtained by transformation) of appropriate linear operators $\hat{\mathcal{J}}$ and $\hat{\mathcal{I}}$ given on the reference interval $[-1, 1]$. Both operators work component-wise when applied to vector-valued functions.

Note that the formulation can be easily extended to the case $k = r + 1$. Then, the variational condition (1.2d) must formally hold for all $\varphi \in P_{-1}(I_n, \mathbb{R}^d)$. This can be interpreted as “there is no variational condition”. Hence, only conditions at both ends of the interval I_n are used.

The \mathbf{VTD}_k^r framework can shortly be described by

$$\begin{array}{l|l} \text{trial space: } P_r, & \text{if } k \geq 1 : \text{initial condition,} \\ \text{test space: } P_{r-k}, & \begin{array}{l} \text{if } k \geq 2 : \text{ODE}^{(i)} \text{ in } t_n^-, \quad i = 0, \dots, \left\lfloor \frac{k}{2} \right\rfloor - 1, \\ \text{if } k \geq 3 : \text{ODE}^{(i)} \text{ in } t_{n-1}^+, \quad i = 0, \dots, \left\lfloor \frac{k-1}{2} \right\rfloor - 1. \end{array} \end{array}$$

The notation $\text{ODE}^{(i)}$ means that the discrete solution fulfills the i th derivative of the system of ordinary differential equations. Obviously, the reduction of the test space for $k \geq 1$ is compensated by other conditions. For a somewhat related approach see [22, (3.3)].

Counting the number of conditions leads for $k \geq 1$ to

$$\dim P_{r-k} + 1 + \left\lfloor \frac{k}{2} \right\rfloor + \left\lfloor \frac{k-1}{2} \right\rfloor = r - k + 1 + 1 + \frac{k}{2} + \frac{k-1}{2} - \frac{1}{2} = r - k + 2 + k - 1 = r + 1$$

while we have also $\dim P_r = r + 1$ conditions if $k = 0$. The number of degrees of freedom equals for all k to $\dim P_r = r + 1$. Hence, in any case the number of conditions coincides with the number of degrees of freedom.

In order to indicate the dependence of the discretization on \mathcal{J}_n and \mathcal{I}_n , we shall refer to the concrete method defined by (1.2) as $\mathcal{J}_n\text{-}\mathbf{VTD}_k^r(\mathcal{I}_n)$. However, we agree that for $\mathcal{J}_n = \int_{I_n}$ and $\mathcal{I}_n = \text{Id}$, respectively, this specification is omitted.

Remark 1.1

The \mathbf{VTD}_k^r framework generalizes two well-known types of variational time discretization methods. The method \mathbf{VTD}_0^r is the discontinuous Galerkin method dG(r), whereas the method \mathbf{VTD}_1^r equates to the continuous Galerkin–Petrov method cGP(r).

On closer considerations we see that methods \mathbf{VTD}_k^r with even k are dG-like since there are point conditions on the $\lfloor \frac{k}{2} \rfloor$ th derivative of the discrete solution, but this derivative might be discontinuous. The methods \mathbf{VTD}_k^r with odd k are cGP-like since there are point conditions up to the $\lfloor \frac{k}{2} \rfloor$ th derivative of the discrete solution and this derivative is continuous if F is sufficiently smooth. We have in detail

$$\mathbf{VTD}_k^r \hat{=} \begin{cases} \text{dG}(r), & k = 0, \\ \text{cGP}(r), & k = 1, \\ \text{dG-C}^{\lfloor \frac{k-1}{2} \rfloor}(r), & k \geq 2, k \text{ even}, \\ \text{cGP-C}^{\lfloor \frac{k-1}{2} \rfloor}(r), & k \geq 3, k \text{ odd}, \end{cases}$$

where we use and generalize the definitions and notation of [46]. Note that there is also another reason to name the methods this way. All methods with odd k share their A-stability with the cGP method while methods with even k are strongly A-stable as the dG method. For details see [14, 17] or Remark 1.39 below. ♣

1.1.1 Global formulation

For $s \in \mathbb{Z}$, $s \geq 0$, we define the space Y_s of \mathbb{R}^d -valued piecewise polynomials of maximal degree s by

$$Y_s := \{ \varphi \in L^2(I, \mathbb{R}^d) : \varphi|_{I_n} \in P_s(I_n, \mathbb{R}^d), n = 1, \dots, N \}.$$

Studying the conditions (1.2a), (1.2b), and (1.2c), we easily see that the solution U of $\mathcal{J}_n\text{-}\mathbf{VTD}_k^r(\mathcal{I}_n)$ is $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable on I if F is globally $(\lfloor \frac{k-1}{2} \rfloor - 1)$ -times continuously differentiable. Furthermore, the condition (1.2b) for $U \in C^{\lfloor \frac{k-1}{2} \rfloor}(I, \mathbb{R}^d)$ then already implies (1.2c) for $n \geq 2$. Consequently, the method could be reformulated as follows

Find $U \in Y_r \cap C^{\lfloor \frac{k-1}{2} \rfloor}(I, \mathbb{R}^d)$ such that

$$U^{(i)}(t_0^+) = U^{(i)}(t_0^-), \quad \text{if } k \geq 1, i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor, \quad (1.3a)$$

$$MU^{(i+1)}(t_n^-) = \frac{d^i}{dt^i} \left(F(t, U(t)) \right) \Big|_{t=t_n^-}, \quad \text{if } k \geq 2, i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1, \quad (1.3b)$$

for all $n = 1, \dots, N$, and

$$\sum_{n=1}^N \left\{ \mathcal{J}_n[(MU' - \mathcal{I}_n F(\cdot, U(\cdot)), \varphi)] + \delta_{0,k}(M[U]_{n-1}, \varphi(t_{n-1}^+)) \right\} = 0 \quad \forall \varphi \in Y_{r-k}, \quad (1.3c)$$

where $U^{(i)}(t_0^-) = u^{(i)}(t_0)$, $0 \leq i \leq \lfloor \frac{k}{2} \rfloor$, which includes the initial value u_0 in the problem formulation. We agree on defining $u^{(j)}(t_0)$ recursively using the differential equation, i.e.,

$$\begin{aligned} u^{(0)}(t_0) &:= u_0, & Mu^{(2)}(t_0) &:= \partial_t F(t_0, u(t_0)) + \partial_u F(t_0, u(t_0))u^{(1)}(t_0), \\ Mu^{(1)}(t_0) &:= F(t_0, u(t_0)), & Mu^{(j)}(t_0) &:= \frac{d^{j-1}}{dt^{j-1}} F(t, u(t)) \Big|_{t=t_0}, \quad j \geq 3. \end{aligned} \quad (1.4)$$

The term $\frac{d^{j-1}}{dt^{j-1}} F(t, u(t)) \Big|_{t=t_0}$ depends only on $u(t_0), \dots, u^{(j-1)}(t_0)$ and can be calculated using some generalization of Faà di Bruno's formula, see e.g. [24, 47]. If F is affine linear in u , i.e., $F(t, u(t)) = f(t) - A(t)u(t)$, then we simply have

$$Mu^{(j)}(t_0) := \frac{d^{j-1}}{dt^{j-1}} F(t, u(t)) \Big|_{t=t_0} = f^{(j-1)}(t_0) - \sum_{l=0}^{j-1} \binom{j-1}{l} A^{(j-1-l)}(t_0) u^{(l)}(t_0), \quad j \geq 1,$$

by Leibniz' rule for the $(j-1)$ th derivative.

Note that, since the test space Y_{r-k} in (1.3c) allows discontinuities at the boundaries of subintervals, the problem can be decoupled by choosing test functions φ supported on a single time interval I_n only. Moreover, exploiting for $k \geq 1$ that $U \in C^{\lfloor \frac{k-1}{2} \rfloor}(I, \mathbb{R}^d)$ as well as (1.3a) and (1.3b), we also obtain (1.2a) and (1.2c). Therefore, the global problem (1.3) can be converted back into a sequence of local problems (1.2) in time on the subintervals I_n , $n = 1, \dots, N$.

1.1.2 Another formulation

In [6] a unified formulation for various time discretization schemes was investigated. Also the dG method ($k = 0$) and the cGP method ($k = 1$) with exact integration and $\mathcal{I}_n = \text{Id}$ were fitted and studied in this framework there. We shall show below, see (1.6), that for $1 \leq k \leq r$ also the $\mathcal{J}_n\text{-VTD}_k^r(\mathcal{I}_n)$ methods (1.2) could be analyzed in the framework of [6].

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. For sufficiently smooth v and under certain assumptions on \mathcal{J}_n , we uniquely define an approximation $\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} v \in P_{r-1}(I_n, \mathbb{R}^d)$ of v by the conditions

$$(\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} v)^{(i)}(t_n^-) = v^{(i)}(t_n^-), \quad \text{if } k \geq 2, \quad i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1, \quad (1.5a)$$

$$(\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} v)^{(i)}(t_{n-1}^+) = v^{(i)}(t_{n-1}^+), \quad \text{if } k \geq 3, \quad i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor - 1, \quad (1.5b)$$

$$\mathcal{J}_n \left[(\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} v(t), \varphi(t)) \right] = \mathcal{J}_n \left[(\mathcal{I}_n v(t), \varphi(t)) \right] \quad \forall \varphi \in P_{r-k}(I_n, \mathbb{R}^d) \text{ with } \delta_{0,k} \varphi(t_{n-1}) = 0, \quad (1.5c)$$

for details see [16, Lemma 17] or Remark 1.2 below.

Using this approximation operator, an equivalent formulation of (1.2) with $1 \leq k \leq r$ reads

Given $U(t_{n-1}^-) \in \mathbb{R}^d$, find $U \in P_r(I_n, \mathbb{R}^d)$ such that $U(t_{n-1}^+) = U(t_{n-1}^-)$ and

$$MU'(t) = \mathcal{P}_n^{\mathcal{J}, \mathcal{I}} F(t, U(t)) \quad \forall t \in I_n, \quad (1.6)$$

where $U(t_0^-) = u_0$.

Indeed, if U solves (1.2), then $MU' \in P_{r-1}(I_n, \mathbb{R}^d)$ obviously satisfies all conditions of (1.5) with $v = F(\cdot, U(\cdot))$. Hence, whenever $\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} v$ is uniquely defined we directly get (1.6).

Otherwise let U solve (1.6). Since there are polynomials on both sides, we can differentiate the equation by any order. With (1.5a) and (1.5b) we have

$$MU^{(i+1)}(\tilde{t}) = \frac{d^i}{dt^i} \left(\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} F(t, U(t)) \right) \Big|_{t=\tilde{t}} = \frac{d^i}{dt^i} \left(F(t, U(t)) \right) \Big|_{t=\tilde{t}}$$

for $\tilde{t} = t_n^-$ and $i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1$, if $k \geq 2$, as well as for $\tilde{t} = t_{n-1}^+$ and $i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor - 1$, if $k \geq 3$, respectively. Hence, the conditions (1.2b) and (1.2c) hold. Taking the inner product of (1.6) with an arbitrary $\varphi \in P_{r-k}(I_n, \mathbb{R}^d)$ and applying \mathcal{J}_n on both sides yield together with (1.5c)

$$\mathcal{J}_n \left[\left(MU', \varphi \right) \right] = \mathcal{J}_n \left[\left(\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} F(\cdot, U(\cdot)), \varphi \right) \right] = \mathcal{J}_n \left[\left(\mathcal{I}_n F(\cdot, U(\cdot)), \varphi \right) \right],$$

which is (1.2d). Hence, a solution of (1.6) also satisfies (1.2).

1.2 Existence, uniqueness, and error estimates

The existence and uniqueness of solutions to (1.2) as well as their error behavior are extensively studied in [16] for non-stiff problems. For the sake of brevity we shall only present the main results here. In order to formulate these results, some more notation and several assumptions need to be introduced.

First of all, recall that \mathcal{J}_n as well as \mathcal{I}_n are supposed to be local versions (obtained by transformation) of appropriate linear operators $\hat{\mathcal{J}}$ and $\hat{\mathcal{I}}$ given on the reference interval $[-1, 1]$. However, \mathcal{J}_n is an approximation of the integral operator while \mathcal{I}_n approximates the identity operator. Thus, the operations scale quite differently under transformation. More precisely, let

$$T_n : [-1, 1] \rightarrow \bar{I}_n, \quad \hat{t} \mapsto t := \frac{t_n + t_{n-1}}{2} + \frac{\tau_n}{2} \hat{t}, \quad (1.7)$$

denote the affine transformation that maps the reference interval $[-1, 1]$ on the closure of an arbitrary mesh interval $I_n = (t_{n-1}, t_n]$. Furthermore, let $k_{\mathcal{J}}$ and $k_{\mathcal{I}}$ be the smallest non-negative integers such that $\hat{\mathcal{J}}$ and $\hat{\mathcal{I}}$ are well-defined for functions in $C^{k_{\mathcal{J}}}([-1, 1])$ and $C^{k_{\mathcal{I}}}([-1, 1])$, respectively. Then, we have for all $\varphi \in C^{k_{\mathcal{J}}}(\bar{I}_n, \mathbb{R}^d)$ and for all $\psi \in C^{k_{\mathcal{I}}}(\bar{I}_n, \mathbb{R}^d)$ that

$$\mathcal{J}_n[\varphi] = \hat{\mathcal{J}}[\varphi \circ T_n] (T_n)' = \frac{\tau_n}{2} \hat{\mathcal{J}}[\varphi \circ T_n] \quad \text{and} \quad \mathcal{I}_n \psi = (\hat{\mathcal{I}}(\psi \circ T_n)) \circ T_n^{-1}.$$

Moreover, we suppose that for all non-negative integers l and $\hat{v} \in C^{\max\{k_{\mathcal{I}}, l\}}([-1, 1])$ it holds $\hat{\mathcal{I}}\hat{v} \in C^l([-1, 1])$, i.e., $\hat{\mathcal{I}}\hat{v}$ is at least as smooth as \hat{v} .

As before let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. The study of existence and uniqueness of solutions to (1.2) as well as the error analysis is strongly connected with the following operator.

Let $\mathcal{J}_n^{\mathcal{J}, \mathcal{I}} : C^{k_{\mathcal{J}}+1}(\bar{I}_n, \mathbb{R}^d) \rightarrow P_r(\bar{I}_n, \mathbb{R}^d)$, $1 \leq n \leq N$, with $k_{\mathcal{J}} := \max \{ \lfloor \frac{k}{2} \rfloor - 1, k_{\mathcal{J}}, k_{\mathcal{I}} \}$ be defined by

$$(\mathcal{J}_n^{\mathcal{J}, \mathcal{I}} v)^{(i)}(t_{n-1}^+) = v^{(i)}(t_{n-1}^+), \quad \text{if } k \geq 1, i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor, \quad (1.8a)$$

$$(\mathcal{J}_n^{\mathcal{J}, \mathcal{I}} v)^{(i)}(t_n^-) = v^{(i)}(t_n^-), \quad \text{if } k \geq 2, i = 1, \dots, \lfloor \frac{k}{2} \rfloor, \quad (1.8b)$$

$$\begin{aligned} & \mathcal{J}_n \left[((\mathcal{J}_n^{\mathcal{J}, \mathcal{I}} v)'(t), \varphi(t)) \right] + \delta_{0,k} \mathcal{J}_n^{\mathcal{J}, \mathcal{I}} v(t_{n-1}^+) \varphi(t_{n-1}^+) \\ &= \mathcal{J}_n \left[(\mathcal{I}_n(v')(t), \varphi(t)) \right] + \delta_{0,k} v(t_{n-1}^+) \varphi(t_{n-1}^+) \quad \forall \varphi \in P_{r-k}(I_n, \mathbb{R}^d). \end{aligned} \quad (1.8c)$$

In [16, cf. (4.1) and Lemma 1] it was shown that $\mathcal{J}_n^{\mathcal{J}, \mathcal{I}}$ is well-defined and that the conditions (1.8) uniquely determine an approximation $\mathcal{J}_n^{\mathcal{J}, \mathcal{I}} v \in P_r(\bar{I}_n, \mathbb{R}^d)$ of $v \in C^{k_{\mathcal{J}}+1}(\bar{I}_n, \mathbb{R}^d)$ if the following assumption is fulfilled.

Assumption 1.1

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, be the parameters of the method. We assume that the reference integrator $\hat{\mathcal{J}}$ is such that $\hat{\psi} \in P_{r-\max\{1,k\}}([-1, 1])$ and

$$\hat{\mathcal{J}} \left[(1 - \hat{t})^{\lfloor \frac{k}{2} \rfloor} (1 + \hat{t})^{\lfloor \frac{k-1}{2} \rfloor} \hat{\psi} \hat{\varphi} \right] = 0 \quad \forall \hat{\varphi} \in P_{r-\max\{1,k\}}([-1, 1])$$

imply $\hat{\psi} \equiv 0$. Note that the absolute value in the exponent is needed only for $k = 0$.

Remark 1.2

The approximation operator $\mathcal{J}_n^{\mathcal{J}, \mathcal{I}}$ is somewhat connected to $\mathcal{P}_n^{\mathcal{J}, \mathcal{I}}$ introduced in (1.5). In fact it holds $(\mathcal{J}_n^{\mathcal{J}, \mathcal{I}} v)' = \mathcal{P}_n^{\mathcal{J}, \mathcal{I}}(v')$ for all $v \in C^{k_{\mathcal{J}}+1}(\bar{I}_n, \mathbb{R}^d)$.

Accordingly, $\mathcal{P}_n^{\mathcal{J}, \mathcal{I}}$ is well-defined for functions in $C^{k_{\mathcal{J}}}(\bar{I}_n, \mathbb{R}^d)$ and uniquely determines an approximation $\mathcal{P}_n^{\mathcal{J}, \mathcal{I}} v \in P_{r-1}(\bar{I}_n, \mathbb{R}^d)$ of $v \in C^{k_{\mathcal{J}}}(\bar{I}_n, \mathbb{R}^d)$ if Assumption 1.1 holds, cf. [16, Lemma 17]. ♣

1.2.1 Unique solvability

First, we have a look on the unique solvability of the local problems (1.2) characterizing the \mathcal{J}_n -VTD $_k^r(I_n)$ method.

Assumption 1.2

We assume that the reference integrator $\hat{\mathcal{J}}$ is a bounded linear operator between $C^{k_{\mathcal{J}}}([-1, 1])$ and \mathbb{R} . So, it satisfies

$$\left| \hat{\mathcal{J}}[\hat{\varphi}] \right| \leq \mathfrak{C}_0 \sum_{j=0}^{k_{\mathcal{J}}} \sup_{\hat{t} \in [-1, 1]} |\hat{\varphi}^{(j)}(\hat{t})| \quad \forall \hat{\varphi} \in C^{k_{\mathcal{J}}}([-1, 1]),$$

where, as before, $k_{\mathcal{J}} \geq 0$ is the smallest non-negative integer such that $\hat{\mathcal{J}}$ is well-defined on $C^{k_{\mathcal{J}}}([-1, 1])$.

Assumption 1.3

We assume that for all $0 \leq l \leq k_{\mathcal{J}}$ the reference approximation operator $\hat{\mathcal{I}}$ is a bounded linear operator between $C^{\max\{k_{\mathcal{I}}, l\}}([-1, 1])$ and $C^l([-1, 1])$. So, for $0 \leq l \leq k_{\mathcal{J}}$ it satisfies

$$\sup_{\hat{t} \in [-1, 1]} |(\hat{\mathcal{I}}\hat{\varphi})^{(l)}(\hat{t})| \leq \mathfrak{C}_1 \sum_{j=0}^{\max\{k_{\mathcal{I}}, l\}} \sup_{\hat{t} \in [-1, 1]} |\hat{\varphi}^{(j)}(\hat{t})| \quad \forall \hat{\varphi} \in C^{\max\{k_{\mathcal{I}}, l\}}([-1, 1]),$$

where, as before, $k_{\mathcal{I}} \geq 0$ is the smallest non-negative integer such that $\hat{\mathcal{I}}$ is well-defined on $C^{k_{\mathcal{I}}}([-1, 1])$.

Assumption 1.4

We assume that for $0 \leq i \leq k_{\mathcal{J}} = \max\{\lfloor \frac{k}{2} \rfloor - 1, k_{\mathcal{J}}, k_{\mathcal{I}}\}$ the condition

$$\left\| \frac{d^i}{dt^i} \left(F(t, v(t)) - F(t, w(t)) \right) \right\|_{t=s} \leq \mathfrak{C}_2 \sum_{l=0}^i \|(v - w)^{(l)}(s)\| \quad \text{for a.e. } s \in \bar{I} = [t_0, t_0 + T]$$

holds for sufficiently smooth functions v, w . Here \mathfrak{C}_2 depends on $k_{\mathcal{J}}$ and F .

Remark 1.3

Sufficient conditions for Assumption 1.4 would be

- (i) for $k_{\mathcal{J}} = 0$: F satisfies a Lipschitz condition on the second variable with constant $L > 0$,
- (ii) for $k_{\mathcal{J}} \geq 1$: F is affine linear in u , i.e., $F(t, u(t)) = A(t)u(t) + f(t)$, and $\|A(\cdot)\|_{C^{k_{\mathcal{J}}}} < \infty$. Then, the inequality follows from Leibniz' rule for the i th derivative.
- (iii) In the literature, see [39, p. 74], there also appear conditions of the form

$$\sup_{t \in I, y \in \mathbb{R}^d} \left\| \frac{\partial}{\partial y} F^{(i)}(t, y) \right\| < \infty, \quad 0 \leq i \leq k_{\mathcal{J}},$$

where $F^{(i)}$ denotes the i th total derivatives of F with respect to t in the sense of [39, p. 65]. These conditions may be weaker in some cases.

Since in general the constant \mathfrak{C}_2 is somewhat connected to the Lipschitz constant and, thus, to the stiffness of the ode system, the dependence of the results on this constant shall be particularly highlighted. ♣

Now, we are ready to state a result on the solvability of the local problem (1.2).

Theorem 1.4 (Existence and uniqueness, cf. [16, Theorem 5])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. We suppose that Assumptions 1.1, 1.2, 1.3, and 1.4 hold. Then, there is a constant $\gamma_{r,k} > 0$, multiplicatively depending on \mathfrak{C}_2^{-1} but independent of n , such that problem (1.2) has a unique solution for all $1 \leq n \leq N$ when $\tau_n \leq \gamma_{r,k}$.

1.2.2 Pointwise error estimates

In order to derive error estimates along the lines of [16, Section 4], the assumptions need to be strengthened. In detail, compared to Theorem 1.4 we replace Assumption 1.3 by Assumption 1.5a or 1.5b (following below). This is necessary since in the proof derivatives can be handled if they are given in certain points but not their supremum. Furthermore, the error analysis exploits an auxiliary interpolation operator \mathcal{I}^{app} defined below, see Definition 1.6, which amongst others is based on these assumptions.

Assumption 1.5a

For $0 \leq l \leq k_{\mathcal{I}}$ we assume that $\widehat{\mathcal{I}}\widehat{\varphi} \in C^l([-1, 1])$ and that there are disjoint points $\hat{t}_m^{\mathcal{I}}, m = 0, \dots, K^{\mathcal{I}}$, in the reference interval $[-1, 1]$ such that

$$\sup_{\hat{t} \in [-1, 1]} |(\widehat{\mathcal{I}}\widehat{\varphi})^{(l)}(\hat{t})| \leq \mathfrak{C}_{1,1} \sum_{m=0}^{K^{\mathcal{I}}} \sum_{j=0}^{\tilde{K}_m^{\mathcal{I}}} |\widehat{\varphi}^{(j)}(\hat{t}_m^{\mathcal{I}})| + \mathfrak{C}_{1,2} \sup_{\hat{t} \in [-1, 1]} |\widehat{\varphi}(\hat{t})| \quad \forall \widehat{\varphi} \in C^{k_{\mathcal{I}}}([-1, 1]).$$

Note that then typically $k_{\mathcal{I}} = \max\{\tilde{K}_m^{\mathcal{I}} : m = 0, \dots, K^{\mathcal{I}}\}$.

Assumption 1.5b

We assume that there are disjoint points $\hat{t}_m^{\mathcal{J}}, m = 0, \dots, K^{\mathcal{J}}$, in the reference interval $[-1, 1]$ such that

$$|\widehat{\mathcal{J}}[\widehat{\varphi}]| \leq \tilde{\mathfrak{C}}_{0,1} \sum_{m=0}^{K^{\mathcal{J}}} \sum_{j=0}^{\tilde{K}_m^{\mathcal{J}}} |\widehat{\varphi}^{(j)}(\hat{t}_m^{\mathcal{J}})| + \tilde{\mathfrak{C}}_{0,2} \sup_{\hat{t} \in [-1, 1]} |\widehat{\varphi}(\hat{t})| \quad \forall \widehat{\varphi} \in C^{k_{\mathcal{J}}}([-1, 1]).$$

Note that then typically $k_{\mathcal{J}} = \max\{\tilde{K}_m^{\mathcal{J}} : m = 0, \dots, K^{\mathcal{J}}\}$.

Moreover, we assume that there are disjoint points $\hat{t}_m^{\mathcal{I}}, m = 0, \dots, K^{\mathcal{I}}$, in the reference interval $[-1, 1]$ such that

$$\begin{aligned} & \sum_{m=0}^{K^{\mathcal{J}}} \sum_{l=0}^{\tilde{K}_m^{\mathcal{J}}} |(\widehat{\mathcal{I}}\widehat{\varphi})^{(l)}(\hat{t}_m^{\mathcal{J}})| + \sup_{\hat{t} \in [-1, 1]} |\widehat{\mathcal{I}}\widehat{\varphi}(\hat{t})| \\ & \leq \tilde{\mathfrak{C}}_{1,1} \sum_{m=0}^{K^{\mathcal{I}}} \sum_{j=0}^{\tilde{K}_m^{\mathcal{I}}} |\widehat{\varphi}^{(j)}(\hat{t}_m^{\mathcal{I}})| + \tilde{\mathfrak{C}}_{1,2} \sup_{\hat{t} \in [-1, 1]} |\widehat{\varphi}(\hat{t})| \quad \forall \widehat{\varphi} \in C^{\max\{k_{\mathcal{J}}, k_{\mathcal{I}}\}}([-1, 1]). \end{aligned}$$

Remark 1.5

Assumption 1.5a is satisfied if $\widehat{\mathcal{I}}$ is a polynomial approximation operator whose defining degrees of freedom only use derivatives in certain points, as, for example, Hermite interpolation operators. Together with Assumption 1.2, then $|\widehat{\mathcal{J}}[\widehat{\mathcal{I}}\widehat{\varphi}]|$ could be estimated by the supremum of $|\widehat{\varphi}|$ in $[-1, 1]$ and certain point values of derivatives of $\widehat{\varphi}$.

However, Assumption 1.5a is not satisfied if $\widehat{\mathcal{I}} = \text{Id}$ and $k_{\mathcal{J}} > 0$. In order to enable a similar estimate for $|\widehat{\mathcal{J}}[\widehat{\mathcal{I}}\widehat{\varphi}]|$ also in this case, Assumption 1.5b is formulated. Here, the requirements on the integrator $\widehat{\mathcal{J}}$ are increased. Of course, the defining degrees of freedom for the integrator now should use derivatives in certain points only. In return, the requirements for $\widehat{\mathcal{I}}$ can be weakened such that they are met for example also by $\widehat{\mathcal{I}} = \text{Id}$. ♣

Definition 1.6 (Auxiliary interpolation operator)

For the error estimation we introduce a special Hermite interpolation operator $\mathcal{I}_n^{\text{app}}$. Concretely, the operator should satisfy the following conditions: $\mathcal{I}_n^{\text{app}}$ preserves derivatives up to order $\lfloor \frac{k}{2} \rfloor - 1$ in t_n^- and up to order $\lfloor \frac{k-1}{2} \rfloor - 1$ in t_{n-1}^+ , i.e.,

$$\begin{aligned} (\mathcal{I}_n^{\text{app}}\varphi)^{(l)}(t_n^-) &= \varphi^{(l)}(t_n^-) & \text{for } 0 \leq l \leq \lfloor \frac{k}{2} \rfloor - 1, \\ (\mathcal{I}_n^{\text{app}}\varphi)^{(l)}(t_{n-1}^+) &= \varphi^{(l)}(t_{n-1}^+) & \text{for } 0 \leq l \leq \lfloor \frac{k-1}{2} \rfloor - 1. \end{aligned} \quad (1.9)$$

Moreover, we suppose that

$$(\mathcal{I}_n^{\text{app}}\varphi)^{(l)}(t_{n,m}^{\mathcal{I}}) = \varphi^{(l)}(t_{n,m}^{\mathcal{I}}) \quad \text{for } 0 \leq m \leq K^{\mathcal{I}}, 0 \leq l \leq \tilde{K}_m^{\mathcal{I}}, \quad (1.10a)$$

with $t_{n,m}^{\mathcal{I}} := \frac{t_n + t_{n-1}}{2} + \frac{\tau_n}{2} \hat{t}_m^{\mathcal{I}}$, where the points $\hat{t}_m^{\mathcal{I}}$ are those of Assumption 1.5a or 1.5b, respectively. If (1.9) and (1.10a) provide r^{app} independent interpolation conditions and $r^{\text{app}} < r + 1$, then we choose $r + 1 - r^{\text{app}}$ further points $\tilde{t}_m^{\mathcal{I}} \in (-1, 1) \setminus \{\tilde{t}_j^{\mathcal{I}} : j = 0, \dots, K^{\mathcal{I}}\}$, $m = K^{\mathcal{I}} + 1, \dots, K^{\mathcal{I}} + r + 1 - r^{\text{app}}$, and suppose

$$(\mathcal{I}_n^{\text{app}}\varphi)(t_{n,m}^{\mathcal{I}}) = \varphi(t_{n,m}^{\mathcal{I}}) \quad \text{for } K^{\mathcal{I}} + 1 \leq m \leq K^{\mathcal{I}} + r + 1 - r^{\text{app}}, \quad (1.10b)$$

where again $t_{n,m}^{\mathcal{I}} := \frac{t_n + t_{n-1}}{2} + \frac{\tau_n}{2} \hat{t}_m^{\mathcal{I}}$. We agree that $\mathcal{I}_n^{\text{app}}$ is applied component-wise to vector-valued functions. Overall, conditions (1.9) and (1.10) uniquely define a Hermite-type interpolation operator of ansatz order $\max\{r^{\text{app}} - 1, r\}$. \clubsuit

Now, we are able to provide an abstract error estimate.

Theorem 1.7 (Cf. [16, Theorem 8])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. We suppose that Assumptions 1.1, 1.2, and 1.4 hold. Moreover, let Assumption 1.5a or 1.5b be satisfied. Denote by u and U the solutions of (1.1) and (1.2), respectively. Then, we have for $1 \leq n \leq N$, sufficiently small τ , and $l = 0, 1$ that

$$\begin{aligned} \sup_{t \in I_n} \|(u - U)^{(l)}(t)\| &\leq C \max_{1 \leq \nu \leq n} \left(\sup_{t \in I_\nu} \|(\text{Id} - \mathcal{I}_\nu^{\text{app}})u(t)\| + \sum_{j=0}^l \sup_{t \in I_\nu} \|(u - \mathcal{J}_\nu^{\mathcal{I}, \mathcal{I}}u)^{(j)}(t)\| \right) \\ &\quad + C \max_{1 \leq \nu \leq n-1} \tau_\nu^{-1} \|(u - \mathcal{J}_\nu^{\mathcal{I}, \mathcal{I}}u)(t_\nu^-)\|, \end{aligned}$$

where the constants C in general exponentially depend on the product of T and \mathfrak{C}_2 .

Remark 1.8 (Cf. [16, Remark 9])

Based on Theorem 1.7 we can also prove abstract estimates for higher order derivatives of the error. Of course, we obtain that

$$\begin{aligned} \sup_{t \in I_n} \|(u - U)^{(l)}(t)\| &\leq \sup_{t \in I_n} \|(u - \mathcal{J}_n^{\mathcal{I}, \mathcal{I}}u)^{(l)}(t)\| + \sup_{t \in I_n} \|(\mathcal{J}_n^{\mathcal{I}, \mathcal{I}}u - U)^{(l)}(t)\| \\ &\leq \sup_{t \in I_n} \|(u - \mathcal{J}_n^{\mathcal{I}, \mathcal{I}}u)^{(l)}(t)\| + C_{\text{inv}} \left(\frac{\tau_n}{2}\right)^{-l} \sup_{t \in I_n} \|(\mathcal{J}_n^{\mathcal{I}, \mathcal{I}}u - U)(t)\| \\ &\leq \sup_{t \in I_n} \|(u - \mathcal{J}_n^{\mathcal{I}, \mathcal{I}}u)^{(l)}(t)\| + C_{\text{inv}} \left(\frac{\tau_n}{2}\right)^{-l} \left(\sup_{t \in I_n} \|(\mathcal{J}_n^{\mathcal{I}, \mathcal{I}}u - u)(t)\| + \sup_{t \in I_n} \|(u - U)(t)\| \right), \end{aligned}$$

where an inverse inequality was used. However, since we only have a non-local error estimate for $\sup_{t \in I_n} \|(u - U)(t)\|$, we cannot expect that the inverse of the local time step length can be compensated in general. So, usually we additionally need to assume that $\tau_\nu \leq \tau_{\nu+1}$ for all ν or alternatively that the mesh is quasi-uniform ($\tau/\tau_\nu \leq C$ for all ν) to get a proper estimate. ♣

Remark 1.9

Note that the estimate of Theorem 1.7 is appropriate for non-stiff problems only. Indeed, since the error constant C exponentially depends on the Lipschitz constant of the problem (hidden in \mathfrak{C}_2), this constant would be excessively large in the case of stiffness such that then the error bound would be useless.

Moreover, for the proof of Theorem 1.7 it is needed that τ_n is smaller than a certain bound which is inversely dependent on the Lipschitz constant. Therefore, stiff problems would force very small time step lengths. For semi-discretizations in space of parabolic time-space problems on shape-regular, quasi-uniform meshes, where the Lipschitz constant is typically proportional to h^{-2} with h denoting the spatial mesh parameter, this would cause upper bounds on the time step length with respect to h similar to CFL conditions. ♣

Of course, Theorem 1.7 provides an abstract bound for the error of the variational time discretization method. However, the order of convergence still is not clear. Since $\mathcal{I}_n^{\text{app}}$ is a Hermite-type interpolator of polynomial ansatz order larger than or equal to r , its approximation order (at least $r + 1$) is known. Suitable bounds on the error of the approximation operator $\mathcal{J}_n^{\mathcal{I}}$ shall be stated below. For their proof we refer to [16, Section 4].

Definition 1.10 (Approximation orders of \mathcal{J}_n and \mathcal{I}_n)

Let $r_{\text{ex}}^{\mathcal{J}}$, $r_{\text{ex}}^{\mathcal{I}}$, $r_{\mathcal{I}}$, and $r_{\mathcal{I},i}^{\mathcal{J}} \in \mathbb{N}_0 \cup \{-1, \infty\}$ denote the largest numbers such that

$$\begin{aligned} \int_{I_n} \varphi(t) dt &= \mathcal{J}_n[\varphi] \quad \forall \varphi \in P_{r_{\text{ex}}^{\mathcal{J}}}(I_n), & \int_{I_n} \varphi(t) dt &= \int_{I_n} \mathcal{I}_n \varphi(t) dt \quad \forall \varphi \in P_{r_{\text{ex}}^{\mathcal{I}}}(I_n), \\ \varphi &= \mathcal{I}_n \varphi \quad \forall \varphi \in P_{r_{\mathcal{I}}}(I_n), & \mathcal{J}_n[\varphi \psi_i] &= \mathcal{J}_n[(\mathcal{I}_n \varphi) \psi_i] \quad \forall \varphi \in P_{r_{\mathcal{I},i}^{\mathcal{J}}}(I_n), \psi_i \in P_i(I_n). \end{aligned}$$

Here, $P_{-1}(I_n)$ is interpreted as $\{0\}$, in which case the respective operator does not provide the corresponding approximation property. For convenience, set $r_{\mathcal{I}}^{\mathcal{J}} := r_{\mathcal{I},r-k}^{\mathcal{J}}$. Moreover, simply write $r_{\mathcal{I},i}^{\mathcal{J}}$ instead of $r_{\mathcal{I},i}^{\mathcal{J}}$ if \mathcal{J}_n represents the (exact) integral over I_n . Note that $r_{\text{ex}}^{\mathcal{I}} \geq r_{\mathcal{I},i}^{\mathcal{I}} \geq r_{\mathcal{I}}$ and $r_{\mathcal{I},i}^{\mathcal{J}} \geq r_{\mathcal{I}}$ hold by definition. ♣

Lemma 1.11 (Cf. [16, Lemma 12])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and suppose that Assumptions 1.1, 1.2, and 1.3 hold. Furthermore, let $l, \tilde{r} \in \mathbb{N}_0$ and define

$$j_{\min, \tilde{r}} := \min\{\tilde{r}, r + 1, r_{\mathcal{I}}^{\mathcal{J}} + 2\}, \quad j_{\max, \tilde{r}} := \max\{k_{\mathcal{J}} + 1, l, j_{\min, \tilde{r}}\}.$$

Then, provided that $v \in C^{j_{\max, \tilde{r}}}(\bar{I}_n, \mathbb{R}^d)$, the error estimate

$$\sup_{t \in I_n} \|(v - \mathcal{J}_n^{\mathcal{I}} v)^{(l)}(t)\| \leq C \sum_{j=j_{\min, \tilde{r}}}^{j_{\max, \tilde{r}}} \left(\frac{\tau_n}{2}\right)^{j-l} \sup_{t \in I_n} \|v^{(j)}(t)\|$$

holds with a constant C independent of τ_n .

Compared to the pointwise estimate of Lemma 1.11, the estimate for the approximation error of $\mathcal{J}_n^{\mathcal{J}, \mathcal{I}}$ in the mesh points t_n^- can even be improved in some cases. In fact, the following statement holds.

Lemma 1.12 (Cf. [16, Lemma 14])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Suppose that the Assumptions 1.1, 1.2, and 1.3 hold. Moreover, assume that $\max\{r_{\text{ex}}^{\mathcal{J}}, r_{\mathcal{I}}^{\mathcal{J}} + 1\} \geq r - 1$. Let $\tilde{r} \in \mathbb{N}_0$ and define

$$\begin{aligned} j_{\min, \tilde{r}}^{\diamond} &:= \min\{\tilde{r}, \max\{r_{\text{ex}}^{\mathcal{J}} + 1, \min\{r, r_{\mathcal{I}}^{\mathcal{J}} + 1\}\} + 1, r_{\mathcal{I}, 0}^{\mathcal{J}} + 2\}, \\ j_{\max, \tilde{r}}^{\diamond} &:= \max\{k_{\mathcal{J}} + 1, j_{\min, \tilde{r}}^{\diamond}\}. \end{aligned}$$

Then, provided that $v \in C^{j_{\max, \tilde{r}}^{\diamond}}(\bar{I}_n, \mathbb{R}^d)$, the error estimate

$$\|(v - \mathcal{J}_n^{\mathcal{J}, \mathcal{I}}v)(t_n^-)\| \leq C \sum_{j=j_{\min, \tilde{r}}^{\diamond}}^{j_{\max, \tilde{r}}^{\diamond}} \left(\frac{\tau_n}{2}\right)^j \sup_{t \in I_n} \|v^{(j)}(t)\|$$

holds for $1 \leq n \leq N$, where the constant C is independent of τ_n .

Finally, summarizing the above results, the guaranteed orders of convergence can now be listed clearly.

Corollary 1.13 (Cf. [16, Corollary 15])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $l \in \{0, 1\}$. Suppose that Assumptions 1.1, 1.2, 1.3, and 1.4 hold. Moreover, let Assumption 1.5a or 1.5b be satisfied. Denote by u and U the solutions of (1.1) and (1.2), respectively. Then, we have for $1 \leq n \leq N$

$$\sup_{t \in I_n} \|(u - U)^{(l)}(t)\| \leq C(F, u) \tau^{\min\{r, r_{\mathcal{I}}^{\mathcal{J}} + 1\}}, \quad (1.11)$$

with $r_{\mathcal{I}}^{\mathcal{J}}$ as defined in Definition 1.10. If in addition $\max\{r_{\text{ex}}^{\mathcal{J}}, r_{\mathcal{I}}^{\mathcal{J}} + 1\} \geq r - 1$, then we even have

$$\sup_{t \in I_n} \|(u - U)(t)\| \leq C(F, u) \tau^{\min\{r+1, r_{\mathcal{I}}^{\mathcal{J}}+2, r_{\mathcal{I}, 0}^{\mathcal{J}}+1, \max\{r_{\text{ex}}^{\mathcal{J}}+1, \min\{r, r_{\mathcal{I}}^{\mathcal{J}}+1\}\}\}}$$

as an improved error estimate.

If $\max\{r_{\text{ex}}^{\mathcal{J}}, r_{\mathcal{I}}^{\mathcal{J}} + 1\} \geq r - 1$ is satisfied, we obtain formally

$$\sup_{t \in I_n} \|(u - U)'(t)\| \leq C(F, u) \tau^{\min\{r, r_{\mathcal{I}}^{\mathcal{J}}+1, r_{\mathcal{I}, 0}^{\mathcal{J}}+1, \max\{r_{\text{ex}}^{\mathcal{J}}+1, \min\{r, r_{\mathcal{I}}^{\mathcal{J}}+1\}\}\}}$$

for the error of the first derivative. However, this gives the same convergence order as (1.11) for $l = 1$.

Remark 1.14

Since the quantity $r_{\mathcal{I}}^{\mathcal{J}} = r_{\mathcal{I}, r-k}^{\mathcal{J}}$ used in the lemmas and the corollary above is quite abstract, we want to provide lower bounds for $r_{\mathcal{I}, i}^{\mathcal{J}}$ based on the more familiar quantities $r_{\mathcal{I}}$, $r_{\text{ex}}^{\mathcal{I}}$,

and $r_{\text{ex}}^{\mathcal{J}}$. For the sake of simplicity, we shall impose somewhat stronger requirements on \mathcal{I}_n than actually necessary. For a proof in a slightly more general setting we refer to [13, Lemma 4.13].

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $i \in \mathbb{N}_0$. Then, $r_{\mathcal{I},i}^{\mathcal{J}} \geq r_{\mathcal{I}}$. So, for $r_{\mathcal{I}} = \infty$ the bound cannot be improved further. Otherwise, supposing that \mathcal{I}_n is a projection onto the space of polynomials of maximal degree $r_{\mathcal{I}} < \infty$, i.e., $\mathcal{I}_n : C^{k_{\mathcal{I}}}(\bar{I}_n) \rightarrow P_{r_{\mathcal{I}}}(\bar{I}_n)$ and $\mathcal{I}_n \varphi = \varphi$ for all $\varphi \in P_{r_{\mathcal{I}}}(\bar{I}_n)$, we even get

$$r_{\mathcal{I},i}^{\mathcal{J}} \geq \max\{r_{\mathcal{I}}, \min\{r_{\text{ex}}^{\mathcal{J}} - i, r_{\mathcal{I},i}^{\mathcal{J}}\}\}.$$

Of course, it holds $r_{\mathcal{I},0}^{\mathcal{J}} = r_{\text{ex}}^{\mathcal{J}}$. In order to simplify the term on the right-hand side for $i \geq 1$, we additionally could assume that \mathcal{I}_n is a Hermite-type interpolation operator. Then, we simply have

$$r_{\mathcal{I},i}^{\mathcal{J}} \geq \max\{r_{\mathcal{I}}, \min\{r_{\text{ex}}^{\mathcal{J}}, r_{\text{ex}}^{\mathcal{I}}\} - i\}$$

since then $r_{\mathcal{I},i}^{\mathcal{J}} \geq \max\{r_{\mathcal{I}}, r_{\text{ex}}^{\mathcal{I}} - i\}$.

Furthermore, under the weaker assumption that $\mathcal{I} = \mathcal{I}^1 \circ \dots \circ \mathcal{I}^l$ is a composition of several Hermite-type interpolation operators \mathcal{I}^j , $1 \leq j \leq l$, we still find

$$r_{\mathcal{I},i}^{\mathcal{J}} \geq \min_{j \in \mathcal{M}_i \cup \{l\}} \{ \max\{r_{\mathcal{I}^j}, r_{\text{ex}}^{\mathcal{I}^j} - i\} \},$$

where $\mathcal{M}_i := \{j \in \mathbb{N} \mid 1 \leq j \leq l-1, \max\{r_{\mathcal{I}^j}, r_{\text{ex}}^{\mathcal{I}^j} - i\} < \min_{j+1 \leq m \leq l} \{r_{\mathcal{I}^m}\}\}$. ♣

1.2.3 Superconvergence in time mesh points

The \mathcal{J}_n -VTD $_k^r(I_n)$ methods described by (1.2) show some superconvergence behavior in the time mesh points. More concretely, in many cases the convergence order of the error in the time mesh points is considerably larger than the convergence order for the pointwise error. The following statement can be proven.

Theorem 1.15 (Superconvergence estimate, cf. [16, Theorem 18])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Suppose that the Assumptions 1.1, 1.2, and 1.3 hold. Moreover, denote by u and U the solutions of (1.1) and (1.2), respectively. Suppose that (for τ sufficiently small) the global error $\sup_{t \in I} \|(u - U)(t)\|$, as well as U and all of its derivatives, can be bounded independent of the mesh parameter. Then, we have for $1 \leq n \leq N$

$$\|(u - U)(t_n^-)\| \leq C(F, u) \left(\sup_{t \in [t_0, t_n]} \|(u - U)(t)\|^2 + \tau^{\min\{2r-k+1, r_{\text{var}}^{\mathcal{J}\mathcal{I}}+1, \max\{r_{\text{ex}}^{\mathcal{J}}+1, \min\{r, r_{\mathcal{I}}^{\mathcal{J}}+1\}\}\}} \right),$$

where $r_{\text{var}}^{\mathcal{J}\mathcal{I}} := \min_{0 \leq i \leq r-k} \{r_{\mathcal{I},i}^{\mathcal{J}} + i\}$.

Remark 1.16

The term $\sup_{t \in [t_0, t_n]} \|(u - U)(t)\|^2$ in the estimate of Theorem 1.15 may be dropped under certain conditions. For more details on this, see [16, Remark 19]. ♣

Remark 1.17

While a bound for the global error $\sup_{t \in I} \|(u - U)(t)\|$ could be derived from Theorem 1.7, also see Corollary 1.13, it is not directly clear how to guarantee in Theorem 1.15 that U and all of its derivatives can be bounded independent of the mesh parameter. However, provided that Assumption 1.1 holds and a uniform bound for the global error is known, it is shown in [16, Lemma 20] that $\sup_{t \in I} \|U^{(l)}(t)\| \leq C(F, u)$ for all $l \geq 0$ if $r_{\mathcal{I}}^{\mathcal{J}} \geq r - 2$. ♣

Remark 1.18 (Superconvergence of derivative(s) in time mesh points, cf. [14, Remark 4.10])

From the point conditions (1.2b) and the bound of Theorem 1.15 we also gain superconvergence estimates up to the $\lfloor \frac{k}{2} \rfloor$ th derivative of the solution U of \mathcal{J}_n -**VTD** $_k^r(\mathcal{I}_n)$ in t_n^- , provided that F satisfies Assumption 1.4. Indeed, we find for $1 \leq n \leq N$

$$\begin{aligned} \|(u - U)^{(i+1)}(t_n^-)\| &= \left\| \frac{d^i}{dt^i} \left(M^{-1}F(t, u(t)) - M^{-1}F(t, U(t)) \right) \right\|_{t=t_n^-} \\ &\leq C \sum_{j=0}^i \|(u - U)^{(j)}(t_n^-)\| \leq \dots \leq C \|(u - U)(t_n^-)\| \end{aligned}$$

by iteration over $i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1$. ♣

Summarizing the above observations, the following estimates in the time mesh points can be stated.

Corollary 1.19 (Cf. [16, Corollary 21])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Suppose that Assumptions 1.1, 1.2, 1.3, and 1.4 hold. Moreover, let Assumption 1.5a or 1.5b be satisfied. Denote by u and U the solutions of (1.1) and (1.2), respectively. Then, if $r_{\mathcal{I}}^{\mathcal{J}} \geq r - 2$, we have for $1 \leq n \leq N$

$$\|(u - U)(t_n^-)\| \leq C(F, u) \left(\tau^{\min\{2r-k+1, r_{\text{var}}^{\mathcal{J}}+1, \max\{r_{\text{ex}}^{\mathcal{J}}+1, \min\{r, r_{\mathcal{I}}^{\mathcal{J}}+1\}\}}\}} + \delta_{0,k} \tau^{2r_{\mathcal{I}}^{\mathcal{J}}+4} \right) \quad (1.12)$$

with $r_{\text{var}}^{\mathcal{J}\mathcal{I}} := \min_{0 \leq i \leq r-k} \{r_{\mathcal{I},i}^{\mathcal{J}} + i\}$, $r_{\mathcal{I}}^{\mathcal{J}} = r_{\mathcal{I},r-k}^{\mathcal{J}}$, and $r_{\mathcal{I},i}^{\mathcal{J}}$ as defined in Definition 1.10.

If $r_{\mathcal{I}}^{\mathcal{J}} < r - 2$, the uniform boundedness of U and all its derivatives cannot be ensured in general. Then, we only have

$$\|(u - U)(t_n^-)\| \leq \sup_{t \in I_n} \|(u - U)(t)\|,$$

where we refer to Corollary 1.13 for bounds on the right-hand side term.

1.2.4 Numerical results

In this subsection, we want to show that the estimates of Corollary 1.13 and Corollary 1.19 are sharp. To this end, the error in the norms

$$\|v\|_{L^\infty} = \text{ess sup}_{t \in I} \|v(t)\|, \quad \|v\|_{W^{1,\infty}} = \max_{0 \leq l \leq 1} \text{ess sup}_{t \in I} \|v^{(l)}(t)\|, \quad \|v\|_{\ell^\infty} = \max_{1 \leq n \leq N} \|v(t_n^-)\|$$

should be investigated numerically. Appropriate numerical studies have been already made in [16, Section 6]. However, for completeness we give a short summary of the obtained numerical results here.

Example

We consider the initial value problem

$$\begin{pmatrix} u_1'(t) \\ u_2'(t) \end{pmatrix} = \begin{pmatrix} -u_1^2(t) - u_2(t) \\ u_1(t) - u_1(t)u_2(t) \end{pmatrix}, \quad t \in (0, 32), \quad u(0) = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \quad (1.13)$$

of a system of nonlinear ordinary differential equations that has

$$u_1(t) = \frac{\cos t}{2 + \sin t}, \quad u_2(t) = \frac{\sin t}{2 + \sin t}$$

as solution.

The appearing nonlinear systems within each time step were solved by Newton's method where a Taylor expansion of the inherited data from the previous time interval was applied to calculate an initial guess for all unknowns on the current interval. If higher order derivatives were needed at initial time $t = 0$, the ode system and its temporal derivatives were used, see (1.4).

According to Corollary 1.13 and Corollary 1.19, we expect the following orders of convergence

$$\text{"}W^{1,\infty}\text{"-order"} = \min\{r, r_{\mathcal{I}}^{\mathcal{J}} + 1\}, \quad (1.14a)$$

$$\text{"}L^{\infty}\text{"-order"} = \min\{r + 1, r_{\mathcal{I}}^{\mathcal{J}} + 2, r_{\mathcal{I},0}^{\mathcal{J}} + 1, \max\{r_{\text{ex}}^{\mathcal{J}} + 1, \min\{r, r_{\mathcal{I}}^{\mathcal{J}} + 1\}\}\}, \quad (1.14b)$$

$$\text{"}\ell^{\infty}\text{"-order"} = \min\{2r - k + 1, r_{\text{var}}^{\mathcal{I}} + 1, \max\{r_{\text{ex}}^{\mathcal{J}} + 1, \min\{r, r_{\mathcal{I}}^{\mathcal{J}} + 1\}\}\} \quad (1.14c)$$

for the error in the $W^{1,\infty}$ -norm, the L^{∞} -norm, and the ℓ^{∞} -norm. However, recall that there are the additional conditions $\max\{r_{\text{ex}}^{\mathcal{J}}, r_{\mathcal{I}}^{\mathcal{J}} + 1\} \geq r - 1$ for the L^{∞} -estimate and $r_{\mathcal{I}}^{\mathcal{J}} \geq r - 2$ for the ℓ^{∞} -estimate, respectively.

In order to verify these theoretical results, a wide variety of integrators and interpolators needs to be studied. Here we always consider \mathcal{J}_n -**VTD**₃⁶(\mathcal{I}_n) methods, which are variants of cGP- $C^1(6)$, as discretization of (1.13) where \mathcal{J}_n and \mathcal{I}_n are obtained from given reference operators $\hat{\mathcal{J}}$ and $\hat{\mathcal{I}}$ via transformation. Each integrator $\hat{\mathcal{J}}$ and each interpolation operator $\hat{\mathcal{I}}$ that is studied is based on Lagrangian interpolation with respect to a specific node set $P_{\hat{\mathcal{J}}}$ and $P_{\hat{\mathcal{I}}}$, respectively. Hence, we have $k_{\mathcal{J}} = k_{\mathcal{I}} = 0$. Both node sets are given for each of the test cases. Since often nodes of quadrature formulas are used, we also write for instance "left Gauss–Radau(k)" to indicate that the nodes of the left-sided Gauss–Radau formula with k points have been used. All upcoming settings fulfill Assumption 1.1.

The different test cases are listed in Table 1.1. Beyond the node sets for integrator and interpolation operator also the associated theoretical expressions for the orders of convergence are presented. Note that the limiting terms are always indicated in boldface. The expressions for the L^{∞} -order or the ℓ^{∞} -order are struck out if the conditions $\max\{r_{\text{ex}}^{\mathcal{J}}, r_{\mathcal{I}}^{\mathcal{J}} + 1\} \geq r - 1$ or $r_{\mathcal{I}}^{\mathcal{J}} \geq r - 2$, respectively, are not fulfilled.

In the first test case both conditions are violated such that only the $W^{1,\infty}$ -estimate holds and gives order 3 while the L^{∞} - and ℓ^{∞} -estimates would yield order 4. The case group 2 provides choices for $P_{\hat{\mathcal{J}}}$ and $P_{\hat{\mathcal{I}}}$ that show that the L^{∞} -convergence order can be limited by each of the three terms occurring in the maximum expression inside the outer minimum

Table 1.1: Test cases with their node sets and theoretically expected orders of convergence

case	node sets			theoretically expected orders of convergence		
	$P_{\hat{\mathcal{G}}}$	$P_{\hat{\mathcal{T}}}$	$W^{1,\infty}$ -order	L^∞ -order	ℓ^∞ -order	
1	$\{-\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{3}{4}\}$	left Gauss–Radau(3)	$\min\{6, 3\}$	$\min\{7, 4, 4, \max\{4, \min\{6, 3\}\}\}$	$\min\{10, 4, \max\{4, \min\{6, 3\}\}\}$	
2a	$\{-\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{3}{4}\}$	Gauss(5)	$\min\{6, 5\}$	$\min\{7, 6, 6, \max\{4, \min\{6, 5\}\}\}$	$\min\{10, 6, \max\{4, \min\{6, 5\}\}\}$	
2a*	$\{-\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{3}{4}\}$	$\{-\frac{5}{6}, -\frac{13}{23}, \frac{1}{10}, \frac{12}{17}, \frac{4}{5}\}$	$\min\{6, 5\}$	$\min\{7, 6, 6, \max\{4, \min\{6, 5\}\}\}$	$\min\{10, 6, \max\{4, \min\{6, 5\}\}\}$	
2b	$\{-\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{3}{4}\}$	$\{-\frac{3}{4}, -\frac{1}{4}, \frac{1}{4}, \frac{3}{4}\}$	$\min\{6, \infty\}$	$\min\{7, \infty, \infty, \max\{4, \min\{6, \infty\}\}\}$	$\min\{10, \infty, \max\{4, \min\{6, \infty\}\}\}$	
2c	$\{-1, -\frac{3}{5}, -\frac{1}{5}, \frac{3}{5}, 1\}$	$\{-1, -\frac{3}{5}, -\frac{1}{5}, \frac{3}{5}, 1\}$	$\min\{6, \infty\}$	$\min\{7, \infty, \infty, \max\{6, \min\{6, \infty\}\}\}$	$\min\{10, \infty, \max\{6, \min\{6, \infty\}\}\}$	
3a	Gauss(6)	$\{-1, -\frac{1}{2}, \frac{1}{4}, \frac{3}{4}, 1\}$	$\min\{6, 5\}$	$\min\{7, 6, 5, \max\{12, \min\{6, 5\}\}\}$	$\min\{10, 5, \max\{12, \min\{6, 5\}\}\}$	
3b	Gauss(6)	left Gauss–Radau(3)	$\min\{6, 3\}$	$\min\{7, 4, 5, \max\{12, \min\{6, 3\}\}\}$	$\min\{10, 5, \max\{12, \min\{6, 3\}\}\}$	
3c	Gauss(6)	Gauss(5)	$\min\{6, 7\}$	$\min\{7, 8, 10, \max\{12, \min\{6, 7\}\}\}$	$\min\{10, 10, \max\{12, \min\{6, 7\}\}\}$	
4a	Gauss(6)	Gauss–Lobatto(5)	$\min\{6, 5\}$	$\min\{7, 6, 8, \max\{12, \min\{6, 5\}\}\}$	$\min\{10, 8, \max\{12, \min\{6, 5\}\}\}$	
4b	Gauss(6)	Gauss(6)	$\min\{6, \infty\}$	$\min\{7, \infty, \infty, \max\{12, \min\{6, \infty\}\}\}$	$\min\{10, \infty, \max\{12, \min\{6, \infty\}\}\}$	
4c	Gauss(4)	Gauss(4)	$\min\{6, \infty\}$	$\min\{7, \infty, \infty, \max\{8, \min\{6, \infty\}\}\}$	$\min\{10, \infty, \max\{8, \min\{6, \infty\}\}\}$	
4d	Gauss(6)	Gauss(3)	$\min\{6, 3\}$	$\min\{7, 4, 6, \max\{12, \min\{6, 3\}\}\}$	$\min\{10, 6, \max\{12, \min\{6, 3\}\}\}$	

Table 1.2: Error of $\mathcal{J}_n\text{-VTD}_3^6(\mathcal{I}_n)$ in different (semi-)norms and associated (experimental) convergence orders

case	$\ u - U\ _{L^\infty}$	eoc (theo)	$\ (u - U)'\ _{L^\infty}$	eoc (theo)	$\ u - U\ _{\ell^\infty}$	eoc (theo)
1	1.615e-06	3.04	2.369e-05	3.00	1.559e-06	3.01
	1.961e-07	(3)	2.965e-06	(3)	1.937e-07	(3)
2a	3.759e-11	6.00	3.863e-09	5.00	7.003e-12	6.00
	5.862e-13	(5)	1.208e-10	(5)	1.096e-13	(5)
2a*	9.412e-11	5.08	4.048e-09	5.00	8.666e-11	5.00
	2.775e-12	(5)	1.266e-10	(5)	2.699e-12	(5)
2b	1.465e-12	6.07	6.841e-11	6.00	1.354e-12	6.00
	2.175e-14	(6)	1.072e-12	(6)	2.122e-14	(6)
2c	6.604e-12	6.00	1.130e-10	5.99	6.601e-12	6.00
	1.034e-13	(6)	1.773e-12	(6)	1.034e-13	(6)
3a	1.716e-10	5.19	1.072e-08	5.00	1.459e-10	5.02
	4.688e-12	(5)	3.353e-10	(5)	4.500e-12	(5)
3b	3.069e-07	4.02	3.581e-05	3.00	6.011e-08	4.12
	1.886e-08	(4)	4.479e-06	(3)	3.455e-09	(4)
3c	4.068e-13	7.00	7.155e-11	6.00	6.689e-19	10.00
	3.181e-15	(7)	1.119e-12	(6)	6.529e-22	(10)
4a	4.464e-11	6.00	5.648e-09	5.00	5.120e-15	8.00
	6.981e-13	(6)	1.766e-10	(5)	2.005e-17	(8)
4b	4.068e-13	7.00	7.155e-11	6.00	5.318e-19	10.00
	3.181e-15	(7)	1.119e-12	(6)	5.192e-22	(10)
4c	8.654e-13	7.00	1.288e-10	6.00	2.565e-15	8.00
	6.759e-15	(7)	2.014e-12	(6)	1.002e-17	(8)
4d	2.028e-07	4.00	2.111e-05	3.00	4.377e-08	4.00
	1.268e-08	(4)	2.641e-06	(3)	2.735e-09	(4)

in (1.14b). Hereby, note that it is not possible that $r_{\text{ex}}^{\mathcal{J}} + 1$ is the only limiting term since the structure of (1.14b) implies that $\min\{r + 1, r_{\mathcal{I}}^{\mathcal{J}} + 2\} \geq r_{\text{ex}}^{\mathcal{J}} + 1 \geq \min\{r, r_{\mathcal{I}}^{\mathcal{J}} + 1\}$ if $r_{\text{ex}}^{\mathcal{J}} + 1$ is limiting. Hence, the integer $r_{\text{ex}}^{\mathcal{J}} + 1$ coincides either with $\min\{r + 1, r_{\mathcal{I}}^{\mathcal{J}} + 2\}$ or $\min\{r, r_{\mathcal{I}}^{\mathcal{J}} + 1\}$. Case group 3 shows that each of the first three expressions in the outer minimum in (1.14b) can bound the L^∞ -order and that the $W^{1,\infty}$ -order can be limited by both occurring terms in (1.14a). With case group 4 we consider settings where the convergence order in the

ℓ^∞ -norm suggested by (1.14c) is strictly greater than the L^∞ -order given by (1.14b). It is shown that the first two arguments in the minimum in (1.14c) and the first argument inside the maximum there can limit the ℓ^∞ -convergence order. Moreover, we consider a case where the higher superconvergence order cannot be expected since $r_{\mathcal{I}}^{\mathcal{J}} < r - 2$.

Computational results for all the different test cases are given in Table 1.2. All calculations were carried out with the software Julia [18] using the floating point data type `BigFloat` with 512 bits. We present the errors in different (semi-)norms obtained for 256 and 512 time steps and also give the experimental orders of convergence (eoc) calculated from these two errors. For comparison, in addition the theoretically predicted convergence orders are given in brackets.

The numerical results confirm the expected convergence behavior. The only exception is case 2a, where we see an experimental order of convergence of 6, which is one order higher than expected. This discrepancy can be explained by a closer look to Lemma 1.12. For its proof a splitting is used whose single terms only vanish for all $v \in P_5(I_n)$. However, in case 2a, due to symmetry reasons, it holds $\int_{I_n} (v - \mathcal{J}_n^{\mathcal{I}} v)'(t) dt = 0$ for all $v \in P_6(I_n)$ and so $(v - \mathcal{J}_n^{\mathcal{I}} v)(t_n^-) = 0$ for all $v \in P_6(I_n)$. Thus, the convergence order of the limiting term is actually better than predicted. For a more detailed discussion of this and all other cases, we refer to [16, Section 6].

1.3 Associated quadrature formulas and their advantages

In order to obtain a fully computable discrete problem, usually a quadrature formula Q_n is chosen as integrator, i.e., $\mathcal{J}_n = Q_n$. To indicate this choice, we simply write $Q_n\text{-}\mathbf{VTD}_k^r(\mathcal{I}_n)$. Moreover, recall that integration over I_n is used if no quadrature rule is specified and that the specification of \mathcal{I}_n is omitted if $\mathcal{I}_n = \text{Id}$. We shall mostly use quadrature rules that are exact for polynomials of degree up to $2r - k$. This ensures in the case of an affine linear right-hand side $F(t, u) = f(t) - Au$ with time-independent A that at least all u depending terms in (1.2d) are integrated exactly.

1.3.1 Special quadrature formulas

The special structure of the method (1.2) motivates to use an assigned interpolation operator that conserves derivatives at the end points of the interval up to a certain order. In detail, we define on $[-1, 1]$ the reference interpolation operator $\hat{\mathcal{I}}_k^r : C^{\lfloor \frac{k}{2} \rfloor}([-1, 1]) \rightarrow P_r([-1, 1])$ that uses the interpolation points

$$\begin{aligned} \text{at the left end:} & \quad \text{derivatives up to order } \lfloor \frac{k-1}{2} \rfloor \text{ in } -1^+, \\ \text{at the right end:} & \quad \text{derivatives up to order } \lfloor \frac{k}{2} \rfloor \text{ in } 1^-, \\ \text{in the interior:} & \quad \text{zeros } \hat{t}_i \in (-1, 1) \text{ of the } (r-k)\text{th Jacobi-polynomial} \\ & \quad \text{with respect to the weight } (1+\hat{t})^{\lfloor \frac{k-1}{2} \rfloor+1} (1-\hat{t})^{\lfloor \frac{k}{2} \rfloor+1}. \end{aligned} \tag{1.15}$$

Note that there is no point evaluation at the left end for $k = 0$. In any case, the number of interpolation conditions is

$$r - k + \left\lfloor \frac{k}{2} \right\rfloor + 1 + \left\lfloor \frac{k-1}{2} \right\rfloor + 1 = r - k + k - 1 + 2 = r + 1$$

and, thus, coincides with the dimension of P_r . The interpolation operator $\hat{\mathcal{I}}_k^r$ is of Hermite-type and provides the standard error estimates for Hermite interpolation, see e.g. [51, (2.1.5.9) Theorem, p. 57].

In addition, we define by

$$\hat{Q}_k^r[\hat{\varphi}] := \int_{-1}^1 (\hat{\mathcal{I}}_k^r \hat{\varphi})(\hat{t}) d\hat{t}$$

a quadrature rule on $[-1, 1]$ that is in a natural way assigned to the method \mathbf{VTD}_k^r . The quadrature rules \hat{Q}_k^r are known in the literature as generalized Gauss–Radau or Gauss–Lobatto formulas, respectively, see e.g. [32, 44]. The weights of the quadrature rule \hat{Q}_k^r could be calculated by integrating the appropriate Hermite basis functions on $[-1, 1]$. Finally, we obtain

$$\int_{-1}^1 \hat{\varphi}(\hat{t}) d\hat{t} \approx \hat{Q}_k^r[\hat{\varphi}] = \int_{-1}^1 (\hat{\mathcal{I}}_k^r \hat{\varphi})(\hat{t}) d\hat{t} = \sum_{i=0}^{\left\lfloor \frac{k-1}{2} \right\rfloor} w_i^L \hat{\varphi}^{(i)}(-1^+) + \sum_{i=1}^{r-k} w_i^I \hat{\varphi}(\hat{t}_i) + \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} w_i^R \hat{\varphi}^{(i)}(+1^-).$$

The quadrature rule \hat{Q}_k^r is exact for polynomials up to degree $2r - k$. It can be shown that all quadrature weights are different from zero, see [44]. More precisely, we have

$$w_j^I > 0, \quad w_j^L > 0, \quad (-1)^j w_j^R > 0, \quad (1.16)$$


so even the sign of the weights is known. Note that in general (for $k \geq 2$) not all weights are positive. Semi-explicit or recursive formulas for the weights of these methods can be found in [48].

Transferring the quadrature rule \hat{Q}_k^r and the interpolation operator $\hat{\mathcal{I}}_k^r$ from $[-1, 1]$ to the interval \bar{I}_n , we obtain $Q_{k,n}^r$ and $\mathcal{I}_{k,n}^r$. We usually skip n in the notation since the relation to I_n will mostly be clear from context. Hence, we have

$$\int_{I_n} \varphi(t) dt \approx Q_k^r[\varphi] = \frac{\tau_n}{2} \left[\sum_{i=0}^{\left\lfloor \frac{k-1}{2} \right\rfloor} w_i^L \left(\frac{\tau_n}{2}\right)^i \varphi^{(i)}(t_{n-1}^+) + \sum_{i=1}^{r-k} w_i^I \varphi(t_{n,i}) + \sum_{i=0}^{\left\lfloor \frac{k}{2} \right\rfloor} w_i^R \left(\frac{\tau_n}{2}\right)^i \varphi^{(i)}(t_n^-) \right],$$

where $t_{n,i} = \frac{t_n + t_{n-1}}{2} + \frac{\tau_n}{2} \hat{t}_i \in I_n$, $i = 1, \dots, r - k$.

Remark 1.20

The quadrature rule Q_0^r is the well-known right-sided Gauss–Radau quadrature formula with $r + 1$ points, which is typically used for the discontinuous Galerkin method dG(r). Q_1^r is the Gauss–Lobatto quadrature rule with $r + 1$ points, which is often used together with the continuous Galerkin–Petrov method cGP(r). 

1.3.2 Postprocessing

The quadrature formulas defined by the quadrature points (1.15) enable a simple postprocessing, which shall be presented in this subsection. Postprocessing techniques for dG and cGP methods have been introduced in [46] and were generalized to the whole family of variational time discretizations in [14]. The postprocessing creates an improved solution where the global smoothness is increased by one differentiation order if F is $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable on I . Moreover, the postprocessing lifts the originally obtained numerical solution on each time subinterval to the polynomial space with one degree higher. This results in an increased accuracy and mostly an improved convergence by one order for the pointwise error.

The postprocessing can be formulated as follows. For the proofs we refer to [14, Section 3].

Theorem 1.21 (Postprocessing $Q_k^r\text{-VTD}_k^r \rightsquigarrow Q_k^r\text{-VTD}_{k+2}^{r+1}$, cf. [14, Theorem 3.1])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and suppose that $U \in Y_r$ solves $Q_k^r\text{-VTD}_k^r$. For every $n = 1, \dots, N$ set

$$\tilde{U}|_{I_n} = U|_{I_n} + a_n \vartheta_n, \quad \vartheta_n \in P_{r+1}(I_n, \mathbb{R}),$$

where ϑ_n vanishes in the $(r+1)$ quadrature points of Q_k^r and satisfies $\vartheta_n^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-) = 1$ while the vector $a_n \in \mathbb{R}^d$ is defined by

$$a_n = M^{-1} \left(\frac{d \lfloor \frac{k}{2} \rfloor}{dt \lfloor \frac{k}{2} \rfloor} F(t, U(t)) \Big|_{t=t_n^-} - MU^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-) \right). \quad (1.17)$$

Moreover, let $\tilde{U}(t_0^-) = U(t_0^-)$. Then, $\tilde{U} \in Y_{r+1}$ solves $Q_k^r\text{-VTD}_{k+2}^{r+1}$.

From the definition (1.17), it seems that a linear system with the mass matrix M has to be solved in every time step in order to obtain the correction vector a_n . However, the computational costs for calculating a_n can be reduced significantly if F is sufficiently smooth as shown in the following proposition.

Proposition 1.22 (Cf. [14, Proposition 3.2])

Suppose that F is $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable on \bar{I} . Then, the correction vectors $a_n \in \mathbb{R}^d$ defined in (1.17) for the postprocessing presented in Theorem 1.21 can be alternatively calculated by

$$a_n = \frac{-1}{\vartheta_n^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+)} \left(U^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+) - \tilde{U}^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^-) \right) \quad \text{for } n > 1,$$

and

$$a_1 = \frac{-1}{\vartheta_1^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_0^+)} \left(U^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_0^+) - u^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_0) \right),$$

where $u^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_0)$ is defined in (1.4).

Note that a_n can be calculated in this way without solving a system of linear equations and, thus, with almost no computational costs. From the structure of a_n we see that the postprocessing can be interpreted as a correction of the jump in the lowest order derivative of the discrete solution that is not continuous by construction.

Since the division by $\vartheta_n^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+)$ changes the normalization of ϑ_n only, we gain the following.

Corollary 1.23 (Alternative postprocessing $Q_k^r\text{-VTD}_k^r \rightsquigarrow Q_k^r\text{-VTD}_{k+2}^{r+1}$, cf. [14, Corollary 3.3])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and suppose that $U \in Y_r$ solves $Q_k^r\text{-VTD}_k^r$. For every $n = 1, \dots, N$ set

$$\tilde{U}|_{I_n} = U|_{I_n} - \tilde{a}_n \tilde{\vartheta}_n, \quad \tilde{\vartheta}_n \in P_{r+1}(I_n, \mathbb{R}),$$

where $\tilde{\vartheta}_n(t) = \vartheta_n(t)/\vartheta_n^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+)$ with ϑ_n from Theorem 1.21, i.e., $\tilde{\vartheta}_n$ vanishes in all $(r+1)$ quadrature points of Q_k^r and satisfies $\tilde{\vartheta}_n^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+) = 1$. The vector $\tilde{a}_n \in \mathbb{R}^d$ is defined by

$$\tilde{a}_n := \begin{cases} U^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_0^+) - u^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_0), & n = 1, \\ U^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+) - \tilde{U}^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^-), & n > 1, \end{cases}$$

where $u^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_0)$ is given by (1.4). Moreover, let $\tilde{U}(t_0^-) = U(t_0^-)$. Then, if F is $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable on \bar{I} , we have that $\tilde{U} \in Y_{r+1}$ solves $Q_k^r\text{-VTD}_{k+2}^{r+1}$.

1.3.3 Connections to collocation methods

In this subsection, we see that the (local) solution of $Q_k^r\text{-VTD}_l^{r+1}$ with $1 \leq l \leq k+2$, which obviously includes $Q_k^r\text{-VTD}_{k+2}^{r+1}$, can be characterized as the solution of the (local) collocation problem with multiple nodes, as known e.g. from [37, p. 275], with respect to the quadrature points of Q_k^r , i.e.,

Given $\tilde{U}(t_{n-1}^-) \in \mathbb{R}^d$, find $\tilde{U} \in P_{r+1}(I_n, \mathbb{R}^d)$ such that $\tilde{U}(t_{n-1}^+) = \tilde{U}(t_{n-1}^-)$ and

$$M\tilde{U}^{(i+1)}(t_n^-) = \frac{d^i}{dt^i} \left(F(t, \tilde{U}(t)) \right) \Big|_{t=t_n^-}, \quad i = 0, \dots, \lfloor \frac{k}{2} \rfloor, \quad (1.18a)$$

$$M\tilde{U}^{(i+1)}(t_{n-1}^+) = \frac{d^i}{dt^i} \left(F(t, \tilde{U}(t)) \right) \Big|_{t=t_{n-1}^+}, \quad \text{if } k \geq 1, i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor, \quad (1.18b)$$

$$M\tilde{U}'(t_{n,i}) = F(t_{n,i}, \tilde{U}(t_{n,i})), \quad i = 1, \dots, r-k, \quad (1.18c)$$

where $\tilde{U}(t_0^-) = u_0$. Here, $t_{n,i} = \frac{t_n + t_{n-1}}{2} + \frac{\tau_n}{2} \hat{t}_i$, where \hat{t}_i denote the zeros of the $(r-k)$ th Jacobi-polynomial with respect to the weight $(1+\hat{t})^{\lfloor \frac{k-1}{2} \rfloor + 1} (1-\hat{t})^{\lfloor \frac{k}{2} \rfloor + 1}$, see also (1.15).

The following connection was found and proven in [14].


Theorem 1.24 (Equivalence to collocation methods, cf. [14, Theorem 4.1])

Let $r, k, l \in \mathbb{Z}$, $0 \leq k \leq r$, and $1 \leq l \leq k + 2$. Then, $\tilde{U} \in P_{r+1}(I_n, \mathbb{R}^d)$ solves $Q_k^r\text{-VTD}_l^{r+1}$ if and only if \tilde{U} solves the collocation method (1.18) with respect to the quadrature points of Q_k^r .

Summarizing, we have that every solution of $Q_k^r\text{-VTD}_{k+2}^{r+1}$ also solves $Q_k^r\text{-VTD}_l^{r+1}$ with $1 \leq l \leq k + 2$ as well as a collocation with respect to the quadrature points of Q_k^r and vice versa. This can be shortly described as

$$\begin{aligned} Q_k^r\text{-VTD}_{k+2}^{r+1} &\hat{=} Q_k^r\text{-VTD}_l^{r+1} \quad \text{with} \quad 1 \leq l \leq k + 2 \\ &\hat{=} \text{collocation with respect to the quadrature points of } Q_k^r. \end{aligned}$$

Remark 1.25

Independent of the above findings, the connection between collocation methods and (post-processed) numerically integrated discontinuous Galerkin methods (using the right-sided Gauss–Radau quadrature), i.e., Theorem 1.24 for $k = 0 \leq r$ and $l = 2$, was already observed in [53]. Moreover, connections between collocation methods and the numerically integrated continuous Galerkin–Petrov methods (using interpolatory quadrature formulas with as many quadrature points as number of independent variational conditions) were shown in [40, 41]. Certain equivalences between collocation methods and dG or cGP methods have also been discussed in [26, Proposition 70.7]. 

1.3.4 Shortcut to error estimates

Error estimates for collocation methods with multiple nodes, as defined e.g. in [37, p. 275], are well-known provided that F and u satisfy certain (regularity) assumptions. Unfortunately, these conditions on F and u are often not explicitly given in the literature. Nevertheless, according to [37, p. 276, pp. 212–214], we shall state various error bounds for the solution of (1.18) without specifying these assumptions. Moreover, global error estimates can be derived by adapting techniques presented in [40, Theorem 2].

Proposition 1.26

Let \tilde{U} denote the solution of the collocation method (1.18) and u the exact solution of (1.1). Then, assuming that F and u satisfy certain (regularity) assumptions, we have

$$\max_{1 \leq n \leq N} \|(u - \tilde{U})(t_n^-)\| \leq C(F, u) \tau^{2r-k+1} \quad (1.19)$$

and

$$\sup_{t \in I_n} \|(u - \tilde{U})^{(l)}(t)\| \leq C(F, u) \tau^{\min\{2r-k+1, (r+1)+1-l\}}, \quad 0 \leq l \leq r + 1, \quad (1.20)$$

for all $1 \leq n \leq N$.

The term $2r - k + 1$ inside the minimum is due to the fact that the convergence order of the collocation method is limited by the accuracy of the underlying quadrature formula Q_k^r that is exactly $2r - k + 1$. Note that the limitation is active for $r = k$ and $l = 0$ only.

Exploiting the equivalence of $Q_k^r\text{-VTD}_{k+2}^{r+1}$ and the collocation method (1.18) as well as the connection between $Q_k^r\text{-VTD}_k^r$ and $Q_k^r\text{-VTD}_{k+2}^{r+1}$ through the postprocessing and its reversion by interpolation in the quadrature points of Q_k^r (for details see [14, Proposition 4.5]), we immediately also gain various results for the $Q_k^r\text{-VTD}_k^r$ method.

Corollary 1.27 (Existence and uniqueness, cf. [14, Corollary 4.6])

If there is a solution $\tilde{U} \in P_{r+1}(I_n, \mathbb{R}^d)$ of the collocation method with multiple nodes defined by (1.18), then $U = \mathcal{I}_k^r \tilde{U} \in P_r(I_n, \mathbb{R}^d)$ solves $Q_k^r\text{-VTD}_k^r$. Furthermore, if \tilde{U} is uniquely defined as solution of (1.18), then so is U as solution of $Q_k^r\text{-VTD}_k^r$.

Corollary 1.28 (Global error estimates, cf. [14, Corollary 4.7])

Let (1.20) hold for the solution \tilde{U} of (1.18) and the exact solution u of (1.1). Then, we have for the solution U of $Q_k^r\text{-VTD}_k^r$ and $0 \leq l \leq r$ that

$$\sup_{t \in I_n} \|(u - U)^{(l)}(t)\| \leq \sup_{t \in I_n} \|(u - \tilde{U})^{(l)}(t)\| + \sup_{t \in I_n} \|(\tilde{U} - \mathcal{I}_k^r \tilde{U})^{(l)}(t)\| \leq C(F, u) \tau^{r+1-l}$$

for all $1 \leq n \leq N$.

Corollary 1.29 (Superconvergence in time mesh points, cf. [14, Corollary 4.8])

Let (1.19) hold for the solution \tilde{U} of (1.18) and the exact solution u of (1.1). Then, we have

$$\max_{1 \leq n \leq N} \|(u - U)(t_n^-)\| = \max_{1 \leq n \leq N} \|(u - \tilde{U})(t_n^-)\| \leq C(F, u) \tau^{2r-k+1}$$

for the solution U of $Q_k^r\text{-VTD}_k^r$.

Remark 1.30 (Superconvergence in quadrature points, cf. [14, Remark 4.9])

We obtain under the assumptions of Corollary 1.28 also a (lower order) superconvergence estimate for the solution U of $Q_k^r\text{-VTD}_k^r$ in the quadrature points of Q_k^r if $0 \leq k < r$. In fact, let $t_{n,i}$, $i = 1, \dots, r - k$, denote the local quadrature points of Q_k^r in the interior of I_n . Then, we have for $1 \leq n \leq N$

$$\|(u - U)(t_{n,i})\| = \|(u - \tilde{U})(t_{n,i})\| \leq C(F, u) \tau^{(r+1)+1}.$$

In addition, we obtain

$$\|(u - U)^{(l)}(t_n^-)\| = \|(u - \tilde{U})^{(l)}(t_n^-)\| \leq C(F, u) \tau^{(r+1)+1-l}, \quad 0 \leq l \leq \left\lfloor \frac{k}{2} \right\rfloor,$$

and

$$\|(u - U)^{(l)}(t_{n-1}^+)\| = \|(u - \tilde{U})^{(l)}(t_{n-1}^+)\| \leq C(F, u) \tau^{(r+1)+1-l}, \quad 0 \leq l \leq \left\lfloor \frac{k-1}{2} \right\rfloor,$$

provided $k \geq 1$.

These superconvergence estimates especially imply

$$\left(\sum_{n=1}^N Q_{k,n}^r [\|u - U\|^2] \right)^{1/2} = \left(\sum_{n=1}^N Q_{k,n}^r [\|u - \tilde{U}\|^2] \right)^{1/2} \leq (t_N - t_0)^{1/2} C(F, u) \tau^{(r+1)+1},$$

which compared to

$$\left(\sum_{n=1}^N \int_{I_n} \|(u - U)(t)\|^2 dt \right)^{1/2} \leq (t_N - t_0)^{1/2} C(F, u) \tau^{r+1}$$

gives an extra order of convergence. ♣

1.3.5 Numerical results

In this subsection, we want to illustrate the effects of postprocessing by some computational results. Hereby, we draw on the numerical data of [15, Section 7]. As in Subsection 1.2.4 we consider the initial value problem (1.13) as test example. Moreover, for the calculations the software Julia [18] have been used with floating point data type `BigFloat` with 512 bits.

We are interested in the error of the discrete solution U and the error of the postprocessed solution \tilde{U} where the postprocessing is determined using the jumps of the derivatives, as given in Corollary 1.23. The errors are measured in the norms

$$\|v\|_{L^2} = \left(\int_{t_0}^{t_N} \|v(t)\|^2 dt \right)^{1/2}, \quad \|v\|_{\ell^\infty} = \max_{1 \leq n \leq N} \|v(t_n^-)\|$$

with $\|\cdot\|$ denoting the Euclidean norm in \mathbb{R}^d .

Numerical results for $Q_k^6\text{-VTD}_k^6$ with $k = 0, 5, 6$ are presented in Table 1.3. The given errors are those obtained for 256 and 512 time steps. In addition, the associated experimental orders of convergence (eoc) and the theoretically predicted convergence orders (theo) are listed.

Overall, our theoretical findings are well confirmed by the numerical data. The error of the discrete solution as well as the error of the postprocessed solution show the expected (super-)convergence orders. Moreover, the postprocessing yields the predicted improvements.

Especially note that, as expected from Remark 1.18, for $Q_0^6\text{-VTD}_0^6$ we have a superconvergence behavior of the derivative in the time mesh points only after postprocessing since only the postprocessed solution satisfies appropriate collocation conditions. On the other hand, for $Q_k^6\text{-VTD}_k^6$, $k = 5, 6$, we see that $\|(u - U)'\|_{\ell^\infty} = \|(u - \tilde{U})'\|_{\ell^\infty}$, which is clear by construction of the postprocessing. Nevertheless, the expected (super-)convergence orders are obtained since collocation conditions are fulfilled already by the discrete solution U in these cases. Furthermore, note that for $Q_6^6\text{-VTD}_6^6$ the postprocessing does not lead to an improvement of the error itself, whereas the L^2 -norm of the time derivative of the error is improved. This also is in agreement with our theory, cf. Corollary 1.28 and (1.20).

For further numerical results and a more detailed discussion we refer to [15, Section 7]. Also note that in [14, Section 6] similar investigations and findings were made for another test problem.

Table 1.3: Error of $Q_k^6\text{-VTD}_k^6$, $k = 0, 5, 6$, in different (semi-)norms and associated (experimental) convergence orders before and after postprocessing

	$Q_0^6\text{-VTD}_0^6$	eoc (theo)	$Q_5^6\text{-VTD}_5^6$	eoc (theo)	$Q_6^6\text{-VTD}_6^6$	eoc (theo)
$\ u - U\ _{L^2}$	2.607e-11 2.042e-13	7.00 (7)	2.828e-10 2.188e-12	7.01 (7)	2.092e-09 1.653e-11	6.98 (7)
$\ u - \tilde{U}\ _{L^2}$	9.898e-13 3.881e-15	7.99 (8)	5.008e-11 1.972e-13	7.99 (8)	1.184e-09 9.295e-12	6.99 (7)
$\ u - U\ _{\ell^\infty}$	1.385e-21 1.685e-25	13.00 (13)	4.552e-12 1.798e-14	7.98 (8)	7.584e-10 5.891e-12	7.01 (7)
$\ (u - U)'\ _{L^2}$	7.699e-09 1.207e-10	6.00 (6)	1.641e-08 2.564e-10	6.00 (6)	3.871e-08 5.954e-10	6.02 (6)
$\ (u - \tilde{U})'\ _{L^2}$	1.531e-10 1.201e-12	6.99 (7)	1.632e-09 1.281e-11	6.99 (7)	7.753e-09 6.120e-11	6.99 (7)
$\ (u - U)'\ _{\ell^\infty}$	3.573e-09 5.605e-11	5.99 (6)	6.361e-12 2.504e-14	7.99 (8)	8.736e-10 7.012e-12	6.96 (7)
$\ (u - \tilde{U})'\ _{\ell^\infty}$	1.522e-21 1.851e-25	13.01 (13)	6.361e-12 2.504e-14	7.99 (8)	8.735e-10 7.012e-12	6.96 (7)

1.4 Results for affine linear problems

The following section is restricted to the study of affine linear problems of the form

Find $u : \bar{I} \rightarrow \mathbb{R}^d$ such that

$$Mu'(t) = f(t) - Au(t), \quad u(t_0) = u_0 \in \mathbb{R}^d, \quad (1.21)$$

where $M, A \in \mathbb{R}^{d \times d}$ are time-independent matrices and M is regular. Thus, in the general setting we have $F(t, u) = f(t) - Au$.

1.4.1 A slight modification of the method

In this subsection, we want to introduce a slight modification of the variational time discretization method for the more structured affine linear problem (1.21). As we will see later, many schemes of practical relevance can be nicely described in the modified structure. Moreover, the modification is also quite interesting from a theoretical point of view.

Let $0 \leq k \leq r$. In order to solve (1.21) numerically, we define the (local) $\mathcal{J}_n\text{-VTD}_k^r(g)$ problem by

Given $U(t_{n-1}^-) \in \mathbb{R}^d$, find $U \in P_r(I_n, \mathbb{R}^d)$ such that

$$U(t_{n-1}^+) = U(t_{n-1}^-), \quad \text{if } k \geq 1, \quad (1.22a)$$

$$MU^{(i+1)}(t_n^-) = g^{(i)}(t_n^-) - AU^{(i)}(t_n^-), \quad \text{if } k \geq 2, i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1, \quad (1.22b)$$

$$MU^{(i+1)}(t_{n-1}^+) = g^{(i)}(t_{n-1}^+) - AU^{(i)}(t_{n-1}^+), \quad \text{if } k \geq 3, i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor - 1, \quad (1.22c)$$

and

$$\mathcal{J}_n[(MU', \varphi)] + \delta_{0,k}(M[U]_{n-1}, \varphi(t_{n-1}^+)) = \mathcal{J}_n[(g - AU, \varphi)] \quad \forall \varphi \in P_{r-k}(I_n, \mathbb{R}^d), \quad (1.22d)$$

where $U(t_0^-) = u_0$ and g is some approximation of f , details will be given later on. As before \mathcal{J}_n denotes an integrator, typically the integral over I_n or a quadrature formula.

First of all, we want to point out the main differences between the methods given by (1.2) and by (1.22). Also note that in the notation $\mathcal{J}_n\text{-VTD}_k^r(g)$ the approximation g itself is indicated instead of an approximation operator.

- Collocation conditions: In (1.2b) and (1.2c) the “real” right-hand side appears while in (1.22b) and (1.22c) usually an approximation of the right-hand side (g instead of f) is used.
- Variational condition: In case that the operator \mathcal{I}_n does not preserve polynomials up to degree r , we have $\mathcal{I}_n(f - AU) = \mathcal{I}_n f - A\mathcal{I}_n U \neq \mathcal{I}_n f - AU$ in general. Thus, the variational conditions (1.2d) and (1.22d) even differ for the affine linear problem (1.21).
- The approximation g of f usually does not provide global regularity properties. Therefore, even if f is sufficiently smooth, the solution U of (1.22) is $(\min\{\lfloor \frac{k-1}{2} \rfloor, k_g + 1\})$ -times continuously differentiable only. Here $k_g \geq -1$ denotes the largest integer such that $g \in C^{k_g}(\bar{I})$.

Remark 1.31

Applied to the affine linear problem (1.21) the methods $\mathcal{J}_n\text{-VTD}_k^r(\mathcal{I}_n)$ defined by (1.2) and $\mathcal{J}_n\text{-VTD}_k^r(\mathcal{I}_n f)$ defined by (1.22) are equivalent if \mathcal{I}_n preserves polynomials of degree less than or equal to r and additionally satisfies that $(v - \mathcal{I}_n v)^{(i)}(t_{n-1}^+) = 0$ for $0 \leq i \leq \lfloor \frac{k-1}{2} \rfloor - 1$ as well as $(v - \mathcal{I}_n v)^{(i)}(t_n^-) = 0$ for $0 \leq i \leq \lfloor \frac{k}{2} \rfloor - 1$.

Especially, we have for example that

$$\begin{aligned} \mathcal{J}_n\text{-VTD}_k^r &\hat{=} \mathcal{J}_n\text{-VTD}_k^r(f), \\ \mathcal{J}_n\text{-VTD}_k^r(\mathcal{I}_k^r) &\hat{=} \mathcal{J}_n\text{-VTD}_k^r(\mathcal{I}_k^r f) && \text{for all } 0 \leq k \leq r, \\ \mathcal{J}_n\text{-VTD}_k^r(\mathcal{I}_{k+2}^{r+1}) &\hat{=} \mathcal{J}_n\text{-VTD}_k^r(\mathcal{I}_{k+2}^{r+1} f) && \text{for all } 0 \leq k \leq r-1, \end{aligned}$$

where \mathcal{I}_k^r is the Hermite interpolation operator associated to the quadrature rule Q_k^r determined by (1.15). Note that for the last equivalence the case $r = k$ needs to be excluded since otherwise $\mathcal{I}_{k+2}^{r+1} f$ would not be well-defined. \clubsuit

1.4.2 Postprocessing for the modified method

Similar to the postprocessing of Subsection 1.3.2 we can also define a postprocessing for the modified method. Recall that Q_k^r denotes the quadrature rule associated to \mathbf{VTD}_k^r determined by (1.15).

Theorem 1.32 (Postprocessing $Q_k^r\text{-}\mathbf{VTD}_k^r(g) \rightsquigarrow Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(g)$)
 Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and suppose that $U \in Y_r$ solves $Q_k^r\text{-}\mathbf{VTD}_k^r(g)$. For $n = 1, \dots, N$ set

$$\check{U}|_{I_n} = U|_{I_n} + \check{a}_n \vartheta_n, \quad \vartheta_n \in P_{r+1}(I_n, \mathbb{R}),$$

where ϑ_n vanishes in the $(r+1)$ quadrature points of Q_k^r and satisfies $\vartheta_n^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-) = 1$ while the vector $\check{a}_n \in \mathbb{R}^d$ is defined by

$$\check{a}_n = M^{-1} \left(g(\lfloor \frac{k}{2} \rfloor)(t_n^-) - AU(\lfloor \frac{k}{2} \rfloor)(t_n^-) - MU(\lfloor \frac{k}{2} \rfloor + 1)(t_n^-) \right). \quad (1.23)$$

Moreover, let $\check{U}(t_0^-) = U(t_0^-)$. Then, $\check{U} \in Y_{r+1}$ solves $Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(g)$.

Proof. The argumentation is quite analog to that of [14, Theorem 3.1]. But for the sake of completeness and clarity we give it here. Especially, it can be seen that we do not need global regularity assumptions on g . We have to verify that \check{U} satisfies all conditions of $Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(g)$ where Q_k^r is the quadrature rule associated to \mathbf{VTD}_k^r , which is exact for polynomials up to degree $2r - k$.

First of all, we show an identity needed later. The special form of ϑ_n , the exactness of Q_k^r , and integration by parts yield

$$\begin{aligned} Q_k^r[\vartheta_n'] &= \int_{I_n} \vartheta_n'(t) \varphi(t) dt = - \int_{I_n} \vartheta_n(t) \varphi'(t) dt + (\vartheta_n \varphi)|_{t_{n-1}^+}^{t_n^-} \\ &= - \underbrace{Q_k^r[\vartheta_n \varphi']}_{=0} - \delta_{0,k}(\vartheta_n \varphi)(t_{n-1}^+) = -\delta_{0,k}(\vartheta_n \varphi)(t_{n-1}^+) \quad \forall \varphi \in P_{r-k}(I_n, \mathbb{R}). \end{aligned} \quad (1.24)$$

Precisely, we used that both $\vartheta_n' \varphi$ and $\vartheta_n \varphi'$ are polynomials of maximal degree $2r - k$ and that ϑ_n vanishes in all quadrature points, especially in t_n^- and for $k \geq 1$ also in t_{n-1}^+ .

For $k \geq 1$ we have $\vartheta_n(t_{n-1}^+) = \vartheta_n(t_n^-) = 0$. Therefore, the initial condition holds due to $\check{U}(t_{n-1}^+) = U(t_{n-1}^+) = U(t_{n-1}^-) = \check{U}(t_{n-1}^-)$. For $k = 0$ it is somewhat more complicated to prove $\check{U}(t_{n-1}^+) = \check{U}(t_{n-1}^-)$, for details see (iii) below. The remaining conditions can be verified as follows.

- (i) Conditions at t_n^- for $0 \leq i \leq \lfloor \frac{k+2}{2} \rfloor - 2 = \lfloor \frac{k}{2} \rfloor - 1$:

We obtain from the definitions of \check{U} and U

$$\begin{aligned} M\check{U}^{(i+1)}(t_n^-) &= MU^{(i+1)}(t_n^-) + M\underbrace{\check{a}_n \vartheta_n^{(i+1)}(t_n^-)}_{=0} = g^{(i)}(t_n^-) - AU^{(i)}(t_n^-) \\ &= g^{(i)}(t_n^-) - A\check{U}^{(i)}(t_n^-) \end{aligned}$$

since the derivatives of U and \check{U} in t_n^- coincide up to order $\lfloor \frac{k}{2} \rfloor$ due to the definition of ϑ_n .

- (ii) Condition at t_n^- for $i = \lfloor \frac{k+2}{2} \rfloor - 1 = \lfloor \frac{k}{2} \rfloor$:
 Just like above we get

$$\begin{aligned} M\check{U}^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-) &= MU^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-) + \underbrace{M\check{a}_n \vartheta_n^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-)}_{=1} \\ &= MU^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-) + g^{(\lfloor \frac{k}{2} \rfloor)}(t_n^-) - AU^{(\lfloor \frac{k}{2} \rfloor)}(t_n^-) - MU^{(\lfloor \frac{k}{2} \rfloor + 1)}(t_n^-) \\ &= g^{(\lfloor \frac{k}{2} \rfloor)}(t_n^-) - AU^{(\lfloor \frac{k}{2} \rfloor)}(t_n^-) = g^{(\lfloor \frac{k}{2} \rfloor)}(t_n^-) - A\check{U}^{(\lfloor \frac{k}{2} \rfloor)}(t_n^-), \end{aligned}$$

where additionally the definition of \check{a}_n was used.

- (iii) Variational condition:

We have to prove that $Q_k^r[(M\check{U}', \varphi)] = Q_k^r[(g - A\check{U}, \varphi)]$ for all $\varphi \in P_{(r+1)-(k+2)}(I_n, \mathbb{R}^d)$. Actually, we can even test with functions $\varphi \in P_{r-k}(I_n, \mathbb{R}^d)$.

We first study the case $k \geq 1$. By the definitions of \check{U} and U , the identity (1.24), and the fact that U and \check{U} coincide at all quadrature points we have

$$\begin{aligned} Q_k^r[(M\check{U}', \varphi)] &= Q_k^r[(MU', \varphi)] + Q_k^r[(M\check{a}_n \vartheta'_n, \varphi)] \\ &= Q_k^r[(g - AU, \varphi)] + \underbrace{Q_k^r[\vartheta'_n(M\check{a}_n, \varphi)]}_{=0, \text{ since } (M\check{a}_n, \varphi) \in P_{r-k}(I_n, \mathbb{R})} \\ &= Q_k^r[(g - A\check{U}, \varphi)] \quad \forall \varphi \in P_{r-k}(I_n, \mathbb{R}^d). \end{aligned}$$

Now let $k = 0$. The same arguments as for $k \geq 1$ here yield for all $\varphi \in P_r(I_n, \mathbb{R}^d)$

$$Q_0^r[(M\check{U}', \varphi)] = Q_0^r[(g - A\check{U}, \varphi)] - (M[U]_{n-1}, \varphi(t_{n-1}^+)) - \vartheta_n(t_{n-1}^+)(M\check{a}_n, \varphi(t_{n-1}^+)).$$

We study the last two terms. Using the definitions of the jump $[U]_{n-1}$ and of \check{U} , we find

$$[U]_{n-1} + \check{a}_n \vartheta_n(t_{n-1}^+) = \check{U}(t_{n-1}^+) - U(t_{n-1}^-) = [\check{U}]_{n-1}, \quad (1.25)$$

where we also exploited that $\vartheta_{n-1}(t_{n-1}^-) = 0$. Hence, we have

$$Q_0^r[(M\check{U}', \varphi)] + (M[\check{U}]_{n-1}, \varphi(t_{n-1}^+)) = Q_0^r[(g - A\check{U}, \varphi)] \quad \forall \varphi \in P_r(I_n, \mathbb{R}^d). \quad (1.26)$$

Choosing the special test functions $\varphi_j \in P_r(I_n, \mathbb{R}^d)$, $1 \leq j \leq d$, that vanish in the r inner quadrature points of Q_0^r and satisfy $\varphi_j(t_{n-1}^+) = e_j$ as well as having in mind (ii), we find $M[\check{U}]_{n-1} = 0$ component by component. Thereby, at once we have proven the initial condition and verified the needed variational condition since now also the jump term in (1.26) can be dropped.

- (iv) Conditions at t_{n-1}^+ for $0 \leq i \leq \lfloor \frac{k+2-1}{2} \rfloor - 2 = \lfloor \frac{k-1}{2} \rfloor - 1$:
 With an argumentation similar to that in (i) we gain

$$\begin{aligned} M\check{U}^{(i+1)}(t_{n-1}^+) &= MU^{(i+1)}(t_{n-1}^+) + M\check{a}_n \underbrace{\vartheta_n^{(i+1)}(t_{n-1}^+)}_{=0} = g^{(i)}(t_{n-1}^+) - AU^{(i)}(t_{n-1}^+) \\ &= g^{(i)}(t_{n-1}^+) - A\check{U}^{(i)}(t_{n-1}^+). \end{aligned}$$

- (v) Condition at t_{n-1}^+ for $i = \lfloor \frac{k+2-1}{2} \rfloor - 1 = \lfloor \frac{k-1}{2} \rfloor$ if $k \geq 1$:
 It remains to prove that

$$M\check{U}^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+) = g^{(\lfloor \frac{k-1}{2} \rfloor)}(t_{n-1}^+) - A\check{U}^{(\lfloor \frac{k-1}{2} \rfloor)}(t_{n-1}^+).$$

We use the variational condition for \check{U} , already shown in (iii), with specially chosen test functions $\varphi_j \in P_{r-k}(I_n, \mathbb{R}^d)$, $1 \leq j \leq d$, that vanish in all inner quadrature points of Q_k^r , i.e.,

$$\varphi_j(t_{n,i}) = 0, \quad i = 1, \dots, r-k, \quad \text{and satisfy} \quad \varphi_j(t_{n-1}^+) = e_j.$$

Since $k \geq 1$ here, we have that

$$Q_k^r[(M\check{U}', \varphi_j)] = Q_k^r[(g - A\check{U}, \varphi_j)], \quad j = 1, \dots, d.$$

The special choices of φ_j , the definition of the quadrature rule, and the already known identities from (i), (ii), and (iv) yield after a short calculation using Leibniz' rule for the i th derivative that

$$\begin{aligned} Q_k^r[(M\check{U}', \varphi_j)] &= Q_k^r[(g - A\check{U}, \varphi_j)], \quad j = 1, \dots, d, \\ \Leftrightarrow w_{\lfloor \frac{k-1}{2} \rfloor}^L M\check{U}^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+) \cdot \underbrace{\varphi_j(t_{n-1}^+)}_{=e_j} \\ &= w_{\lfloor \frac{k-1}{2} \rfloor}^L \left(g^{(\lfloor \frac{k-1}{2} \rfloor)}(t_{n-1}^+) - A\check{U}^{(\lfloor \frac{k-1}{2} \rfloor)}(t_{n-1}^+) \right) \cdot \underbrace{\varphi_j(t_{n-1}^+)}_{=e_j}, \quad j = 1, \dots, d, \\ \Leftrightarrow M\check{U}^{(\lfloor \frac{k-1}{2} \rfloor + 1)}(t_{n-1}^+) &= g^{(\lfloor \frac{k-1}{2} \rfloor)}(t_{n-1}^+) - A\check{U}^{(\lfloor \frac{k-1}{2} \rfloor)}(t_{n-1}^+). \end{aligned}$$

Note that here we also used that $w_{\lfloor \frac{k-1}{2} \rfloor}^L \neq 0$, cf. (1.16).

Collecting the above arguments, we see that \check{U} solves $Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(g)$. \square

As we already noticed above, one further issue is that in general g is not globally smooth and so also the discrete solution does not possess higher regularity properties. Therefore, we cannot expect that the alternative definition of the postprocessing, based on the correction of jumps, is equivalent to the postprocessing of Theorem 1.32 anymore.

Proposition 1.33

The correction vectors $\check{a}_n \in \mathbb{R}^d$ defined in (1.23) for the postprocessing presented in Theorem 1.32 could be alternatively calculated for $n > 1$ by

$$\begin{aligned} \check{a}_n &= \frac{-1}{\vartheta_n(t_{n-1}^+)} \left(U(t_{n-1}^+) - \check{U}(t_{n-1}^-) \right), & \text{if } k = 0, \\ \check{a}_n &= \frac{-1}{\vartheta_n(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^+)} \left(U(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^+) - \check{U}(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^-) \right. \\ &\quad \left. - M^{-1} \left[g(\lfloor \frac{k-1}{2} \rfloor) \right]_{n-1} + M^{-1} A \left[U(\lfloor \frac{k-1}{2} \rfloor) \right]_{n-1} \right), & \text{if } k \geq 1, \end{aligned}$$

and for $n = 1$ by

$$\begin{aligned} \check{a}_1 &= \frac{-1}{\vartheta_1(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0^+)} \left(U(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0^+) - u(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0) \right. \\ &\quad \left. - \sum_{j=0}^{\lfloor \frac{k-1}{2} \rfloor} (-M^{-1}A)^j M^{-1}(g - f)(\lfloor \frac{k-1}{2} \rfloor - j)(t_0^+) \right), \end{aligned}$$

where $u(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0)$ is defined via (1.4).

The proposition shows that in general the correction vector cannot be determined without solving a linear equation system with system matrix M if g is not globally smooth. However, if g is at least $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable and preserves f and its derivatives up to order $\lfloor \frac{k-1}{2} \rfloor$ at t_0^+ , then \check{a}_n , $n \geq 1$, can still be easily calculated as jump correction.

Proof. The basic ideas of the proof can be adopted from the proof of [14, Proposition 3.2]. However, many details have to be adapted since g cannot be assumed to be globally sufficiently smooth.

For $k = 0$, we get from (1.25) combined with $[\check{U}]_{n-1} = 0$, which was shown just below (1.26), that $\check{a}_n = \frac{-1}{\vartheta_n(t_{n-1}^+)} [U]_{n-1} = \frac{-1}{\vartheta_n(t_{n-1}^+)} (U(t_{n-1}^+) - \check{U}(t_{n-1}^-))$. Taking into account that $\check{U}(t_0^-) = U(t_0^-) = u(t_0) = u_0$, we are done in this case.

Otherwise, for $k \geq 1$, using the definition of the postprocessing and (v) of the proof of Theorem 1.32, we obtain that

$$\begin{aligned} MU(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^+) + M\check{a}_n \vartheta_n(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^+) &= M\check{U}(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^+) \\ &= g(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^+) - A\check{U}(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^+). \end{aligned} \tag{1.27}$$

Furthermore, we have $\vartheta_n^{(i)}(t_{n-1}^+) = 0$ for $i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor$ and therefore

$$\check{U}(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^+) = U(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^+).$$

If g is an approximation of f which is not globally smooth, then also the derivatives of U are not necessarily continuous up to a sufficiently high order anymore. So, we get for $n > 1$

$$\begin{aligned}
 & g(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^+) - AU(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^+) \\
 &= g(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^-) - AU(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^-) + \left[g(\lfloor \frac{k-1}{2} \rfloor) - AU(\lfloor \frac{k-1}{2} \rfloor) \right]_{n-1} \\
 &= g(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^-) - A\check{U}(\lfloor \frac{k-1}{2} \rfloor)(t_{n-1}^-) + \left[g(\lfloor \frac{k-1}{2} \rfloor) - AU(\lfloor \frac{k-1}{2} \rfloor) \right]_{n-1} \\
 &= M\check{U}(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^-) + \left[g(\lfloor \frac{k-1}{2} \rfloor) - AU(\lfloor \frac{k-1}{2} \rfloor) \right]_{n-1},
 \end{aligned}$$

where also $\vartheta_{n-1}^{(i)}(t_{n-1}^-) = 0$ for $i = 0, \dots, \lfloor \frac{k}{2} \rfloor$ and (i) or (ii), respectively, of the proof of Theorem 1.32 were used. Altogether, exploiting that M is regular, an easy manipulation of the identities yields

$$\begin{aligned}
 \check{a}_n = \frac{-1}{\vartheta_n(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^+)} & \left(U(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^+) - \check{U}(\lfloor \frac{k-1}{2} \rfloor + 1)(t_{n-1}^-) \right. \\
 & \left. - M^{-1} \left[g(\lfloor \frac{k-1}{2} \rfloor) \right]_{n-1} + M^{-1} A \left[U(\lfloor \frac{k-1}{2} \rfloor) \right]_{n-1} \right) \quad \text{for } n > 1.
 \end{aligned}$$

It remains to derive the formula for $n = 1$. Since U satisfies (1.22c) and recalling the definition (1.4) of $u^{(i)}(t_0)$, we have

$$(U - u)^{(i)}(t_0^+) = M^{-1}(g - f)^{(i-1)}(t_0^+) - M^{-1}A(U - u)^{(i-1)}(t_0^+) \quad \text{for } i = 1, \dots, \lfloor \frac{k-1}{2} \rfloor.$$

By recursion and exploiting that $U(t_0^+) = U(t_0^-) = u_0 = u(t_0)$, we obtain

$$(U - u)^{(i)}(t_0^+) = \sum_{j=1}^i (-M^{-1}A)^{j-1} M^{-1}(g - f)^{(i-j)}(t_0^+) \quad \text{for } i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor.$$

Therefore, the right-hand side of (1.27) can be rewritten for $n = 1$ as follows

$$\begin{aligned}
 & g(\lfloor \frac{k-1}{2} \rfloor)(t_0^+) - A\check{U}(\lfloor \frac{k-1}{2} \rfloor)(t_0^+) = g(\lfloor \frac{k-1}{2} \rfloor)(t_0^+) - AU(\lfloor \frac{k-1}{2} \rfloor)(t_0^+) \\
 &= Mu(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0) + (g - f)(\lfloor \frac{k-1}{2} \rfloor)(t_0^+) - A(U - u)(\lfloor \frac{k-1}{2} \rfloor)(t_0^+) \\
 &= Mu(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0) + M \sum_{j=0}^{\lfloor \frac{k-1}{2} \rfloor} (-M^{-1}A)^j M^{-1}(g - f)(\lfloor \frac{k-1}{2} \rfloor - j)(t_0^+).
 \end{aligned}$$

This results in

$$\begin{aligned}
 \check{a}_1 = \frac{-1}{\vartheta_1(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0^+)} & \left(U(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0^+) - u(\lfloor \frac{k-1}{2} \rfloor + 1)(t_0) \right. \\
 & \left. - \sum_{j=0}^{\lfloor \frac{k-1}{2} \rfloor} (-M^{-1}A)^j M^{-1}(g - f)(\lfloor \frac{k-1}{2} \rfloor - j)(t_0^+) \right),
 \end{aligned}$$

which completes the proof. \square

Remark 1.34

Theorem 1.32 also enables us to apply the postprocessing properly to the exactly integrated variational time discretization method $\mathbf{VTD}_k^r(f)$. However, some trick is necessary.

For $f \in C^{\lfloor \frac{k-1}{2} \rfloor}(\bar{I}_n, \mathbb{R}^d)$ define $\Pi_k^r f \in P_r(I_n, \mathbb{R}^d)$ by

$$(\Pi_k^r f)^{(i)}(t_{n-1}^+) = f^{(i)}(t_{n-1}^+), \quad \text{if } k \geq 1, i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor, \quad (1.28a)$$

$$(\Pi_k^r f)^{(i)}(t_n^-) = f^{(i)}(t_n^-), \quad \text{if } k \geq 2, i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1, \quad (1.28b)$$

$$\int_{I_n} (\Pi_k^r f(t), \varphi(t)) dt = \int_{I_n} (f(t), \varphi(t)) dt \quad \forall \varphi \in P_{r-k}(I_n, \mathbb{R}^d). \quad (1.28c)$$

Then, it obviously holds

$$\mathbf{VTD}_k^r(f) \quad \hat{=} \quad \mathbf{VTD}_k^r(\Pi_k^r f) \quad \hat{=} \quad Q_k^r\text{-}\mathbf{VTD}_k^r(\Pi_k^r f),$$

where for the last equivalence we used that all terms in the variational condition are of maximal polynomial degree $2r - k$ and, thus, are integrated exactly by Q_k^r .

The application of the postprocessing of Theorem 1.32 therefore yields

$$\mathbf{VTD}_k^r(f) \quad \hat{=} \quad Q_k^r\text{-}\mathbf{VTD}_k^r(\Pi_k^r f) \quad \rightsquigarrow \quad Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(\Pi_k^r f) \quad \hat{=} \quad \mathbf{VTD}_{k+2}^{r+1}(\Pi_k^r f)$$

for all $0 \leq k \leq r$. For the above argument it is not needed that (1.28a) holds also for $i = \lfloor \frac{k-1}{2} \rfloor$. However, this additional feature of Π_k^r guarantees that the postprocessed solution always is $\lfloor \frac{k}{2} \rfloor$ -times continuously differentiable if f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable. \clubsuit

1.4.3 Interpolation cascade

Recall that \mathcal{I}_k^r is the Hermite interpolation operator associated to the quadrature rule Q_k^r determined by (1.15). \mathcal{I}_k^r is a projection operator onto polynomials of maximal degree r . The quadrature rule Q_k^r is exact for polynomials up to degree $2r - k$.

The presented postprocessing techniques essentially use that the quadrature formula Q_k^r is well-suited to the \mathbf{VTD}_k^r method. After one postprocessing step, however, we stay with Q_k^r , but the basic method has changed to \mathbf{VTD}_{k+2}^{r+1} . This does not match anymore. Therefore, we ask whether the quadrature rule can be changed and readjusted.

In a first step, we will consider $Q_k^r\text{-}\mathbf{VTD}_k^r(\mathcal{I}_{k+2}^{r+1}f)$ for $0 \leq k \leq r-1$. Here the case $r = k$ is excluded in order to ensure that $\mathcal{I}_{k+2}^{r+1}f$ is well-defined. We observe the following interesting property.

Theorem 1.35 (Cf. [14, Theorem 5.1])

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r-1$. Suppose that $U \in Y_r$ solves $Q_k^r\text{-}\mathbf{VTD}_k^r(\mathcal{I}_{k+2}^{r+1}f)$. Determine $\tilde{U} \in Y_{r+1}$ by the postprocessing of Theorem 1.32. Then, \tilde{U} solves $Q_{k+2}^{r+1}\text{-}\mathbf{VTD}_{k+2}^{r+1}(f)$.

Proof. Let U solve $Q_k^r\text{-}\mathbf{VTD}_k^r(\mathcal{I}_{k+2}^{r+1}f)$. Then, by Theorem 1.32 the postprocessed solution \tilde{U} solves $Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(\mathcal{I}_{k+2}^{r+1}f)$. It remains to prove that

$$Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(\mathcal{I}_{k+2}^{r+1}f) \quad \hat{=} \quad Q_{k+2}^{r+1}\text{-}\mathbf{VTD}_{k+2}^{r+1}(\mathcal{I}_{k+2}^{r+1}f) \quad \hat{=} \quad Q_{k+2}^{r+1}\text{-}\mathbf{VTD}_{k+2}^{r+1}(f).$$

Since $\mathcal{I}_{k+2}^{r+1}f \in P_{r+1}(I_n, \mathbb{R}^d)$, all terms of the variational condition are integrated exactly by quadrature formulas that are exact for polynomials up to degree $2r - k$, so especially by Q_k^r and Q_{k+2}^{r+1} . Thus, the first equivalence is shown.

Moreover, on the one hand, \mathcal{I}_{k+2}^{r+1} preserves all derivatives of f that occur in the collocation conditions and so can be dropped there. On the other hand, \mathcal{I}_{k+2}^{r+1} is not seen by Q_{k+2}^{r+1} since both are defined by the same points. Hence, \mathcal{I}_{k+2}^{r+1} can also be dropped in the variational condition, which verifies the second equivalence. \square

Remark 1.36 (Cf. [14, Remark 5.2])

Within the above argument we proved that the method $Q_{k+2}^{r+1}\text{-VTD}_{k+2}^{r+1}(f)$ and the method $Q_k^r\text{-VTD}_{k+2}^{r+1}(\mathcal{I}_{k+2}^{r+1}f)$ are equivalent for $0 \leq k \leq r - 1$.

Similarly, one can show that $Q_k^r\text{-VTD}_{k+2}^{r+1}(f)$ is equivalent to $Q_{k+2}^{r+1}\text{-VTD}_{k+2}^{r+1}(\mathcal{I}_k^r f)$ for $0 \leq k \leq r - 1$. Note that also \mathcal{I}_k^r preserves all derivatives that appear in the point conditions at both ends of the interval. \clubsuit

Having a closer look at the result of Theorem 1.35, we see that the postprocessed solution of the modified discrete problem also solves a numerically integrated variational time discretization method but with the “right” associated quadrature rule. This enables one further postprocessing step.

For $1 \leq j \leq r - k$, using an interpolation cascade, we even could enable up to $j + 1$ postprocessing steps. More concretely, we have (where \rightsquigarrow denotes the postprocessing steps as given by Theorem 1.32)

$$\begin{aligned} Q_k^r\text{-VTD}_k^r(\mathcal{I}_{k+2}^{r+1} \circ \mathcal{I}_{k+4}^{r+2} \circ \dots \circ \mathcal{I}_{k+2j}^{r+j} f) &\rightsquigarrow Q_{k+2}^{r+1}\text{-VTD}_{k+2}^{r+1}(\mathcal{I}_{k+4}^{r+2} \circ \dots \circ \mathcal{I}_{k+2j}^{r+j} f) \\ &\rightsquigarrow \dots \\ &\rightsquigarrow Q_{k+2(j-1)}^{r+j-1}\text{-VTD}_{k+2(j-1)}^{r+j-1}(\mathcal{I}_{k+2j}^{r+j} f) \rightsquigarrow Q_{k+2j}^{r+j}\text{-VTD}_{k+2j}^{r+j}(f) \\ &\rightsquigarrow Q_{k+2j}^{r+j}\text{-VTD}_{k+2(j+1)}^{r+j+1}(f). \end{aligned}$$

Note that f itself can be used in each postprocessing step to calculate the correction vector $\check{a}_n \in \mathbb{R}^d$ (cf. Theorem 1.32) since in each step the occurring derivative of f at t_n^- is preserved by the respective interpolation cascade.

As abbreviation, we write $\mathcal{C}_k^r := \mathcal{I}_k^r \circ \mathcal{I}_{k+2}^{r+1} \circ \dots \circ \mathcal{I}_{2r-k}^{2r-k}$ for the longest interpolation cascade (for which $j = r - k$) in the following.

Remark 1.37

If g is locally on I_n an approximation of f of maximal polynomial degree $r + 1$, then similar to the proof of Theorem 1.35 we have that $Q_k^r\text{-VTD}_{k+2}^{r+1}(g)$ is equivalent to $Q_{k+2}^{r+1}\text{-VTD}_{k+2}^{r+1}(g)$. Hence, if $g|_{I_n} \in P_{r+1}(I_n, \mathbb{R}^d)$, we are always able to perform up to $r - k + 1$ postprocessing steps. We find (where \rightsquigarrow denotes the postprocessing steps as given by Theorem 1.32)

$$\begin{aligned} Q_k^r\text{-VTD}_k^r(g) &\rightsquigarrow Q_{k+2}^{r+1}\text{-VTD}_{k+2}^{r+1}(g) \rightsquigarrow \dots \rightsquigarrow Q_{2r-k}^{2r-k}\text{-VTD}_{2r-k}^{2r-k}(g) \rightsquigarrow Q_{2r-k}^{2r-k}\text{-VTD}_{2r-k+2}^{2r-k+1}(g). \end{aligned}$$

However, while $g|_{I_n} \in P_{r+1}(I_n, \mathbb{R}^d)$ is needed to enable the change of the quadrature rule after the first postprocessing step, this entails that after several postprocessing steps the

approximation of f by g is of lower order than the ansatz order of the variational time discretization method. Therefore, we cannot expect an improvement of the convergence order after two or more postprocessing steps in general. The interpolation cascade does not have this issue.

Moreover, from Proposition 1.33 it is obvious that in general (for arbitrary g) the postprocessing by (modified) residuals and the usual postprocessing by jumps do not provide the same correction anymore.

A more detailed analysis shows that applying two postprocessing steps based on residuals on the solution of $Q_k^r\text{-VTD}_k^r(f)$ yields the solution of $Q_{k+4}^{r+2}\text{-VTD}_{k+4}^{r+2}(\mathcal{I}_{k,\otimes}^{r+1}f)$ where $\mathcal{I}_{k,\otimes}^{r+1}f$ interpolates f in the quadrature points of Q_k^r and additionally preserves its $(\lfloor \frac{k}{2} \rfloor + 1)$ th derivative in t_n^- . Similarly (at least) for dG-like methods (characterized by even k) it can be shown that applying two postprocessing steps based on jumps on the solution of $Q_k^r\text{-VTD}_k^r(f)$ gives the solution of $Q_{k+4}^{r+2}\text{-VTD}_{k+4}^{r+2}(\mathcal{I}_{k,*}^{r+1}f)$ where $\mathcal{I}_{k,*}^{r+1}f$ interpolates f in the quadrature points of Q_k^r and additionally preserves its $(\lfloor \frac{k-1}{2} \rfloor + 1)$ th derivative in t_{n-1}^+ . ♣

Remark 1.38

Since the postprocessing does not change the function value in t_n^- , $1 \leq n \leq N$, we have for $1 \leq j \leq r - k$ that the solutions of $Q_k^r\text{-VTD}_k^r(\mathcal{I}_{k+2}^{r+1} \circ \mathcal{I}_{k+4}^{r+2} \circ \dots \circ \mathcal{I}_{k+2j}^{r+j}f)$ and of $Q_{k+2j}^{r+j+1}\text{-VTD}_{k+2j}^{r+j+1}(f)$ coincide in the end points of the intervals. Hence, the pointwise error estimates for the latter method immediately imply superconvergence in the time mesh points for $Q_k^r\text{-VTD}_k^r(\mathcal{I}_{k+2}^{r+1} \circ \mathcal{I}_{k+4}^{r+2} \circ \dots \circ \mathcal{I}_{k+2j}^{r+j}f)$. ♣

Remark 1.39 (Cf. [14, Remark 5.3])

For Dahlquist's stability equation

$$u'(t) = \lambda u(t), \quad u(t_0) = u_0 \in \mathbb{R}, \quad (1.29)$$

i.e., $d = 1$, $M = 1$, $A = -\lambda \in \mathbb{C}$, and $f = 0$ in (1.21), we easily see that

$$\text{VTD}_{k-2j}^{r-j}(f) \hat{=} Q_{k-2j}^{r-j}\text{-VTD}_{k-2j}^{r-j}(f) \hat{=} Q_{k-2j}^{r-j}\text{-VTD}_{k-2j}^{r-j}(\mathcal{I}_{k-2j+2}^{r-j+1} \circ \mathcal{I}_{k-2j+4}^{r-j+2} \circ \dots \circ \mathcal{I}_k^r f)$$

for all $j = 0, \dots, \lfloor \frac{k}{2} \rfloor$. Thus, j postprocessing steps can be applied for this equation. Since the postprocessing does not change the function value in the end points of the intervals, the stability function does not change either. Therefore, VTD_k^r as well as $Q_k^r\text{-VTD}_k^r$ provide the same stability function as VTD_{k-2j}^{r-j} . With the special choice $j = \lfloor \frac{k}{2} \rfloor$, we immediately find that VTD_k^r shares its stability properties with

$$\text{VTD}_{k-2\lfloor \frac{k}{2} \rfloor}^{r-\lfloor \frac{k}{2} \rfloor} \hat{=} \begin{cases} \text{VTD}_0^{r-\lfloor \frac{k}{2} \rfloor} \hat{=} \text{dG}(r - \lfloor \frac{k}{2} \rfloor), & \text{if } k \text{ is even,} \\ \text{VTD}_1^{r-\lfloor \frac{k}{2} \rfloor} \hat{=} \text{cGP}(r - \lfloor \frac{k}{2} \rfloor), & \text{if } k \text{ is odd,} \end{cases}$$

also cf. Remark 1.1 and [14, 17]. Thus, the VTD_k^r methods are A -stable for k odd while they are even strongly A -stable if k is even. ♣

1.4.4 Derivatives of solutions

In this subsection, the derivatives of solutions to $\text{VTD}_k^r(g)$ methods are studied. We see that the conditions of (1.22) are somewhat nested and that the derivatives also are solutions of certain variational time discretization schemes.

Theorem 1.40

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and suppose that $U \in Y_r$ solves $\mathbf{VTD}_k^r(g)$. Then, it holds

$$\int_{I_n} (M(U^{(j)})' + AU^{(j)}, \varphi) dt = \int_{I_n} (g^{(j)}, \varphi) dt$$

for all $0 \leq j \leq \lfloor \frac{k-1}{2} \rfloor$ and $\varphi \in P_{r-k+j}(I_n, \mathbb{R}^d)$. Further, for $j = \lfloor \frac{k}{2} \rfloor$ with k even we find that

$$\int_{I_n} (M(U^{(j)})' + AU^{(j)}, \varphi) dt + \delta_{0,k-2j}(M[U^{(j)}]_{n-1}, \varphi(t_{n-1}^+)) = \int_{I_n} (g^{(j)}, \varphi) dt$$

for all $\varphi \in P_{r-k+j}(I_n, \mathbb{R}^d)$, if g is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable and, for $j \geq 1$, $U^{(j)}(t_0^-)$ is determined by $MU^{(j)}(t_0^-) + AU^{(j-1)}(t_0^+) = g^{(j-1)}(t_0^+)$.

Proof. The proof of [14, Theorem 5.5] can be directly adopted replacing $\mathcal{I}f$ by g . Since, however, g does not provide global smoothness, we shortly recapitulate the key arguments of the proof in order to reveal where smoothness is actually necessary.

Let us start with the case $0 \leq j \leq \lfloor \frac{k-1}{2} \rfloor$. Integration by parts several times and exploiting (1.22), we find

$$\begin{aligned} & \int_{I_n} (M(U^{(j)})' + AU^{(j)}, \varphi) dt \\ &= (-1)^j \int_{I_n} (MU' + AU, \varphi^{(j)}) dt + \sum_{l=0}^{j-1} (-1)^l [(\partial_t^{j-1-l}(MU' + AU), \varphi^{(l)})]_{t_{n-1}^+}^{t_n^-} \\ &= (-1)^j \int_{I_n} (g, \varphi^{(j)}) dt + \sum_{l=0}^{j-1} (-1)^l [(g^{(j-1-l)}, \varphi^{(l)})]_{t_{n-1}^+}^{t_n^-} = \int_{I_n} (g^{(j)}, \varphi) dt \end{aligned}$$

for all $\varphi \in P_{r-k+j}(I_n, \mathbb{R}^d)$. Here, we do not need any global smoothness of g .

Now, let $j = \lfloor \frac{k}{2} \rfloor$ with $k \geq 2$ even. With the same arguments as just above we obtain

$$\begin{aligned} & \int_{I_n} (M(U^{(j)})' + AU^{(j)}, \varphi) dt \\ &= \int_{I_n} (g^{(j)}, \varphi) dt - (MU^{(j)}(t_{n-1}^+) + AU^{(j-1)}(t_{n-1}^+), \varphi(t_{n-1}^+)) + (g^{(j-1)}(t_{n-1}^+), \varphi(t_{n-1}^+)) \end{aligned}$$

for all $\varphi \in P_{r-k+j}(I_n, \mathbb{R}^d)$. Note that there are these extra terms on the right-hand side since an appropriate collocation condition is missing here. However, by the additional assumption that g is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable, which thereby also holds for U , the desired jump term can be derived for $n > 1$ from (1.22b) and for $n = 1$ using the special definition of $U^{(j)}(t_0^-)$, respectively. \square

Remark 1.41

If g preserves f and its derivatives up to order $\lfloor \frac{k}{2} \rfloor - 1$ at t_0^+ , then $U^{(j)}(t_0^-) = u^{(j)}(t_0)$ for $0 \leq j \leq \lfloor \frac{k}{2} \rfloor$. \clubsuit

Using an appropriate initial condition, derivatives of \mathbf{VTD} solutions are themselves solutions of \mathbf{VTD} methods.

Corollary 1.42

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Furthermore, suppose that f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable and that $U \in Y_r$ solves $\mathbf{VTD}_k^r(\mathcal{I}f)$ with $\mathcal{I} \in \{\text{Id}, \Pi_k^r, \mathcal{I}_k^r, \mathcal{C}_k^r\}$. Then, $U^{(j)} \in Y_{r-j}$, $0 \leq j \leq \lfloor \frac{k}{2} \rfloor$, solves $\mathbf{VTD}_{k-2j}^{r-j}((\mathcal{I}f)^{(j)})$ if $u^{(j)}(t_0)$ is used as initial condition.

Proof. Because of Theorem 1.40, it only remains to prove the needed conditions at t_{n-1}^+ and t_n^- . Since we have by construction that U is $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable, the desired identities follow from the fact that $U^{(j)}$ is continuous for $0 \leq j \leq \lfloor \frac{k-1}{2} \rfloor$ together with (1.22a), (1.22b), and (1.22c) with $g = \mathcal{I}f$. \square

Remark 1.43

For a convenient interpretation of Corollary 1.42 note that for the affine linear problems of the form (1.21) we have that $\mathbf{VTD}_k^r \hat{=} \mathbf{VTD}_k^r(f) \hat{=} \mathbf{VTD}_k^r(\Pi_k^r f)$ and

$$Q_k^r\text{-}\mathbf{VTD}_k^r \hat{=} Q_k^r\text{-}\mathbf{VTD}_k^r(f) \hat{=} Q_k^r\text{-}\mathbf{VTD}_k^r(\mathcal{I}_k^r f) \hat{=} \mathbf{VTD}_k^r(\mathcal{I}_k^r f)$$

for all $0 \leq k \leq r$. Moreover, recall that $\mathcal{C}_k^r = \mathcal{I}_k^r \circ \mathcal{I}_{k+2}^{r+1} \circ \dots \circ \mathcal{I}_{2r-k}^{2r-k}$.

Thus, $\mathcal{I} = \text{Id}$ and $\mathcal{I} = \Pi_k^r$ model the case of exact integration, $\mathcal{I} = \mathcal{I}_k^r$ models the case of numerical integration by the Q_k^r quadrature formula, and $\mathcal{I} = \mathcal{C}_k^r$ models the case where the interpolation cascade \mathcal{C}_k^r is used. \clubsuit

1.4.5 Numerical results

Our theoretical investigations suggest that the application of cascadic interpolation to the right-hand side f allows multiple postprocessing steps. This should be illustrated by some computational results. Besides we want to have a look on the differences between postprocessing based on jumps and postprocessing based on residuals when more than one postprocessing step is applied. Since appropriate numerical studies were made in [14, Section 6], also see [15, Section 7], we only give a short summary of the obtained results here.

Example

We consider the affine linear initial value problem

$$\begin{pmatrix} 10 & -20 \\ -10 & 20 \end{pmatrix} \begin{pmatrix} u_1'(t) \\ u_2'(t) \end{pmatrix} = \begin{pmatrix} -10e^{-10t} \\ 0 \end{pmatrix} - \begin{pmatrix} 1 & -101 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} u_1(t) \\ u_2(t) \end{pmatrix}, \quad t \in (0, 40), \quad (1.30a)$$

with initial condition

$$u_1(0) = 2, \quad u_2(0) = 1. \quad (1.30b)$$

Then, the solution components are given by

$$u_1(t) = e^{-t/10} + (1+t)e^{-10t}, \quad u_2(t) = (1+t)e^{-10t}.$$

Note that test problem (1.30) is a slight modification of [28, Example 7.3]. In particular, a non-trivial mass matrix was introduced. Furthermore, a non-vanishing right-hand side function $f = (-10e^{-10t}, 0)^T$ was added since otherwise the effects of the interpolation cascade cannot be studied.

Again all calculations were carried out with the software Julia [18], where the floating point data type `BigFloat` with 512 bits was used. For clarity, the function obtained after an application of s postprocessing steps starting from the discrete solution U is denoted by $\text{PP}_s U$ in the following.

In Table 1.4 the experimental orders of convergence of $\|(u - \text{PP}_s U)'\|_{L^2}$ for $Q_k^9\text{-VTD}_k^9$, $k = 3, \dots, 7$, after $s = 0, \dots, r + 1 - k = 10 - k$ postprocessing steps are presented. Hereby, the experimental orders of convergence were calculated from the errors obtained for 1024 and 2048 uniform time steps. Results are given for three different settings. In setting (i) cascadic interpolation is applied to the right-hand side f , i.e., we use $g = \mathcal{C}_k^9 f$. In this case both types of postprocessing are equivalent and, thus, lead to identical results. This changes substantially if no cascadic interpolation is used, i.e., for $g = f$. Therefore, for this case, the postprocessing based on jumps and the postprocessing based on residuals are considered separately in setting (ii) and (iii), respectively.

Table 1.4: Experimental orders of convergence for $\|(u - \text{PP}_s U)'\|_{L^2}$ using $Q_k^9\text{-VTD}_k^9$ with $k = 3, \dots, 7$, and s postprocessing steps

k	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$	$s = 6$	$s = 7$
(i) cascadic interpolation of f								
3	9.000	9.949	10.970	11.977	12.980	13.981	14.981	15.979
4	9.000	9.930	10.954	11.963	12.967	13.969	14.967	
5	9.000	9.949	10.970	11.976	12.979	13.977		
6	9.001	9.927	10.952	11.961	12.962			
7	9.001	9.950	10.969	11.972				
(ii) postprocessing based on jumps								
3	9.000	9.948	9.994	8.994	7.994	6.994	5.994	4.994
4	9.000	9.932	10.966	11.059	10.971	10.934	10.905	
5	9.000	9.949	10.981	9.981	8.980	7.980		
6	9.000	9.929	10.957	10.954	10.991			
7	9.000	9.949	9.995	8.994				
(iii) postprocessing based on residuals								
3	9.000	9.948	10.962	10.966	10.977	10.986	10.982	10.988
4	9.000	9.932	10.952	10.958	10.968	10.978	10.973	
5	9.000	9.949	10.956	10.958	10.963	10.987		
6	9.000	9.929	10.940	10.945	10.945			
7	9.000	9.949	10.946	10.947				

The numerical data are in good agreement with our theoretical expectations. If cascadic interpolation of the right-hand side f is used, it can be clearly seen that the convergence order increases by one with each additional postprocessing step. Moreover, if no interpolation cascade is applied, the computational results nicely verify the improvements by the first postprocessing step independent of the type of postprocessing. However, the situation

changes substantially when at least two postprocessing steps are used. While, independent of k , two or more postprocessing steps based on residuals always increase the convergence order by two compared to the results without postprocessing, such improvements can only be observed for dG-like methods (characterized by even k) if the postprocessing is based on jumps. In contrast, for cGP-like methods (corresponding to odd k) the second postprocessing step based on jumps leads to an additional improvement only for $k \equiv 1 \pmod{4}$, whereas for $k \equiv 3 \pmod{4}$ the convergence order is not increased further. Besides, for all cGP-like methods the convergence orders start to decrease if three or more postprocessing steps based on jumps are applied. Note that in calculations for $Q_k^{10}\text{-}\mathbf{VTD}_k^{10}$, $k = 0, \dots, 10$, the roles of $k \equiv 1 \pmod{4}$ and $k \equiv 3 \pmod{4}$ were switched. For further discussions and results we refer to [14, Section 6] and [15, Section 7].

2 Error Analysis for Stiff Systems

The error analysis for numerical methods applied to stiff ordinary differential equations is strongly connected to the concept of B -convergence introduced in [29]. The main object of this concept, developed for general nonlinear differential equations satisfying a certain one-sided Lipschitz condition, is the derivation of error bounds that only depend on the one-sided Lipschitz constant. A dependence of the error constant on the two-sided Lipschitz constant, which might be disproportionately large due to the stiffness of the problem, is explicitly avoided.

It is well-known that $Q_1^r\text{-VTD}_1^r$ and $Q_0^r\text{-VTD}_0^r$ can be interpreted as $(r + 1)$ -stage Lobatto IIIA and Radau IIA methods, respectively, as it was exemplarily shown in [46, p. 8, p. 13]. But Lobatto IIIA methods are not B -convergent for the general class of nonlinear problems, see [38, p. 231]. Therefore, it cannot be possible to prove a general B -convergence result for the variational time discretization methods (1.2). However, for certain classes of semilinear initial value problems, Lobatto IIIA methods and others nevertheless can be B -convergent, as it was shown for example in [20, see Theorem 3.4 and Lemma 2.3]. Thus, we have the reasonable hope that at least for affine linear problems with time-independent constants of the form (1.21) the VTD_k^r methods provide an error estimate independent of the stiffness.

In order to study the variational time discretization methods, we shall write them in a way similar to Runge–Kutta methods. For this purpose, first of all a Runge–Kutta-like framework is presented that enables an easy fitting of the VTD methods. In the end, the reformulation allows us to adapt and generalize many ideas and results that are usually used in the (stiff) error analysis for Runge–Kutta methods. In this context we particularly refer to [19], where the B -convergence of Runge–Kutta methods was studied for a semilinear problem which has the stiffness contained in a constant coefficient linear part. This situation is slightly more general than our setting but still so simple that most technical details can be avoided.

2.1 Runge–Kutta-like discretization framework

The aim of this section is to establish a Runge–Kutta-like framework that easily allows the representation of VTD methods but still enables typical convergence analysis approaches for Runge–Kutta methods. Here the well-known connection between collocation and Runge–Kutta methods can be used as motivation and inspiration. With these insights the Runge–Kutta formulation is extended in such a way that appropriate characteristics of the VTD methods can nicely be covered.

2.1.1 Connection between collocation and Runge–Kutta methods and its extension

It is well-known that collocation methods with s points can be easily written as s -stage Runge–Kutta methods, see e.g. [37, Theorem II.7.7, p. 212]. The Runge–Kutta coefficients then are given as certain integrals over the Lagrangian basis functions with respect to the collocation points. For the proof it is used that, given $U(t_{n-1}^-) \in \mathbb{R}^d$, the collocation polynomial $U \in P_s(I_n, \mathbb{R}^d)$ is determined by

$$U(t_{n-1}^+) = U(t_{n-1}^-), \quad MU'(\cdot) = \mathcal{P}_n F(\cdot, U(\cdot)) \quad (2.1)$$

where \mathcal{P}_n is the polynomial interpolation operator into $P_{s-1}(I_n)$ with respect to the s collocation points, which is applied component-wise here.

For a first extension, we now assume that \mathcal{P}_n is the local (transformed) version of a more general projection operator $\hat{\mathcal{P}}$ on $[-1, 1]$. More concretely, for sufficiently smooth functions on $[-1, 1]$ let $\hat{\mathcal{P}}$ be a projection operator onto $P_{s-1}([-1, 1])$ uniquely defined by the s linear functionals $\hat{N}_i^{\ell[i]}$, $i = 1, \dots, s$. The upper script $\ell[i] \geq 0$ here denotes the smallest derivative included in the definition of $\hat{N}_i^{\ell[i]}$, so with suitable functionals \hat{N}_i^0 we could simply write $\hat{N}_i^{\ell[i]}(\hat{v}) = \hat{N}_i^0(\hat{v}^{(\ell[i])})$. Typical examples are, of course, functionals based on point evaluations of functions or derivatives as $\hat{N}_i^{\ell[i]}(\hat{v}) = \hat{v}^{(\ell[i])}(\tilde{t})$ for some $\tilde{t} \in [-1, 1]$ but also functionals based on integrals as $\hat{N}_i^{\ell[i]}(\hat{v}) = \int_{-1}^1 \hat{v}^{(\ell[i])}(\hat{s}) d\hat{s}$. The associated basis functions $\hat{B}_i \in P_{s-1}([-1, 1])$, $i = 1, \dots, s$, should be chosen such that $\hat{N}_j^{\ell[j]}(\hat{B}_i) = \delta_{i,j}$. Thus, for sufficiently smooth functions \hat{v} on $[-1, 1]$ the projection $\hat{\mathcal{P}}\hat{v}$ can be written as

$$\hat{v} \mapsto (\hat{\mathcal{P}}\hat{v})(\cdot) = \sum_{i=1}^s \hat{N}_i^{\ell[i]}(\hat{v}) \hat{B}_i(\cdot).$$

The local versions \mathcal{P}_n on I_n , $n = 1, \dots, N$, are then given by

$$v \mapsto \mathcal{P}_n v = (\hat{\mathcal{P}}(v \circ T_n)) \circ T_n^{-1} = \sum_{i=1}^s \hat{N}_i^{\ell[i]}(v \circ T_n) (\hat{B}_i \circ T_n^{-1})$$

with T_n from (1.7). For brevity, we set $N_{i,n}^{\ell[i]}(v) = \hat{N}_i^{\ell[i]}(v \circ T_n)$.

We start to review the proof of the connection between collocation and Runge–Kutta methods step by step and extend the ideas if necessary. For U satisfying (2.1) the fundamental theorem of calculus implies for all $t \in I_n$ that

$$MU(t) = MU(t_{n-1}^+) + \int_{t_{n-1}}^t MU'(\tilde{s}) d\tilde{s} = MU(t_{n-1}^-) + \int_{t_{n-1}}^t \mathcal{P}_n F(\tilde{s}, U(\tilde{s})) d\tilde{s}.$$

Additionally using the (extended) definition of \mathcal{P}_n , we find

$$\begin{aligned} MU(t) &= MU(t_{n-1}^-) + \sum_{j=1}^s N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) \int_{t_{n-1}}^t (\hat{B}_j \circ T_n^{-1})(\tilde{s}) d\tilde{s} \\ &= MU(t_{n-1}^-) + \sum_{j=1}^s N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) \frac{\tau_n}{2} \int_{-1}^{T_n^{-1}(t)} \hat{B}_j(\hat{s}) d\hat{s}. \end{aligned}$$

Therefore, it follows for $i = 1, \dots, s$ that

$$\begin{aligned} MN_{i,n}^{\ell[i]}(U(\cdot)) &= MN_{i,n}^{\ell[i]}(U(t_{n-1}^-)) + \frac{\tau_n}{2} \sum_{j=1}^s N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) N_{i,n}^{\ell[i]} \left(\int_{-1}^{T_n^{-1}(\cdot)} \widehat{B}_j(\hat{s}) d\hat{s} \right) \\ &= MN_{i,n}^{\ell[i]}(U(t_{n-1}^-)) + \frac{\tau_n}{2} \sum_{j=1}^s N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) \widehat{N}_i^{\ell[i]} \left(\int_{-1}^{\cdot} \widehat{B}_j(\hat{s}) d\hat{s} \right). \end{aligned} \quad (2.2a)$$

Here, $N_{i,n}^{\ell[i]}(U(t_{n-1}^-))$ means the application of $N_{i,n}^{\ell[i]}$ to the constant function $t \mapsto U(t_{n-1}^-)$. Now, in the case of collocation methods we have that

$$N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) = N_{j,n}^{\ell[j]}(F(\cdot, N_{j,n}^{\ell[j]}(U(\cdot)))) \quad (2.2b)$$

since for those methods $N_{j,n}^{\ell[j]} = N_{j,n}^0$ is just a function evaluation at a single point. Then, (2.2) gives a nonlinear equation system for $N_{i,n}^{\ell[i]}(U(\cdot))$, $i = 1, \dots, s$, because $N_{i,n}^{\ell[i]}(U(t_{n-1}^-))$ can be calculated from known data. But, if $N_{j,n}^{\ell[j]}$ for example represents the integral mean over I_n , then (2.2b) does not hold in general. Therefore, we will restrict ourselves to affine linear problems with time-independent coefficients of the form (1.21), i.e., we assume that $F(t, u) = f(t) - Au$. Then, for any linear functional $N_{j,n}^{\ell[j]}$ it holds

$$N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) = N_{j,n}^{\ell[j]}(f(\cdot)) - AN_{j,n}^{\ell[j]}(U(\cdot)).$$

Thus, (2.2b) is not necessary since only terms of the desired form $N_{j,n}^{\ell[j]}(U(\cdot))$ appear on the right-hand side anyway.

Altogether, for U solving (2.1) with $F(t, u) = f(t) - Au$ we have found that

$$\begin{aligned} MN_{i,n}^{\ell[i]}(U(\cdot)) &= MN_{i,n}^{\ell[i]}(U(t_{n-1}^-)) + \frac{\tau_n}{2} \sum_{j=1}^s \left(N_{j,n}^{\ell[j]}(f(\cdot)) - AN_{j,n}^{\ell[j]}(U(\cdot)) \right) \widehat{N}_i^{\ell[i]} \left(\int_{-1}^{\cdot} \widehat{B}_j(\hat{s}) d\hat{s} \right) \end{aligned} \quad (2.3a)$$

for all $i = 1, \dots, s$. Moreover, it easily follows

$$MU(t_n^-) = MU(t_{n-1}^-) + \frac{\tau_n}{2} \sum_{j=1}^s \left(N_{j,n}^{\ell[j]}(f(\cdot)) - AN_{j,n}^{\ell[j]}(U(\cdot)) \right) \int_{-1}^1 \widehat{B}_j(\hat{s}) d\hat{s}. \quad (2.3b)$$

These equations (2.3) could be seen as some generalization of a Runge–Kutta scheme in the style of [38, Proposition IV.3.1, p. 40] for the affine linear problem (1.21). Here, the equations (2.3a) could be interpreted as generalized stage equations for the “internal stages” $N_{i,n}^{\ell[i]}(U(\cdot))$, $i = 1, \dots, s$.

For a second extension, we recall that according to (1.3) not only the function value of U at t_{n-1} can be inherited from the previous interval but (depending on k) also evaluations of derivatives. In the derivation of “stage” and “iteration” equations we thus could afford also higher derivatives at t_{n-1} . So, in generalization of (2.1) we suppose that $U \in P_s(I_n, \mathbb{R}^d)$ satisfies

$$U^{(l)}(t_{n-1}^+) = U^{(l)}(t_{n-1}^-), \quad 0 \leq l \leq \bar{l}, \quad MU'(\cdot) = \mathcal{P}_n F(\cdot, U(\cdot)) \quad (2.4)$$

for some $\bar{l} \geq 0$. Then, similar as above, we also gain for $0 \leq l \leq \bar{l}$ and $t \in I_n$ that

$$\begin{aligned} MU^{(l)}(t) &= MU^{(l)}(t_{n-1}^+) + \int_{t_{n-1}}^t MU^{(l+1)}(\tilde{s}) d\tilde{s} \\ &= MU^{(l)}(t_{n-1}^-) + \int_{t_{n-1}}^t (\mathcal{P}_n F(\cdot, U(\cdot)))^{(l)}(\tilde{s}) d\tilde{s} \\ &= MU^{(l)}(t_{n-1}^-) + \sum_{j=1}^s N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) \left(\frac{\tau_n}{2}\right)^{1-l} \int_{-1}^{T_n^{-1}(t)} \widehat{B}_j^{(l)}(\hat{s}) d\hat{s}. \end{aligned}$$

Choosing for every $i = 1, \dots, s$ an $\ell_{[i]} \in \{0, \dots, \min\{\bar{l}, \ell^{[i]}\}\}$, we therefore have

$$\begin{aligned} \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}} MN_{i,n}^{\ell_{[i]}}(U(\cdot)) &= MN_{i,n}^{\ell_{[i]}-\ell_{[i]}}(U^{(\ell_{[i]})}(\cdot)) \\ &= MN_{i,n}^{\ell_{[i]}-\ell_{[i]}}(U^{(\ell_{[i]})}(t_{n-1}^-)) + \left(\frac{\tau_n}{2}\right)^{1-\ell_{[i]}} \sum_{j=1}^s N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) \widehat{N}_i^{\ell_{[i]}-\ell_{[i]}} \left(\int_{-1}^{\cdot} \widehat{B}_j^{(\ell_{[i]})}(\hat{s}) d\hat{s} \right) \end{aligned} \quad (2.5a)$$

and

$$MU^{(l)}(t_n^-) = MU^{(l)}(t_{n-1}^-) + \left(\frac{\tau_n}{2}\right)^{1-l} \sum_{j=1}^s N_{j,n}^{\ell[j]}(F(\cdot, U(\cdot))) \int_{-1}^1 \widehat{B}_j^{(l)}(\hat{s}) d\hat{s} \quad (2.5b)$$

for $0 \leq l \leq \bar{l}$.

2.1.2 A Runge–Kutta-like scheme

Let $s \in \mathbb{N}$, $\ell \in \mathbb{N}_0$, and fix for every $i = 1, \dots, s$ an $\ell_{[i]} \in \{0, \dots, \ell\}$. In addition, let $\{\widehat{N}_i^* \mid i = 1, \dots, s\}$ denote a set of linear functionals which are defined for sufficiently smooth functions on $[-1, 1]$. Usually, but not necessarily, those functionals are chosen such that a polynomial $\hat{v} \in P_{s-1}([-1, 1])$ is uniquely determined by the s values $\widehat{N}_i(\hat{v}) := \widehat{N}_i^*(\hat{v}^{(\ell_{[i]})})$, $i = 1, \dots, s$. Using the transformation T_n of (1.7) with $n = 1, \dots, N$, we set for sufficiently smooth functions v on \bar{I}_n

$$\begin{aligned} \widetilde{N}_{i,n}^*(v) &:= \widehat{N}_i^*(v \circ T_n), \quad \widetilde{N}_{i,n}(v) := \widetilde{N}_{i,n}^*(v^{(\ell_{[i]})}) = \widehat{N}_i^*(v^{(\ell_{[i]})} \circ T_n) \\ &= \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}} \widehat{N}_i^*((v \circ T_n)^{(\ell_{[i]})}) = \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}} \widehat{N}_i(v \circ T_n). \end{aligned}$$

Moreover, let g be an approximation of f which can be used in a local scheme instead of f .

Motivated by (2.5) and in the style of [38, Proposition IV.3.1, p. 40], the local version (on I_n) of an s -stage Runge–Kutta-like formulation for the discretization of (1.21) given an approximation g of f as well as function value and ℓ derivatives at t_{n-1}^- , for short we write (s, ℓ) -**RKl**(g), should have the form

$$\begin{aligned} Mg_{i,n}^{\text{RKl}} &= M\widetilde{N}_{i,n}^*(U^{(\ell_{[i]})}(t_{n-1}^-)) + \frac{\tau_n}{2} \sum_{j=1}^s a_{ij}^{\text{RKl}} \left(\frac{\tau_n}{2}\right)^{\ell[j]-\ell_{[i]}} \left(\widetilde{N}_{j,n}(g(\cdot)) - Ag_{j,n}^{\text{RKl}} \right), \quad i = 1, \dots, s, \\ MU^{(i)}(t_n^-) &= MU^{(i)}(t_{n-1}^-) + \frac{\tau_n}{2} \sum_{j=1}^s b_{(i+1)j}^{\text{RKl}} \left(\frac{\tau_n}{2}\right)^{\ell[j]-i} \left(\widetilde{N}_{j,n}(g(\cdot)) - Ag_{j,n}^{\text{RKl}} \right), \quad i = 0, \dots, \ell, \end{aligned}$$

where all a_{ij}^{RKl} and b_{ij}^{RKl} are real coefficients. Here, the $g_{i,n}^{\text{RKl}}$, $i = 1, \dots, s$, could be interpreted as generalized “internal stages”.

The coefficients are compressed in two matrices $A^{\text{RKl}} \in \mathbb{R}^{s \times s}$ and $B^{\text{RKl}} \in \mathbb{R}^{(\ell+1) \times s}$ that are given by $(A^{\text{RKl}})_{ij} = a_{ij}^{\text{RKl}}$ and $(B^{\text{RKl}})_{ij} = b_{ij}^{\text{RKl}}$, respectively. Moreover, a diagonal scaling matrix is defined by

$$S_n^{\text{RKl}} = \text{diag} \left(\left(\frac{\tau_n}{2} \right)^{\ell_{[1]}}, \dots, \left(\frac{\tau_n}{2} \right)^{\ell_{[s]}} \right) \in \mathbb{R}^{s \times s}.$$

For brevity, we further set $A_n^{\text{RKl}} \in \mathbb{R}^{s \times s}$ and $B_n^{\text{RKl}} \in \mathbb{R}^{(\ell+1) \times s}$ as

$$A_n^{\text{RKl}} = (S_n^{\text{RKl}})^{-1} A^{\text{RKl}} S_n^{\text{RKl}} \quad \text{and} \quad B_n^{\text{RKl}} = \text{diag} \left(1, \left(\frac{\tau_n}{2} \right)^{-1}, \dots, \left(\frac{\tau_n}{2} \right)^{-\ell} \right) B^{\text{RKl}} S_n^{\text{RKl}}.$$

Then, the Runge–Kutta-like formulation simply reads

$$\begin{pmatrix} MU(t_n^-) \\ \vdots \\ MU^{(\ell)}(t_n^-) \end{pmatrix} = \begin{pmatrix} MU(t_{n-1}^-) \\ \vdots \\ MU^{(\ell)}(t_{n-1}^-) \end{pmatrix} + \frac{\tau_n}{2} (B_n^{\text{RKl}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(g(\cdot)) - A g_{j,n}^{\text{RKl}} \\ \vdots \end{pmatrix} \quad (2.6a)$$

where

$$\begin{pmatrix} \vdots \\ M g_{i,n}^{\text{RKl}} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ M \tilde{N}_{i,n}^*(U^{(\ell_{[i]})}(t_{n-1}^-)) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} (A_n^{\text{RKl}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(g(\cdot)) - A g_{j,n}^{\text{RKl}} \\ \vdots \end{pmatrix}. \quad (2.6b)$$

Here, \otimes denotes the Kronecker product and $I_{d,d}$ the identity matrix in $\mathbb{R}^{d \times d}$.


Especially, we observe that the iteration equation (2.6a) uses and returns not just point values but also derivatives at t_{n-1}^- and t_n^- , respectively. Therefore, in general B_n^{RKl} is not a vector as for Runge–Kutta methods but a matrix.

Remark 2.1


Setting $Mk_{i,n}^{\text{RKl}} = \tilde{N}_{i,n}(g(\cdot)) - A g_{i,n}^{\text{RKl}}$, we could rewrite (s, ℓ) -**RKl**(g) as

$$Mk_{i,n}^{\text{RKl}} = \tilde{N}_{i,n}(g(\cdot)) - A \left(\tilde{N}_{i,n}^*(U^{(\ell_{[i]})}(t_{n-1}^-)) + \frac{\tau_n}{2} \sum_{j=1}^s a_{ij}^{\text{RKl}} \left(\frac{\tau_n}{2} \right)^{\ell_{[j]} - \ell_{[i]}} k_{j,n}^{\text{RKl}} \right), \quad i = 1, \dots, s,$$

$$MU^{(i)}(t_n^-) = MU^{(i)}(t_{n-1}^-) + \frac{\tau_n}{2} \sum_{j=1}^s b_{(i+1)j}^{\text{RKl}} \left(\frac{\tau_n}{2} \right)^{\ell_{[j]} - i} Mk_{j,n}^{\text{RKl}}, \quad i = 0, \dots, \ell,$$

which nicely shows the similarity to the other frequently used formulation of Runge–Kutta methods, see e.g. [37, Definition II.7.1, p. 205]. 

Remark 2.2

The classical Runge–Kutta method for the discretization of (1.21), completely described by the coefficient matrices $(A^{\text{RK}}, b^{\text{RK}}, c^{\text{RK}})$, is obtained for $\ell = \ell_{[i]} = 0$, $g = f$, $A^{\text{RKl}} = A^{\text{RK}}$, $B^{\text{RKl}} = b^{\text{RK}}$, and $\tilde{N}_i^*(\hat{v}) = \hat{v}(-1 + 2c_i^{\text{RK}})$ for all i . 

Remark 2.3

In order to fit (2.5) for problems of the form (1.21) in the Runge–Kutta-like framework, we can set

$$\hat{N}_i^*(\hat{v}) = \hat{N}_i^{\ell^{[i]} - \ell_{[i]}}(\hat{v}) \quad \text{and} \quad \hat{N}_i(\hat{v}) = \hat{N}_i^*(\hat{v}^{(\ell_{[i]})}) = \hat{N}_i^{\ell^{[i]} - \ell_{[i]}}(\hat{v}^{(\ell_{[i]})}) = \hat{N}_i^{\ell^{[i]}}(\hat{v}).$$

Note that then

$$\begin{aligned} \tilde{N}_{i,n}^*(v) &= \hat{N}_i^*(v \circ T_n) = \hat{N}_i^{\ell^{[i]} - \ell_{[i]}}(v \circ T_n) = N_{i,n}^{\ell^{[i]} - \ell_{[i]}}(v), \\ \tilde{N}_{i,n}(v) &= \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}} \hat{N}_i(v \circ T_n) = \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}} \hat{N}_i^{\ell^{[i]}}(v \circ T_n) = \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}} N_{i,n}^{\ell^{[i]}}(v). \end{aligned}$$

Therefore, we have that

$$\begin{aligned} g_{i,n}^{\text{RKL}} &\hat{=} \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}} N_{i,n}^{\ell^{[i]}}(U(\cdot)) = \tilde{N}_{i,n}(U(\cdot)), \\ a_{ij}^{\text{RKL}} &\hat{=} \hat{N}_i^{\ell^{[i]} - \ell_{[i]}} \left(\int_{-1}^{\cdot} \hat{B}_j^{(\ell_{[i]})}(\hat{s}) d\hat{s} \right) = \hat{N}_i^* \left(\int_{-1}^{\cdot} \hat{B}_j^{(\ell_{[i]})}(\hat{s}) d\hat{s} \right), \\ b_{(i+1)j}^{\text{RKL}} &\hat{=} \int_{-1}^1 \hat{B}_j^{(i)}(\hat{s}) d\hat{s}, \end{aligned}$$

where $\hat{B}_i \in P_{s-1}([-1, 1])$, $i = 1, \dots, s$, are determined by $\hat{N}_j(\hat{B}_i) = \hat{N}_j^{\ell^{[j]}}(\hat{B}_i) = \delta_{i,j}$. ♣

2.1.3 Existence and uniqueness

We now ask under which conditions the Runge–Kutta-like discretization scheme (2.6) has an appropriate solution. Of course, for a proper description of the solution, the involved “stage equations” (2.6b) should be uniquely solvable. Rewriting (2.6b), we easily see that this is guaranteed if the system matrix

$$((I_{s,s} \otimes M) + \frac{\tau_n}{2}(A_n^{\text{RKL}} \otimes A))$$

is regular.

Since M is regular by assumption, the system matrix also could be split in different ways, for example, as

- (i) $(I_{s,s} \otimes M)((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A_n^{\text{RKL}} \otimes M^{-1}A))$ or
- (ii) $((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A_n^{\text{RKL}} \otimes AM^{-1}))(I_{s,s} \otimes M)$.

Another, more symmetric splitting can be carried out if M is not just regular but symmetric and positive definite. Then, the square root $M^{1/2}$ of M is uniquely defined and also symmetric and positive definite. So, we have $M = M^{1/2}M^{1/2}$. In this case the system matrix can be rewritten as

- (iii) $(I_{s,s} \otimes M^{1/2})((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A_n^{\text{RKL}} \otimes M^{-1/2}AM^{-1/2}))(I_{s,s} \otimes M^{1/2})$.

For brevity, we will write

$$\overline{M} := \begin{cases} I_{d,d}, & \text{for splitting (i),} \\ M, & \text{for splitting (ii),} \\ M^{1/2}, & \text{for splitting (iii),} \end{cases} \quad \overline{A} := \overline{M}M^{-1}A\overline{M}^{-1} = \begin{cases} M^{-1}A, & \text{for (i),} \\ AM^{-1}, & \text{for (ii),} \\ M^{-1/2}AM^{-1/2}, & \text{for (iii).} \end{cases}$$

Then, all three splittings could be written in a unified form as

$$(I_{s,s} \otimes M \bar{M}^{-1})((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A_n^{\text{RKl}} \otimes \bar{A}))(I_{s,s} \otimes \bar{M}).$$

Setting $\bar{g}_{i,n}^{\text{RKl}} := \bar{M} g_{i,n}^{\text{RKl}}$ and left multiplying the equation system (2.6a) by $(I_{\ell+1,\ell+1} \otimes \bar{M} M^{-1})$ and the equation system (2.6b) by $(I_{s,s} \otimes \bar{M} M^{-1})$, we find

$$\begin{pmatrix} \bar{U}(t_n^-) \\ \vdots \\ \bar{U}^{(\ell)}(t_n^-) \end{pmatrix} = \begin{pmatrix} \bar{U}(t_{n-1}^-) \\ \vdots \\ \bar{U}^{(\ell)}(t_{n-1}^-) \end{pmatrix} + \frac{\tau_n}{2}(B_n^{\text{RKl}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(\bar{M} M^{-1} g(\cdot)) - \bar{A} \bar{g}_{j,n}^{\text{RKl}} \\ \vdots \end{pmatrix} \quad (2.7a)$$

and

$$\begin{pmatrix} \vdots \\ \bar{g}_{i,n}^{\text{RKl}} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \tilde{N}_{i,n}^*(\bar{U}^{(\ell[i])}(t_{n-1}^-)) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2}(A_n^{\text{RKl}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(\bar{M} M^{-1} g(\cdot)) - \bar{A} \bar{g}_{j,n}^{\text{RKl}} \\ \vdots \end{pmatrix}, \quad (2.7b)$$

which could be interpreted as the “iteration” and “stage equations” associated to the solution $\bar{U} = \bar{M} U$ of (s, ℓ) -**RKl** $(\bar{M} M^{-1} g)$ as approximation to the solution $\bar{u} = \bar{M} u$ of

$$\bar{u}'(t) = \bar{F}(t, \bar{u}(t)) := \bar{M} M^{-1} f(t) - \bar{A} \bar{u}(t), \quad \bar{u}(t_0) = \bar{M} u_0. \quad (2.8)$$


Due to the regularity (or even the symmetry and positive definiteness) of M , both equation systems (2.6) and (2.7) as well as both problems (1.21) and (2.8) are equivalent. Therefore, we will concentrate on the discretization of the somewhat more simple problem (2.8).

In the following, let $M \in \mathbb{R}^{d \times d}$ always be regular and \bar{M}, \bar{A} matrices in $\mathbb{R}^{d \times d}$. For the study of the existence and uniqueness of solutions to (2.7b), we need more notation.

Definition 2.4

For $\Lambda \in \mathbb{R}^{d \times d}$ we set

$$\mu[\Lambda] := \sup \{ (v, \Lambda v) : v \in \mathbb{R}^d, \|v\| = 1 \},$$

which is the largest eigenvalue of the symmetric part $\frac{1}{2}(\Lambda^T + \Lambda)$ of Λ . Also note that $\mu[\Lambda]$ is the logarithmic norm of Λ with respect to the Euclidean inner product, see [42, (2.1.2)] or [37, Theorem I.10.5, p. 61]. 

Remark 2.5


For $\mu \geq \mu[-\bar{A}]$ we easily verify that

$$(-\bar{A}v, v) = (v, -\bar{A}v) \leq \mu \|v\|^2 \quad \forall v \in \mathbb{R}^d. \quad (2.9)$$

Therefore, \bar{F} (with $\bar{F}(t, v) = \bar{M} M^{-1} f(t) - \bar{A} v$) satisfies the one-sided Lipschitz condition

$$(\bar{F}(t, \tilde{v}) - \bar{F}(t, v), \tilde{v} - v) \leq \mu \|\tilde{v} - v\|^2 \quad \forall t \in \bar{I}, \quad \forall \tilde{v}, v \in \mathbb{R}^d$$

with one-sided Lipschitz constant $\mu \in \mathbb{R}$.

Note that there is no restriction on the sign of μ . Indeed, in many situations of practical relevance μ is negative. For example, if \bar{A} is the stiffness matrix of a semi-discretization in space of a parabolic problem, then μ is a negative multiple of the coercivity constant. 

Further, we introduce some notation on functions with matrix arguments following [42, Subsection 2.2.2, Subsection 2.4.3].

Denote by ϕ_{denom} and ϕ_{num} two polynomials without common non-trivial factors and consider the rational function ϕ given by $\phi(z) = (\phi_{\text{denom}}(z))^{-1} \phi_{\text{num}}(z)$ for all $z \in \mathbb{C}$ with $\phi_{\text{denom}}(z) \neq 0$. Let $\Lambda \in \mathbb{R}^{d \times d}$. Then, provided $\phi_{\text{denom}}(\Lambda)$ is regular, we say that $\phi(\Lambda)$ exists and is given by $\phi(\Lambda) = (\phi_{\text{denom}}(\Lambda))^{-1} \phi_{\text{num}}(\Lambda)$.

Moreover, let Φ be a matrix-valued function given by $\Phi(z) = (\phi_{ij}(z)) \in \mathbb{C}^{s \times s}$ whenever $z \in \mathbb{C}$ and all $\phi_{ij}(z)$ are well-defined where ϕ_{ij} are rational functions with real coefficients. Then, for $\Lambda \in \mathbb{R}^{d \times d}$, we denote by $\Phi(\Lambda)$ the $sd \times sd$ matrix with block-entries $\phi_{ij}(\Lambda) \in \mathbb{R}^{d \times d}$. Of course, we say that $\Phi(\Lambda)$ exists if all $\phi_{ij}(\Lambda)$ exist.

Using [42, Lemma 2.4.6, Lemma 2.2.6] it can be shown that the regularity of the system matrix $((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A_n^{\text{RKl}} \otimes \bar{A}))$ is strongly connected to the properties of the matrix-valued function given by $z \in \mathbb{C} \mapsto (I_{s,s} + A^{\text{RKl}}z)$. In fact, the following lemma holds.

Lemma 2.6

Let $\bar{\tau} > 0$ and $\mu \geq \mu[-\bar{A}]$. Then, the system matrix $((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A_n^{\text{RKl}} \otimes \bar{A}))$ is regular for all $\tau_n \in (0, \bar{\tau}]$ if

$$(I_{s,s} - A^{\text{RKl}}z) \text{ is regular for all } z \in \mathbb{C} \text{ with } \operatorname{Re} z \leq \max\{0, \frac{1}{2}\bar{\tau}\mu\}.$$

Proof. First of all, $A_n^{\text{RKl}} = (S_n^{\text{RKl}})^{-1} A^{\text{RKl}} S_n^{\text{RKl}}$ implies

$$\begin{aligned} ((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A_n^{\text{RKl}} \otimes \bar{A})) \\ = ((S_n^{\text{RKl}})^{-1} \otimes I_{d,d}) ((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A^{\text{RKl}} \otimes \bar{A})) (S_n^{\text{RKl}} \otimes I_{d,d}). \end{aligned}$$

Hence, it suffices to study the regularity of $((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A^{\text{RKl}} \otimes \bar{A}))$, which we will call the main part of the system matrix.

For $z \in \mathbb{C}$, let $V(z) = (v_{ij}(z)) = (I_{s,s} - A^{\text{RKl}}z)$ and $W(z) = (w_{ij}(z)) = V(z)^{-1}$ if $V(z)$ is regular. Recalling the notation for matrix-valued functions, the main part of the system matrix $((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A^{\text{RKl}} \otimes \bar{A}))$ can be shortly written as $V(-\frac{\tau_n}{2}\bar{A})$. Now, according to [42, Lemma 2.4.6], we have that $V(-\frac{\tau_n}{2}\bar{A})$ is regular if and only if $W(-\frac{\tau_n}{2}\bar{A})$ exists. So, of course, we shall ask whether all $w_{ij}(-\frac{\tau_n}{2}\bar{A})$ exist.

Now, it is well-known that (if existing) the inverse matrix of $V(z)$, $z \in \mathbb{C}$, can be written as $V(z)^{-1} = \frac{1}{\det(V(z))} \operatorname{adj}(V(z))$ where $\operatorname{adj}(V(z))$ denotes the adjugate of $V(z)$. Exploiting this representation, we find that w_{ij} is a rational function with $w_{ij}(z) = \frac{(\operatorname{adj}(V(z)))_{ij}}{\det(V(z))}$. Obviously, $w_{ij}(z)$ exists for $z \in \mathbb{C}$ if $\det(V(z)) \neq 0$, i.e., if $V(z)$ is regular. Moreover, because of $\mu \geq \mu[-\bar{A}]$ (also cf. (2.9)), we obtain from [42, Lemma 2.2.6] that all $w_{ij}(-\frac{\tau_n}{2}\bar{A})$ exist and, thus, immediately get that $((I_{s,s} \otimes I_{d,d}) + \frac{\tau_n}{2}(A^{\text{RKl}} \otimes \bar{A}))^{-1} = V(-\frac{\tau_n}{2}\bar{A})^{-1} = W(-\frac{\tau_n}{2}\bar{A})$ exists if $\det(V(z))$ has no zeros in $\{z \in \mathbb{C} : \operatorname{Re} z \leq \frac{\tau_n}{2}\mu\}$. This completes the proof. \square

Remark 2.7

The statement of Lemma 2.6 is quite interesting in many ways. First of all, we find that in order to guarantee that (2.7b) (and consequently also (2.6b)) has a unique solution, $(I_{s,s} - A^{\text{RKl}}z)$ should be regular at least on $\mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re}(z) \leq 0\}$. Indeed, then $(I_{s,s} - A^{\text{RKl}}z)$ would even be regular on $\{z \in \mathbb{C} : \operatorname{Re}(z) \leq \omega\}$ for some $\omega > 0$ and, thus,

also the system matrix is regular for $\tau_n \in (0, \bar{\tau}]$ when $\bar{\tau}\mu \leq 2\omega$. In general this yields an upper bound for the time step length. However, we observe that for $\mu \leq 0$ no restrictions on $\tau_n > 0$ are necessary. ♣

As we shall prove below, the regularity of $(I_{s,s} - A^{\text{RKl}}z)$ is strongly connected to the eigenvalues of A^{RKl} . Let $\sigma(\Lambda)$ denote the spectrum of the matrix Λ . Moreover, we define some special regions of the complex plane \mathbb{C} by

$$\begin{aligned}\mathbb{C}^- &:= \{z \in \mathbb{C} : \operatorname{Re}(z) \leq 0\}, \\ \mathbb{C}_0^+ &:= \{z \in \mathbb{C} : z = 0 \text{ or } \operatorname{Re}(z) > 0\}.\end{aligned}$$

Then, we have the following statements that especially also hold for $\Lambda = A^{\text{RKl}}$.

Lemma 2.8

Let $\Lambda \in \mathbb{R}^{s \times s}$ and $z \in \mathbb{C} \setminus \{0\}$. Then, $(I_{s,s} - \Lambda z)$ is regular if and only if $z^{-1} \notin \sigma(\Lambda)$.

Proof. The matrix $(I_{s,s} - \Lambda z)$ with $z \in \mathbb{C} \setminus \{0\}$ is singular if and only if

$$\det(I_{s,s} - \Lambda z) = 0 \quad \Leftrightarrow \quad \det(z^{-1}I_{s,s} - \Lambda) = 0 \quad \Leftrightarrow \quad z^{-1} \in \sigma(\Lambda),$$

which immediately gives the desired statement. □

Corollary 2.9 (Cf. [19, Lemma 4.1])

Let $\Lambda \in \mathbb{R}^{s \times s}$. Then, $(I_{s,s} - \Lambda z)$ is regular on \mathbb{C}^- if and only if $\sigma(\Lambda) \subset \mathbb{C}_0^+$.

Proof. For $z = 0$ the matrix is always regular. Otherwise, we gain by Lemma 2.8 that

$$\begin{aligned}(I_{s,s} - \Lambda z) \text{ is regular on } \mathbb{C}^- \setminus \{0\} &\Leftrightarrow \{z^{-1} : z \in \mathbb{C}^- \setminus \{0\}\} \subset \mathbb{C} \setminus \sigma(\Lambda) \\ &\Leftrightarrow \sigma(\Lambda) \subset \mathbb{C} \setminus \{z^{-1} : z \in \mathbb{C}^- \setminus \{0\}\} = \mathbb{C}_0^+, \end{aligned}$$

where we also used that $\{z^{-1} : z \in \mathbb{C}^- \setminus \{0\}\} = \mathbb{C}^- \setminus \{0\} = \mathbb{C} \setminus \mathbb{C}_0^+$. □

2.1.4 Stability properties

In order to enable a proper approximation of the global error, especially for the discretization of stiff problems, a discretization method needs to fulfill certain stability properties. Since we consider affine linear problems, we do not need (and in general also do not have) B -stability for the methods we are particularly interested in. However, we shall have a look on some similar stability concepts in analogy to [19, Subsection 3.1].

Definition 2.10

The Runge–Kutta-like method (2.7) is called *ASI-stable* if $(I_{s,s} - A^{\text{RKl}}z)$ is regular as well as $(I_{s,s} - A^{\text{RKl}}z)^{-1}$ is uniformly bounded for all $z \in \mathbb{C}^-$. ♣

Definition 2.11

The Runge–Kutta-like method (2.7) is called *AS-stable* if $(I_{s,s} - A^{\text{RKl}}z)$ is regular as well as $B^{\text{RKl}}z(I_{s,s} - A^{\text{RKl}}z)^{-1}$ is uniformly bounded for all $z \in \mathbb{C}^-$. ♣

The Definitions 2.10 and 2.11 simply transfer the concepts of *ASI*- and *AS*-stability, respectively, which are typically studied for Runge–Kutta methods, to the Runge–Kutta-like method (2.7).

Analogously to [19, Lemma 4.3] it can be shown that the *ASI*-stability is strongly connected to the spectrum of the matrix A^{Rkl} . In fact, the following lemma holds.

Lemma 2.12

The Runge–Kutta-like method (2.7) is ASI-stable if $\sigma(A^{\text{Rkl}}) \subset \mathbb{C}_0^+$ and zero is at most a simple eigenvalue of A^{Rkl} .

Proof. First of all, Corollary 2.9 yields that the matrix $(I_{s,s} - A^{\text{Rkl}}z)$ is regular for all $z \in \mathbb{C}^-$ if and only if $\sigma(A^{\text{Rkl}}) \subset \mathbb{C}_0^+$. Furthermore, according to [19, Lemma 4.3], we have that $(I_{s,s} - A^{\text{Rkl}}z)^{-1}$ is uniformly bounded if $\sigma(A^{\text{Rkl}}) \subset \mathbb{C}_0^+$ and zero is at most a simple eigenvalue of A^{Rkl} . This completes the proof. \square

Furthermore, under certain conditions it can be shown that for Runge–Kutta-like methods of the form (2.7) the *ASI*-stability already implies the *AS*-stability. This result and its proof are quite similar to that of [19, Lemma 4.4].

Lemma 2.13

Assume that $B^{\text{Rkl}} = CA^{\text{Rkl}}$ for some matrix C . Then, any ASI-stable Runge–Kutta-like method of form (2.7) also is AS-stable.

Proof. Because of $B^{\text{Rkl}} = CA^{\text{Rkl}}$, we gain that

$$\begin{aligned} B^{\text{Rkl}}z(I_{s,s} - A^{\text{Rkl}}z)^{-1} &= C((A^{\text{Rkl}}z - I_{s,s}) + I_{s,s})(I_{s,s} - A^{\text{Rkl}}z)^{-1} \\ &= C(I_{s,s} - A^{\text{Rkl}}z)^{-1} - C. \end{aligned}$$

Now, if the method is *ASI*-stable, we easily verify that this matrix is well-defined and uniformly bounded for $z \in \mathbb{C}^-$, and thus *AS*-stable. \square

Since A^{Rkl} is not necessarily regular, the condition $B^{\text{Rkl}} = CA^{\text{Rkl}}$ really is an additional assumption. However, in many cases this assumption is fulfilled, especially also for the methods of our interest, see Corollary 2.16 below.

2.2 VTD methods as Runge–Kutta-like discretizations

This section aims to give a convenient description for the discrete **VTD** solution \bar{U} in terms of a Runge–Kutta-like formulation. Here, unlike for collocation methods and their Runge–Kutta formulation, the internal stages will in general not represent function values of the discrete solution at intermediate time points but certain (other) degrees of freedom. For convenience we will consider the somewhat simpler problem representation (2.8) only.

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. For $i = 1, \dots, r+1$ set $\ell_{[i]} := \min\{i-1, \lfloor \frac{k}{2} \rfloor\}$. Moreover, define

$$\hat{N}_i^*(\hat{v}) = \hat{v}(1^-), \quad i = 1, \dots, \lfloor \frac{k}{2} \rfloor, \quad \hat{N}_{\lfloor \frac{k}{2} \rfloor + i}^*(\hat{v}) = \hat{v}(\tilde{t}_i), \quad i = 1, \dots, r - \lfloor \frac{k}{2} \rfloor + 1, \quad (2.10)$$

where $\tilde{t}_i \in [-1, 1]$, $i = 1, \dots, r - \lfloor \frac{k}{2} \rfloor + 1$, denote the quadrature points of $Q_{k-2\lfloor k/2 \rfloor}^{r-\lfloor k/2 \rfloor}$, which is Gauss–Radau for k even or Gauss–Lobatto for k odd, respectively. This implies that

$$\hat{N}_{r+1}^*(\hat{v}) = \hat{v}(1^-) \quad \text{and} \quad \hat{N}_{\lfloor \frac{k}{2} \rfloor + 1}^*(\hat{v}) = \hat{v}(-1^+), \quad \text{if } k \text{ is odd.}$$

Also note that these functionals are such that for a constant function $\hat{v} \equiv c$ they return c .

It can be easily shown that a function $\hat{v} \in P_r((-1, 1])$ can be uniquely described by the $r + 1$ degrees of freedom $\hat{N}_i(\hat{v}) := \hat{N}_i^*(\hat{v}^{(\ell_{[i]})})$, $i = 1, \dots, r + 1$. Thus, any function $\hat{v} \in P_r((-1, 1])$ can be written as

$$\hat{v}(\hat{t}) = \sum_{i=1}^{r+1} \hat{N}_i(\hat{v}) \hat{B}_i(\hat{t}) \quad \forall \hat{t} \in (-1, 1],$$

where the $\hat{B}_i \in P_r((-1, 1])$, $i = 1, \dots, r + 1$, denote the associated basis functions, i.e., it holds $\hat{N}_i(\hat{B}_j) = \hat{N}_i^*(\hat{B}_j^{(\ell_{[i]})}) = \delta_{i,j}$ for all $i, j = 1, \dots, r + 1$.

Because of the special choice of the degrees of freedom, we have that the basis functions \hat{B}_j , $1 \leq j \leq \lfloor \frac{k}{2} \rfloor$, which are associated to the function and derivative values at 1^- , are simply given by $\hat{B}_j(\hat{t}) = \frac{(-1)^{j-1}}{(j-1)!} (1 - \hat{t})^{j-1}$. This implies that for $\hat{v} \in P_r((-1, 1])$ and $l = 0, \dots, \lfloor \frac{k}{2} \rfloor$

$$\hat{v}^{(l)}(\hat{t}) = \sum_{i=1}^{r+1} \hat{N}_i(\hat{v}) \hat{B}_i^{(l)}(\hat{t}) = \sum_{i=l+1}^{r+1} \hat{N}_i(\hat{v}) \hat{B}_i^{(l)}(\hat{t}) \quad \forall \hat{t} \in (-1, 1].$$

Also note that by construction $\hat{N}_i^*(\hat{B}_j^{(\ell_{[i]})}) = \delta_{i,j}$ for $i, j = \lfloor \frac{k}{2} \rfloor + 1, \dots, r + 1$, which especially implies that $\hat{B}_i^{(\ell_{[i]})} \in P_{r-\lfloor k/2 \rfloor}((-1, 1])$, $i = \lfloor \frac{k}{2} \rfloor + 1, \dots, r + 1$, are just the nodal basis functions associated to the set of linear functionals $\{\hat{N}_i^* \mid i = \lfloor \frac{k}{2} \rfloor + 1, \dots, r + 1\}$.

For the sake of clarity, we concretize the general operators of Subsection 2.1.2 in our current setting. We have for sufficiently smooth functions v on \bar{I}_n

$$\begin{aligned} \tilde{N}_{i,n}^*(v) &:= \hat{N}_i^*(v \circ T_n) = v(t_n^-), & i = 1, \dots, \lfloor \frac{k}{2} \rfloor, \\ \tilde{N}_{\lfloor \frac{k}{2} \rfloor + i, n}^*(v) &:= \hat{N}_{\lfloor \frac{k}{2} \rfloor + i}^*(v \circ T_n) = v(T_n(\tilde{t}_i)), & i = 1, \dots, r - \lfloor \frac{k}{2} \rfloor + 1, \end{aligned} \quad (2.11a)$$

where the transformation T_n is given by (1.7). Moreover, it holds

$$\begin{aligned} \tilde{N}_{i,n}(v) &:= \tilde{N}_{i,n}^*(v^{(\ell_{[i]})}) = v^{(\ell_{[i]})}(t_n^-) = v^{(i-1)}(t_n^-), & i = 1, \dots, \lfloor \frac{k}{2} \rfloor, \\ \tilde{N}_{\lfloor \frac{k}{2} \rfloor + i, n}(v) &:= \tilde{N}_{\lfloor \frac{k}{2} \rfloor + i, n}^*(v^{(\lfloor \frac{k}{2} \rfloor)}) = v^{(\lfloor \frac{k}{2} \rfloor)}(T_n(\tilde{t}_i)), & i = 1, \dots, r - \lfloor \frac{k}{2} \rfloor + 1. \end{aligned} \quad (2.11b)$$

Of course, the l th derivative, $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$, of a function $v \in P_r(I_n)$ then is represented by

$$v^{(l)}(t) = \sum_{i=l+1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell_{[i]}-l} \tilde{N}_{i,n}(v) (\hat{B}_i^{(l)} \circ T_n^{-1})(t), \quad \forall t \in I_n. \quad (2.12)$$

We now are ready to reveal a Runge–Kutta-like formulation of the **VTD** methods. Recall that, for convenience, we only consider the more simple problem (2.8). Moreover, we need some assumptions on g .

Assumption 2.1

We assume that g is $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable on \bar{I} . Moreover, we suppose that g satisfies

$$g|_{I_n} \in P_r(I_n, \mathbb{R}^d) \quad \text{for all } 1 \leq n \leq N$$

as well as

$$g^{(i)}(t_n) = f^{(i)}(t_n) \quad \text{for all } 0 \leq n \leq N \text{ and } i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1$$

if the right-hand side function $f : \bar{I} \rightarrow \mathbb{R}^d$ is sufficiently smooth.

Remark 2.14

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Typical choices for g that fulfill Assumption 2.1 (for sufficiently smooth f) are

- $g = \Pi_k^r f$: Here, Π_k^r is defined in (1.28).
- $g = \mathcal{I}_k^r f$: For a definition of \mathcal{I}_k^r compare (1.15).
- $g = \mathcal{C}_k^r f$: Here, $\mathcal{C}_k^r = \mathcal{I}_k^r \circ \mathcal{I}_{k+2}^{r+1} \circ \dots \circ \mathcal{I}_{2r-k}^{2r-k}$ is the interpolation cascade known from Subsection 1.4.3.

For a convenient interpretation of these choices see Remark 1.43.

Note that, for $k \geq 2$, also the situation after a postprocessing of $Q_{k-2}^{r-1} \text{-VTD}_{k-2}^{r-1}(f)$ can be described by a g satisfying Assumption 2.1. Indeed, for this case set $g = \mathcal{I}_{k-2,*}^r f$, where $\mathcal{I}_{k-2,*}^r$ is, in accordance with Remark 1.37, the interpolation operator which interpolates in the quadrature points of Q_{k-2}^{r-1} and additionally preserves the $\lfloor \frac{k-1}{2} \rfloor$ th derivative in t_{n-1}^+ . ♣

Proposition 2.15

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Moreover, let $\{\hat{N}_i^*\}$ be specified by (2.10), $\ell_{[i]} = \min\{i-1, \lfloor \frac{k}{2} \rfloor\}$, and $\ell = \lfloor \frac{k}{2} \rfloor$. Furthermore, suppose that f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable and that $\bar{U} \in Y_r$ solves $\text{VTD}_k^r(\bar{M}M^{-1}g)$ with g fulfilling Assumption 2.1. Then, $\bar{g}_{i,n}^{\text{VTD}} = \tilde{N}_{i,n}(\bar{U}(\cdot)) \in \mathbb{R}^d$, $i = 1, \dots, r+1$, satisfy

$$\begin{pmatrix} \vdots \\ \bar{g}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \bar{U}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(\bar{M}M^{-1}g(\cdot)) - \bar{A}\bar{g}_{j,n}^{\text{VTD}} \\ \vdots \end{pmatrix} \quad (2.13)$$

with $\bar{U}^{(\ell_{[i]})}(t_0^-) = \bar{u}^{(\ell_{[i]})}(t_0)$. Here, we have

$$A_n^{\text{VTD}} := S_n^{-1} A^{\text{VTD}} S_n,$$

where $A^{\text{VTD}} \in \mathbb{R}^{(r+1) \times (r+1)}$ is given by

$$(A^{\text{VTD}})_{ij} = \begin{cases} \hat{N}_i^* \left(\int_{-1}^1 \hat{B}_j^{(\ell_{[i]})}(\hat{s}) d\hat{s} \right), & 1 \leq i \leq r+1, \ell_{[i]} + 1 = \min\{i, \lfloor \frac{k}{2} \rfloor + 1\} \leq j \leq r+1, \\ 0, & \text{otherwise,} \end{cases} \quad (2.14)$$

and $S_n \in \mathbb{R}^{(r+1) \times (r+1)}$ is the diagonal scaling matrix $S_n = \text{diag} \left(1, \left(\frac{\tau_n}{2} \right)^{\ell_{[2]}}, \dots, \left(\frac{\tau_n}{2} \right)^{\ell_{[r+1]}} \right)$.

Proof. The argument strongly uses various properties of the postprocessed solution and its connection to the actual solution which therefore shall be summarized at first.

The postprocessing of Theorem 1.32 can be applied to the solution \bar{U} of $\mathbf{VTD}_k^r(\bar{M}M^{-1}g)$ with g satisfying Assumption 2.1. This procedure then yields a solution $\check{U} \in Y_{r+1}$ of the $Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(\bar{M}M^{-1}g)$ method that coincides with \bar{U} in the $r+1$ quadrature points of Q_k^r . More concrete, there is a vector $c_n \in \mathbb{R}^d$ such that

$$\check{U}|_{I_n} = \bar{U}|_{I_n} + c_n(\phi \circ T_n^{-1})$$

$$\text{with } \phi(\hat{t}) = (1 - \hat{t})^{\lfloor \frac{k}{2} \rfloor + 1} (1 + \hat{t})^{\lfloor \frac{k-1}{2} \rfloor + 1} P_{r-k}^{\lfloor \frac{k}{2} \rfloor + 1, \lfloor \frac{k-1}{2} \rfloor + 1}(\hat{t}) \quad \forall \hat{t} \in [-1, 1],$$

where $P_{r-k}^{\lfloor \frac{k}{2} \rfloor + 1, \lfloor \frac{k-1}{2} \rfloor + 1}$ denotes the $(r-k)$ th Jacobi-polynomial with respect to the weighting function $(1 - \hat{t})^{\lfloor \frac{k}{2} \rfloor + 1} (1 + \hat{t})^{\lfloor \frac{k-1}{2} \rfloor + 1}$ in the interval $(-1, 1)$, cf. Appendix A.2. Therefore, as a consequence of Rodrigues' formula we further gain that $\bar{U}(\lfloor \frac{k}{2} \rfloor)|_{I_n}$ and $\check{U}(\lfloor \frac{k}{2} \rfloor)|_{I_n}$ also coincide in the quadrature points of $Q_{k-2\lfloor k/2 \rfloor}^{r-\lfloor k/2 \rfloor}$ since

$$\phi(\lfloor \frac{k}{2} \rfloor)(\hat{t}) = C (1 - \hat{t}) (1 + \hat{t})^{\lfloor \frac{k-1}{2} \rfloor + 1 - \lfloor \frac{k}{2} \rfloor} P_{r-\lfloor \frac{k-1}{2} \rfloor - 1}^{(1, \lfloor \frac{k-1}{2} \rfloor + 1 - \lfloor \frac{k}{2} \rfloor)}(\hat{t}) \quad \forall \hat{t} \in [-1, 1]$$

due to (A.2). Altogether this implies that

$$\tilde{N}_{i,n}(\bar{U}(\cdot)) = \tilde{N}_{i,n}(\check{U}(\cdot)) \quad \text{for all } i = 1, \dots, r+1. \quad (2.15a)$$

Moreover, recalling Theorem 1.24, we could also interpret \check{U} as a solution of a certain collocation method. Then, in particular, we find that $\mathcal{I}_k^r(\check{U}')|_{I_n} = \mathcal{I}_k^r(\bar{M}M^{-1}g - \bar{A}\check{U})|_{I_n}$. Using that M , \bar{M} , and \bar{A} are independent of t as well as that \check{U}' and g are polynomials of maximal degree r , we thus conclude

$$\begin{aligned} \check{U}'(t) &= \mathcal{I}_k^r(\check{U}')(t) = \mathcal{I}_k^r(\bar{M}M^{-1}g - \bar{A}\check{U})(t) = (\bar{M}M^{-1}\mathcal{I}_k^r g - \bar{A}\mathcal{I}_k^r \check{U})(t) \\ &= (\bar{M}M^{-1}g - \bar{A}\bar{U})(t) \end{aligned} \quad (2.15b)$$

for all $t \in I_n$. Last but not least, by construction the solution \check{U} of $Q_k^r\text{-}\mathbf{VTD}_{k+2}^{r+1}(\bar{M}M^{-1}g)$ is at least $\lfloor \frac{k}{2} \rfloor$ -times continuously differentiable since f and g are sufficiently smooth by assumption. Hence, we obtain

$$\check{U}^{(i)}(t_{n-1}^+) = \check{U}^{(i)}(t_{n-1}^-) = \begin{cases} \bar{U}^{(i)}(t_{n-1}^-), & n > 1, \\ \bar{u}^{(i)}(t_0), & n = 1, \end{cases} \quad \text{for } i = 0, \dots, \lfloor \frac{k}{2} \rfloor, \quad (2.15c)$$

where, for $n > 1$, we used that the derivatives of ϕ up to order $\lfloor \frac{k}{2} \rfloor$ vanish at 1^- and, for $n = 1$, we used the definition of $\bar{u}^{(i)}(t_0)$, the collocation conditions at t_0^+ , and that $g^{(i)}(t_0^+) = f^{(i)}(t_0^+)$ for $i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1$.

Now, we are ready to start the actual proof. For any $l = 0, \dots, \lfloor \frac{k}{2} \rfloor$ and $t \in I_n$ we gain for $\check{U} \in Y_{r+1}$ from the fundamental theorem of calculus and (2.12) that

$$\begin{aligned} \check{U}^{(l)}(t) &= \check{U}^{(l)}(t_{n-1}^+) + \int_{t_{n-1}}^t \check{U}^{(l+1)}(\tilde{s}) \, d\tilde{s} \\ &= \check{U}^{(l)}(t_{n-1}^+) + \sum_{j=l+1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell_{[j]}-l} \tilde{N}_{j,n}(\check{U}'(\cdot)) \int_{t_{n-1}}^t (\hat{B}_j^{(l)} \circ T_n^{-1})(\tilde{s}) \, d\tilde{s} \\ &= \check{U}^{(l)}(t_{n-1}^+) + \frac{\tau_n}{2} \sum_{j=l+1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell_{[j]}-l} \tilde{N}_{j,n}(\check{U}'(\cdot)) \int_{-1}^{T_n^{-1}(t)} \hat{B}_j^{(l)}(\hat{s}) \, d\hat{s}, \end{aligned} \quad (2.16)$$

where the last identity follows from integration by substitution.

For $i = 1, \dots, r+1$ applying $\tilde{N}_{i,n}^*$ to (2.16) with $l = \ell_{[i]} = \min\{i-1, \lfloor \frac{k}{2} \rfloor\}$, we obtain

$$\begin{aligned} \tilde{N}_{i,n}^*(\check{U}^{(\ell_{[i]})}(\cdot)) \\ = \tilde{N}_{i,n}^*(\check{U}^{(\ell_{[i]})}(t_{n-1}^+)) + \frac{\tau_n}{2} \sum_{j=\ell_{[i]}+1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell_{[j]}-\ell_{[i]}} \tilde{N}_{j,n}(\check{U}'(\cdot)) \tilde{N}_{i,n}^*\left(\int_{-1}^{T_n^{-1}(\cdot)} \hat{B}_j^{(\ell_{[i]})}(\hat{s}) \, d\hat{s}\right). \end{aligned}$$

So, from the definitions and properties of the linear functionals, we further conclude that

$$\tilde{N}_{i,n}(\check{U}(\cdot)) = \check{U}^{(\ell_{[i]})}(t_{n-1}^+) + \frac{\tau_n}{2} \sum_{j=\ell_{[i]}+1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell_{[j]}-\ell_{[i]}} \tilde{N}_{j,n}(\check{U}'(\cdot)) \hat{N}_i^*\left(\int_{-1}^{T_n^{-1}(\cdot)} \hat{B}_j^{(\ell_{[i]})}(\hat{s}) \, d\hat{s}\right)$$

for $i = 1, \dots, r+1$. Recalling the definition of A^{VTD} , this identity simply reads

$$\tilde{N}_{i,n}(\check{U}(\cdot)) = \check{U}^{(\ell_{[i]})}(t_{n-1}^+) + \frac{\tau_n}{2} \sum_{j=1}^{r+1} (A^{\text{VTD}})_{ij} \left(\frac{\tau_n}{2}\right)^{\ell_{[j]}-\ell_{[i]}} \tilde{N}_{j,n}(\check{U}'(\cdot)).$$

Noting that the scaling matrices S_n and S_n^{-1} are defined such that

$$(S_n)_{ij} = \begin{cases} \left(\frac{\tau_n}{2}\right)^{\ell_{[i]}}, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases} \quad \text{and} \quad (S_n^{-1})_{ij} = \begin{cases} \left(\frac{\tau_n}{2}\right)^{-\ell_{[i]}}, & \text{if } i = j, \\ 0, & \text{otherwise,} \end{cases}$$

respectively, we gain the equation system

$$\begin{pmatrix} \vdots \\ \tilde{N}_{i,n}(\check{U}(\cdot)) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \check{U}^{(\ell_{[i]})}(t_{n-1}^+) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} \underbrace{(S_n^{-1} A^{\text{VTD}} S_n)}_{=: A_n^{\text{VTD}}} \otimes I_{d,d} \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(\check{U}'(\cdot)) \\ \vdots \end{pmatrix}. \quad (2.17)$$

Finally, recalling $\bar{U}^{(l)}(t_0^-) = \bar{u}^{(l)}(t_0)$ for $l = 0, \dots, \lfloor \frac{k}{2} \rfloor$ and noting (2.15), we have

$$\begin{aligned} \tilde{N}_{i,n}(\check{U}(\cdot)) &= \tilde{N}_{i,n}(\bar{U}(\cdot)), \quad \check{U}^{(\ell_{[i]})}(t_{n-1}^+) = \bar{U}^{(\ell_{[i]})}(t_{n-1}^-), \quad \text{and} \\ \tilde{N}_{j,n}(\check{U}'(\cdot)) &= \tilde{N}_{j,n}(\bar{M} M^{-1} g(\cdot) - \bar{A} \bar{U}(\cdot)) = \tilde{N}_{j,n}(\bar{M} M^{-1} g(\cdot)) - \bar{A} \tilde{N}_{j,n}(\bar{U}(\cdot)), \end{aligned} \quad (2.18)$$

where we also used that \bar{A} is time-independent. From this and (2.17) we easily conclude that $\bar{g}_{i,n}^{\text{VTD}} = \tilde{N}_{i,n}(\bar{U}(\cdot))$, $i = 1, \dots, r+1$, solves the equation system (2.13) as desired. \square

Corollary 2.16 (Runge–Kutta-like formulation)

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Moreover, let $\{\hat{N}_i^*\}$ be specified by (2.10), $\ell_{[i]} = \min\{i-1, \lfloor \frac{k}{2} \rfloor\}$, and $\ell = \lfloor \frac{k}{2} \rfloor$. Furthermore, suppose that f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable and that $\bar{U} \in Y_r$ solves $\mathbf{VTD}_k^r(\bar{M}M^{-1}g)$ with g fulfilling Assumption 2.1. Then, \bar{U} satisfies (2.7) with $\bar{g}_{i,n}^{\text{RKl}} = \bar{g}_{i,n}^{\text{VTD}} = \tilde{N}_{i,n}(\bar{U}(\cdot))$, $i = 1, \dots, r+1$, where

$$A_n^{\text{RKl}} = A_n^{\text{VTD}} \in \mathbb{R}^{(r+1) \times (r+1)} \quad \text{and} \quad B_n^{\text{RKl}} = B_n^{\text{VTD}} := \begin{pmatrix} I_{\lfloor \frac{k}{2} \rfloor, r+1} \\ e_{r+1}^T \end{pmatrix} A_n^{\text{VTD}} \in \mathbb{R}^{(\ell+1) \times (r+1)}$$

with A_n^{VTD} as defined in Proposition 2.15. In this spirit \bar{U} can be viewed as solution of an $(r+1, \lfloor \frac{k}{2} \rfloor)$ -**RKl**($\bar{M}M^{-1}g$) scheme.

Proof. The vectors $\bar{g}_{i,n}^{\text{VTD}}$ in Proposition 2.15 represent the evaluation of the linear operators $\tilde{N}_{i,n}$ for \bar{U} . Therefore, (2.13) could be interpreted as generalized stage equations for the “internal stages” $\bar{g}_{i,n}^{\text{RKl}} = \bar{g}_{i,n}^{\text{VTD}} = \tilde{N}_{i,n}(\bar{U}(\cdot))$, $i = 1, \dots, r+1$, of an $(r+1, \lfloor \frac{k}{2} \rfloor)$ -**RKl**($\bar{M}M^{-1}g$) scheme, cf. (2.7b). Here, also recall that $\bar{U}^{(\ell_{[i]})}(t_{n-1}^-) = \tilde{N}_{i,n}^*(\bar{U}^{(\ell_{[i]})}(t_{n-1}^-))$ for all i .

Additionally noting that $\bar{g}_{i,n}^{\text{VTD}} = \tilde{N}_{i,n}(\bar{U}(\cdot)) = \bar{U}^{(\ell_{[i]})}(t_n^-)$ for $i = 1, \dots, \lfloor \frac{k}{2} \rfloor$ and $i = r+1$ with $\ell_{[i]} = \min\{i-1, \lfloor \frac{k}{2} \rfloor\}$, the first $\lfloor \frac{k}{2} \rfloor$ rows and the last row of (2.13) give that

$$\begin{pmatrix} \bar{U}(t_n^-) \\ \vdots \\ \bar{U}(\lfloor \frac{k}{2} \rfloor)(t_n^-) \end{pmatrix} = \begin{pmatrix} \bar{U}(t_{n-1}^-) \\ \vdots \\ \bar{U}(\lfloor \frac{k}{2} \rfloor)(t_{n-1}^-) \end{pmatrix} + \frac{\tau_n}{2} \underbrace{\begin{pmatrix} I_{\lfloor \frac{k}{2} \rfloor, r+1} \\ e_{r+1}^T \end{pmatrix} A_n^{\text{VTD}}}_{=: B_n^{\text{VTD}}} \otimes I_{d,d} \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(\bar{M}M^{-1}g(\cdot)) - \bar{A} \bar{g}_{j,n}^{\text{VTD}} \\ \vdots \end{pmatrix}, \quad (2.19)$$

which could be seen as some generalized iteration equation as occurring in Runge–Kutta-like methods, cf. (2.7a). \square

Remark 2.17

For clarity, $I_{\lfloor \frac{k}{2} \rfloor, r+1} \in \mathbb{R}^{\lfloor \frac{k}{2} \rfloor \times (r+1)}$ denotes the generalized identity matrix in $\mathbb{R}^{\lfloor \frac{k}{2} \rfloor \times (r+1)}$, i.e., $(I_{\lfloor \frac{k}{2} \rfloor, r+1})_{ij} = \delta_{i,j}$ for all $i = 1, \dots, \lfloor \frac{k}{2} \rfloor$, $j = 1, \dots, r+1$, and e_{r+1} is the $(r+1)$ th standard unit vector in \mathbb{R}^{r+1} . So, B_n^{VTD} only contains the first $\lfloor \frac{k}{2} \rfloor$ rows and the last row of A_n^{VTD} . \clubsuit

2.2.1 Block structure of A^{VTD}

Taking a closer look on its definition (2.14), we note that the matrix $A^{\text{VTD}} \in \mathbb{R}^{(r+1) \times (r+1)}$ has a special block structure. Indeed, it holds

$$A^{\text{VTD}} = \begin{pmatrix} \boxed{B_{11}} & \boxed{B_{12}} \\ 0 & \boxed{B_{22}} \end{pmatrix} \quad (2.20)$$

with $B_{11} \in \mathbb{R}^{\lfloor \frac{k}{2} \rfloor \times \lfloor \frac{k}{2} \rfloor}$, $B_{12} \in \mathbb{R}^{\lfloor \frac{k}{2} \rfloor \times (r+1-\lfloor \frac{k}{2} \rfloor)}$, and $B_{22} \in \mathbb{R}^{(r+1-\lfloor \frac{k}{2} \rfloor) \times (r+1-\lfloor \frac{k}{2} \rfloor)}$, where B_{11} is an upper triangular matrix. Obviously, sums of matrices with this particular structure also have this structure. Moreover, if the inverse matrix exists, then it also has this structure.

We shall now investigate the matrix blocks B_{11} and B_{22} in more detail.

Lemma 2.18

It holds $(A^{\text{VTD}})_{ij} = \frac{2 \cdot (-2)^{j-i}}{(j-i+1)!}$ for $1 \leq i \leq j \leq \lfloor \frac{k}{2} \rfloor$. This completely determines the upper left (triangular) matrix block B_{11} of A^{VTD} , see (2.20). Especially, note that B_{11} is 2 on the main diagonal.

Proof. Let $1 \leq i \leq j \leq \lfloor \frac{k}{2} \rfloor$ throughout the whole proof. Then, we have

$$(A^{\text{VTD}})_{ij} = \hat{N}_i^* \left(\int_{-1}^1 \hat{B}_j^{(i-1)}(\hat{s}) d\hat{s} \right) = \int_{-1}^1 \hat{B}_j^{(i-1)}(\hat{s}) d\hat{s}.$$

The occurring basis functions \hat{B}_j are those associated to the function and derivative values at 1^- . Therefore, as we already noted earlier, it holds $\hat{B}_j(\hat{t}) = \frac{(-1)^{j-1}}{(j-1)!} (1 - \hat{t})^{j-1}$. From this it is easy to verify that $\hat{B}_j^{(i-1)}(\hat{t}) = \frac{(-1)^{j-i}}{(j-i)!} (1 - \hat{t})^{j-i}$. Hence, we find

$$(A^{\text{VTD}})_{ij} = \frac{(-1)^{j-i}}{(j-i)!} \int_{-1}^1 (1 - \hat{s})^{j-i} d\hat{s} = \frac{(-1)^{j-i}}{(j-i)!} \left[\frac{-1}{j-i+1} (1 - \hat{s})^{j-i+1} \right]_{\hat{s}=-1}^1 = \frac{2 \cdot (-2)^{j-i}}{(j-i+1)!},$$

which finishes the proof. \square

Lemma 2.19

The lower right matrix block B_{22} of A^{VTD} , see (2.20), equates to the Runge–Kutta matrix of the $(r+1 - \lfloor \frac{k}{2} \rfloor)$ -stage Radau IIA method if k is even or the $(r+1 - \lfloor \frac{k}{2} \rfloor)$ -stage Lobatto IIIA method if k is odd, respectively. However, note that the Runge–Kutta matrices are typically defined with respect to the reference interval $[0, 1]$, whereas A^{VTD} is defined with respect to $[-1, 1]$, which causes a transformation factor of 2. Concretely, B_{22} is twice the corresponding Runge–Kutta matrix as given e.g. in [38, pp. 74–75].

Proof. The functionals \hat{N}_i^* , $i = \lfloor \frac{k}{2} \rfloor + 1, \dots, r+1$, are exactly the function evaluations at the quadrature points of $Q_{k-2\lfloor \frac{k}{2} \rfloor}^{r-\lfloor \frac{k}{2} \rfloor}$, which is right-sided Gauss–Radau for k even or Gauss–Lobatto for k odd, respectively. The implicit Runge–Kutta methods that are equivalent to the collocation methods based on these quadrature points are the $(r+1 - \lfloor \frac{k}{2} \rfloor)$ -stage Radau IIA method (k even) and the $(r+1 - \lfloor \frac{k}{2} \rfloor)$ -stage Lobatto IIIA method (k odd), respectively, see [37, Theorem II.7.7, p. 212] and [38, pp. 72–77]. Since disregarding transformation the coefficients of the matrix block B_{22} of A^{VTD} are defined in the same way, we are done. \square

Remark 2.20

Lemma 2.19 already indicates some connection between **VTD** methods and Runge–Kutta methods of type Radau IIA or Lobatto IIIA. In fact, taking a closer look at the derivation of the Runge–Kutta-like formulation, it can be seen that $Q_0^r\text{-VTD}_0^r$ is equivalent to the $(r+1)$ -stage Radau IIA method, whereas $Q_1^r\text{-VTD}_1^r$ is equivalent to the $(r+1)$ -stage Lobatto IIIA method. This was already exemplarily shown in [46, p. 8, p. 13].

The connection between Radau IIA and numerically integrated dG methods with quadrature at the right-sided Gauss–Radau nodes was earlier proven in [45, Lemma 2.3], also see [26, Lemma 69.11]. Moreover, an equivalence of numerically integrated cGP methods with quadrature at the Gauss–Legendre nodes and certain Kuntzmann–Butcher methods has been observed in [26, Lemma 70.5]. However, note that we use cGP methods together with quadrature at the Gauss–Lobatto nodes. ♣

2.2.2 Eigenvalue structure of A^{VTD}

In Section 2.1 we noticed that for Runge–Kutta-like methods a good knowledge of various properties of the method matrix A^{RKI} is needed to answer questions on the solvability and stability of the discrete method. For this reason and since **VTD** methods can be viewed as Runge–Kutta-like methods (as proven in Corollary 2.16), the eigenvalue structure of A^{VTD} shall be studied in detail in this subsection.

Recalling the special block structure of A^{VTD} , see Subsection 2.2.1, $\lfloor \frac{k}{2} \rfloor$ eigenvalues can be directly read from the first $\lfloor \frac{k}{2} \rfloor$ entries on the main diagonal. Because of Lemma 2.18, we thus have that $\lambda = 2$ is an eigenvalue of A^{VTD} with algebraic multiplicity greater than or equal to $\lfloor \frac{k}{2} \rfloor$.

The remaining eigenvalues are those of the lower right matrix block B_{22} of A^{VTD} , cf. (2.20), which can be nicely interpreted due to Lemma 2.19. It is easy to verify that the first row of B_{22} is zero if k is odd. In this case $\lambda = 0$ is a simple eigenvalue of A^{VTD} since all further eigenvalues have a real part greater than zero as we shall see below. To verify the latter statement, the following auxiliary lemma is quite useful.

Lemma 2.21

Let $\Lambda \in \mathbb{R}^{s \times s}$ and $D = \text{diag}(d_1, \dots, d_s)$ with $d_i > 0$ for all $i = 1, \dots, s$. Then, for every eigenvalue λ_Λ of Λ it holds

$$\text{Re}(\lambda_\Lambda) \geq \left(\max_{i=1, \dots, s} d_i \right)^{-1} \lambda_{\min} \left(\frac{1}{2} (D\Lambda + \Lambda^T D) \right),$$

where $\lambda_{\min}(B)$ denotes the smallest eigenvalue of the symmetric matrix $B \in \mathbb{R}^{s \times s}$.

Proof. Since similar matrices share their eigenvalues, every eigenvalue λ_Λ of Λ equals to an eigenvalue $\lambda_{D^{1/2}\Lambda D^{-1/2}}$ of $D^{1/2}\Lambda D^{-1/2}$. Because of [31, Theorem C1], we thus have

$$\text{Re}(\lambda_\Lambda) = \text{Re}(\lambda_{D^{1/2}\Lambda D^{-1/2}}) \geq \lambda_{\min} \left(\frac{1}{2} (D^{1/2}\Lambda D^{-1/2} + D^{-1/2}\Lambda^T D^{1/2}) \right).$$

Since for symmetric matrices the smallest eigenvalue can be calculated by minimizing the Rayleigh quotient (see e.g. [31, p. 32]), it follows

$$\begin{aligned} \lambda_{\min} \left(\frac{1}{2} (D^{1/2}\Lambda D^{-1/2} + D^{-1/2}\Lambda^T D^{1/2}) \right) &= \min_{x \in \mathbb{R}^s, x \neq 0} \frac{x^T \left(\frac{1}{2} (D^{1/2}\Lambda D^{-1/2} + D^{-1/2}\Lambda^T D^{1/2}) \right) x}{x^T x} \\ &= \min_{x \in \mathbb{R}^s, x \neq 0} \frac{x^T D^{1/2}\Lambda D^{-1/2} x}{x^T x} = \min_{y \in \mathbb{R}^s, y \neq 0} \frac{y^T D \Lambda y}{y^T D y}, \end{aligned}$$

where $y = D^{-1/2}x$. We further gain

$$\begin{aligned} \min_{y \in \mathbb{R}^s, y \neq 0} \frac{y^T D \Lambda y}{y^T D y} &= \min_{y \in \mathbb{R}^s, y \neq 0} \frac{y^T \left(\frac{1}{2} (D \Lambda + \Lambda^T D) \right) y}{y^T y} \frac{y^T y}{y^T D y} \\ &\geq \min_{y \in \mathbb{R}^s, y \neq 0} \frac{y^T \left(\frac{1}{2} (D \Lambda + \Lambda^T D) \right) y}{y^T y} \min_{y \in \mathbb{R}^s, y \neq 0} \frac{y^T y}{y^T D y}. \end{aligned}$$

The first minimum on the right-hand side just is $\lambda_{\min} \left(\frac{1}{2} (D \Lambda + \Lambda^T D) \right)$. The second minimum can be bounded from below as follows

$$\min_{y \in \mathbb{R}^s, y \neq 0} \frac{y^T y}{y^T D y} \geq \min_{y \in \mathbb{R}^s, y \neq 0} \frac{y^T y}{\left(\max_{i=1, \dots, s} d_i \right) y^T y} = \left(\max_{i=1, \dots, s} d_i \right)^{-1}.$$

Combining the above estimates, we easily complete the proof. \square

Lemma 2.22

Let $B_{22} \in \mathbb{R}^{s \times s}$ with $s = r + 1 - \lfloor \frac{k}{2} \rfloor$ denote the lower right matrix block of A^{VTD} , see (2.20). Moreover, let $\tilde{t}_i \in [-1, 1]$, $i = 1, \dots, s$, denote the quadrature points of $Q_{k-2\lfloor k/2 \rfloor}^{r-\lfloor k/2 \rfloor}$ and \tilde{b}_i the associated weights. Set $D = \text{diag}(d_1, \dots, d_s)$ with

$$d_i = \tilde{b}_i (1 + \tilde{t}_i)^{-(\sigma_k + 1)} \quad \text{for } i = \sigma_k + 1, \dots, s \quad \text{and} \quad d_1 = 1 \quad \text{if } k \text{ is odd}, \quad (2.21)$$

where $\sigma_k := k - 2 \lfloor \frac{k}{2} \rfloor = \begin{cases} 1, & \text{if } k \text{ is odd,} \\ 0, & \text{if } k \text{ is even.} \end{cases}$ Then, it holds

$$x^T \left(\frac{1}{2} (D B_{22} + B_{22}^T D) \right) x > 0 \quad \text{for all } x \in \mathbb{R}^s \setminus \{0\} \text{ with } x_1 = 0 \text{ if } k \text{ is odd.}$$

Proof. First of all, we note that $Q_{k-2\lfloor k/2 \rfloor}^{r-\lfloor k/2 \rfloor}$ is the Gauss–Radau quadrature if k is even or the Gauss–Lobatto quadrature if k is odd, respectively. Therefore, for odd k we especially have that $\tilde{t}_1 = -1$, which makes clear why another definition for d_1 is needed in this case. Moreover, for these quadrature rules it is well-known that the weights are positive and are given as integrals over the associated basis functions. So, we have $\tilde{b}_i = \int_{-1}^1 \hat{B}_{\lfloor k/2 \rfloor + i}^{(\lfloor k/2 \rfloor)}(\hat{s}) d\hat{s} > 0$, which also implies that $d_i > 0$, $i = 1, \dots, s$.

Let $x \in \mathbb{R}^s$ be arbitrary and assume that $x_1 = 0$ if k is odd. We define a polynomial $\hat{p}' \in P_{s-1}((-1, 1])$ by

$$\hat{p}'(\hat{t}) = \sum_{j=1}^s x_j \hat{B}_{\lfloor k/2 \rfloor + j}^{(\lfloor k/2 \rfloor)}(\hat{t}) \quad \text{for all } \hat{t} \in (-1, 1],$$

i.e., \hat{p}' is the Gauss–Radau (k even) or Gauss–Lobatto (k odd) interpolant satisfying

$$x_i = \hat{p}'(\tilde{t}_i) = \hat{N}_{\lfloor k/2 \rfloor + i}^*(\hat{p}') \quad \text{for all } i = 1, \dots, s.$$

Obviously, $\hat{p}'(-1) = \hat{p}'(\tilde{t}_1) = x_1 = 0$ if k is odd.

Further, we denote by \hat{p} the particular antiderivative of \hat{p}' satisfying $\hat{p}(-1) = 0$. Then, $\hat{p} \in P_s((-1, 1])$ allows the following representations

$$\hat{p}(\hat{t}) = \int_{-1}^{\hat{t}} \hat{p}'(\hat{s}) d\hat{s} = (1 + \hat{t})^{\sigma_k + 1} \hat{q}(\hat{t}) \quad \text{with } \hat{q} \in P_{s-1-\sigma_k}((-1, 1]).$$

Of course, we then have

$$\hat{p}'(\hat{t}) = (\sigma_k + 1)(1 + \hat{t})^{\sigma_k} \hat{q}(\hat{t}) + (1 + \hat{t})^{\sigma_k+1} \hat{q}'(\hat{t}).$$

Now, for $i = 1, \dots, s$ we obtain

$$\begin{aligned} (B_{22}x)_i &= \sum_{j=1}^s (A^{\text{VTD}})_{[\frac{k}{2}] + i, [\frac{k}{2}] + j} x_j = \sum_{j=1}^s \left(\int_{-1}^{\tilde{t}_i} \hat{B}_{[\frac{k}{2}] + j}^{([k/2])}(\hat{s}) d\hat{s} \right) x_j \\ &= \int_{-1}^{\tilde{t}_i} \left(\sum_{j=1}^s x_j \hat{B}_{[\frac{k}{2}] + j}^{([k/2])}(\hat{s}) \right) d\hat{s} = \int_{-1}^{\tilde{t}_i} \hat{p}'(\hat{s}) d\hat{s} = \hat{p}(\tilde{t}_i). \end{aligned}$$

Using this, it is easy to verify that

$$x^T \left(\frac{1}{2} (DB_{22} + B_{22}^T D) \right) x = x^T (DB_{22}) x = \sum_{i=1}^s \hat{p}'(\tilde{t}_i) d_i \hat{p}(\tilde{t}_i) = \sum_{i=1+\sigma_k}^s \hat{p}'(\tilde{t}_i) d_i \hat{p}(\tilde{t}_i),$$

where for the last identity we have exploited that $\hat{p}'(\tilde{t}_1) = x_1 = 0$ if k is odd. Recalling the definition of d_i and the alternative representations of \hat{p} and \hat{p}' via \hat{q} , we further conclude

$$\begin{aligned} &\sum_{i=1+\sigma_k}^s \hat{p}'(\tilde{t}_i) d_i \hat{p}(\tilde{t}_i) \\ &= \sum_{i=1+\sigma_k}^s \tilde{b}_i (1 + \tilde{t}_i)^{-(\sigma_k+1)} \left((\sigma_k + 1)(1 + \tilde{t}_i)^{2\sigma_k+1} \hat{q}^2(\tilde{t}_i) + (1 + \tilde{t}_i)^{2(\sigma_k+1)} \hat{q}'(\tilde{t}_i) \hat{q}(\tilde{t}_i) \right) \\ &= \sum_{i=1}^s \tilde{b}_i \left((\sigma_k + 1)(1 + \tilde{t}_i)^{\sigma_k} \hat{q}^2(\tilde{t}_i) + (1 + \tilde{t}_i)^{\sigma_k+1} \hat{q}'(\tilde{t}_i) \hat{q}(\tilde{t}_i) \right). \end{aligned}$$

Here, for the last step we have used that the summand for $i = 1$ is zero if k is odd since in this case the factor $(1 + \tilde{t}_1)^{\sigma_k} = (1 + (-1))^1$ is vanishing.

The function $\hat{t} \mapsto ((\sigma_k + 1)(1 + \hat{t})^{\sigma_k} \hat{q}^2(\hat{t}) + (1 + \hat{t})^{\sigma_k+1} \hat{q}'(\hat{t}) \hat{q}(\hat{t}))$ is a polynomial of maximal degree $(2s - 2 - \sigma_k) = (2r - k)$ and, thus, is exactly integrated by $Q_{k-2[k/2]}^{r-[k/2]}$. Therefore, combining the above identities, we gain

$$\begin{aligned} x^T \left(\frac{1}{2} (DB_{22} + B_{22}^T D) \right) x &= \int_{-1}^1 \left((\sigma_k + 1)(1 + \hat{t})^{\sigma_k} \hat{q}^2(\hat{t}) + (1 + \hat{t})^{\sigma_k+1} \hat{q}'(\hat{t}) \hat{q}(\hat{t}) \right) d\hat{t} \\ &= \frac{1}{2} \int_{-1}^1 (\sigma_k + 1)(1 + \hat{t})^{\sigma_k} \hat{q}^2(\hat{t}) d\hat{t} + 2^{\sigma_k} \hat{q}^2(1) \geq 0, \end{aligned} \quad (2.22)$$

where we used integration by parts to rewrite the integral over the second summand. Indeed, it follows

$$\begin{aligned} \int_{-1}^1 (1 + \hat{t})^{\sigma_k+1} \hat{q}'(\hat{t}) \hat{q}(\hat{t}) d\hat{t} &= \int_{-1}^1 (1 + \hat{t})^{\sigma_k+1} \frac{1}{2} (\hat{q}^2)'(\hat{t}) d\hat{t} \\ &= -\frac{1}{2} \int_{-1}^1 (\sigma_k + 1)(1 + \hat{t})^{\sigma_k} \hat{q}^2(\hat{t}) d\hat{t} + \frac{1}{2} \left[(1 + \hat{t})^{\sigma_k+1} \hat{q}^2(\hat{t}) \right]_{\hat{t}=-1}^1. \end{aligned}$$

It only remains to prove that the term (2.22) is zero only if $x = 0$. Now, if the expression in (2.22) vanishes, then $\hat{q} \equiv 0$. But this directly implies $\hat{p} \equiv 0$ and therefore also $\hat{p}' \equiv 0$. Hence, $x_i = \hat{p}'(\tilde{t}_i) = 0$ for all $i = 1, \dots, s$ and we are done. \square

Bringing together the above results, we obtain the following statement.

Corollary 2.23

It holds $\sigma(A^{\text{VTD}}) \subset \mathbb{C}_0^+$. Moreover, $\lambda = 0$ is an eigenvalue of A^{VTD} if and only if k is odd. In this case $\lambda = 0$ is a simple eigenvalue. Thus, for even k all eigenvalues have a positive real part while for odd k zero is a simple eigenvalue and all further eigenvalues have a positive real part.

Proof. We have already seen that the first $\lfloor \frac{k}{2} \rfloor$ eigenvalues of A^{VTD} are $\lambda = 2$ and, thus, obviously have a positive real part. Moreover, we noted that $\lambda = 0$ is an eigenvalue of A^{VTD} if k is odd since then the first row of the matrix block B_{22} of A^{VTD} , see (2.20), vanishes.

Therefore, bringing to mind the structure of A^{VTD} , the remaining eigenvalues are just those of $\tilde{B}_{22} = (B_{22})_{i,j=1+\sigma_k, \dots, s}$ where $s = r + 1 - \lfloor \frac{k}{2} \rfloor$ and $\sigma_k = k - 2\lfloor \frac{k}{2} \rfloor$. But, setting $\tilde{D} = \text{diag}(d_{\sigma_k+1}, \dots, d_s)$ with $d_i > 0$ as defined in (2.21), we conclude from Lemma 2.22 that $(\frac{1}{2}(\tilde{D}\tilde{B}_{22} + \tilde{B}_{22}^T\tilde{D}))$ is positive definite and, thus, only has positive eigenvalues. Hence, according to Lemma 2.21, all eigenvalues of \tilde{B}_{22} have positive real part. \square

2.2.3 Solvability and stability

Knowing the eigenvalue structure of the matrix A^{VTD} , we are now able to assess the solvability of (2.13) and the stability of the Runge–Kutta-like formulation of the **VTD** method.

Proposition 2.24

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $\mu \geq \mu[-A]$. Moreover, let $\{\hat{N}_i^\}$ be specified by (2.10), $\ell_{[i]} = \min\{i - 1, \lfloor \frac{k}{2} \rfloor\}$, and $\ell = \lfloor \frac{k}{2} \rfloor$. Furthermore, suppose that f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable.*

*Then, the Runge–Kutta-like formulation associated by Corollary 2.16 to $\mathbf{VTD}_k^r(\overline{M}M^{-1}g)$ with g fulfilling Assumption 2.1 is uniquely solvable for time step lengths $\tau_n \in (0, \bar{\tau}]$ with $\bar{\tau} > 0$ sufficiently small. If $\mu \leq 0$, the unique solvability holds without restriction on the (maximal) mesh interval length. Moreover, in either case the formulation is *ASI-stable* and *AS-stable*.*

Proof. From Corollary 2.23 we know that $\sigma(A^{\text{VTD}}) \subset \mathbb{C}_0^+$ and that $\lambda = 0$ is at most a simple eigenvalue of A^{VTD} . Therefore, the solvability follows from Lemma 2.6 and Corollary 2.9, also note Remark 2.7. Furthermore, the *ASI-stability* holds due to Lemma 2.12. Since $B^{\text{VTD}} = \begin{pmatrix} I_{\lfloor \frac{k}{2} \rfloor, r+1} \\ e_{r+1}^T \end{pmatrix} A^{\text{VTD}}$, we also gain *AS-stability* from Lemma 2.13. \square

We already know that the **VTD** methods are *A-stable* (cf. Remark 1.1 and Remark 1.39). However, their special construction involving collocation conditions yields in combination with the special structure of their associated Runge–Kutta-like formulation some more quite interesting consequences with respect to stability.

Lemma 2.25

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and suppose that $\bar{U} \in Y_r$ is the $\mathbf{VTD}_k^r(0)$ approximation to the solution of Dahlquist's stability equation (1.29). Then, there is a (stability) function R (defined on \mathbb{C} a.e.) such that

$$\bar{U}^{(l)}(t_n^-) = R\left(\frac{\tau_n}{2}\lambda\right)\bar{U}^{(l)}(t_{n-1}^-)$$

for all $l = 0, \dots, \lfloor \frac{k}{2} \rfloor$. Further, we have that R is just the $(r - \lfloor \frac{k}{2} \rfloor, r - \lfloor \frac{k-1}{2} \rfloor)$ Padé approximation of $\exp(2z)$, so especially satisfies

$$|R(z)| \leq 1 \quad \text{for all } z \in \mathbb{C}^- = \{z \in \mathbb{C} : \operatorname{Re}(z) \leq 0\}.$$

Proof. According to Corollary 1.42, we have that $\bar{U}^{(l)}$, $l = 0, \dots, \lfloor \frac{k}{2} \rfloor$, solve $\mathbf{VTD}_{k-2l}^{r-l}(0)$ when appropriate initial conditions are used. Because of Remark 1.39, we further know that all these methods share their stability properties with $\text{dG}(r - \lfloor \frac{k}{2} \rfloor)$ if k is even or $\text{cGP}(r - \lfloor \frac{k}{2} \rfloor)$ if k is odd, respectively. Hence, they especially have the same stability function R , which yields the first statement. Moreover, $|R(z)| \leq 1$ on \mathbb{C}^- immediately follows from the fact that dG methods as well as cGP methods are A -stable.

Furthermore, because of Remark 2.20, the stability functions of $\text{dG}(s)$ (with $s \geq 0$) and $\text{cGP}(s)$ (with $s \geq 1$) are essentially the same as for the $(s+1)$ -stage Radau IIA method and for the $(s+1)$ -stage Lobatto IIIA method, respectively, thus certain Padé approximations, cf. [38, Table IV.5.13, p. 77]. Therefore, we find that R is the $(r - \lfloor \frac{k}{2} \rfloor, r - \lfloor \frac{k-1}{2} \rfloor)$ Padé approximation of $\exp(2z)$. \square

As for Runge–Kutta methods it would be nice to have some representation of the (stability) function R in terms of the method parameters, i.e., in terms of A^{VTD} . This is provided by the following lemma.

Lemma 2.26

Let $0 \leq i \leq \lfloor \frac{k}{2} \rfloor$. The (stability) function R of Lemma 2.25 provides the following representations

$$R(z) = e_{\sigma[i]}^T (I_{r+1, r+1} - A^{\text{VTD}} z)^{-1} \left(\sum_{j=i+1}^{r+1} z^{\min\{j-1, \lfloor \frac{k}{2} \rfloor\} - i} e_j \right),$$

where e_j denotes the j th standard unit vector in \mathbb{R}^{r+1} and $\sigma[i] = (i+1) + \delta_{i, \lfloor \frac{k}{2} \rfloor} (r-i)$. This also implies that for all i the expressions on the right-hand side are the same.

Proof. Applied to Dahlquist's stability equation (1.29), i.e., to problem (1.21) with $d = 1$, $M = 1 = \bar{M}$, $A = -\lambda = \bar{A}$, and $f = 0$, the “stage” equations (2.13) of $\mathbf{VTD}_k^r(0)$ simply read

$$\begin{pmatrix} \vdots \\ \bar{g}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \bar{U}^{(\ell[i])}(t_{n-1}^-) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} A_n^{\text{VTD}} \begin{pmatrix} \vdots \\ \lambda \bar{g}_{j,n}^{\text{VTD}} \\ \vdots \end{pmatrix}.$$

Rewriting this and recalling that $\bar{g}_{i,n}^{\text{VTD}} = \tilde{N}_{i,n}(\bar{U}(\cdot)) = \bar{U}^{(\ell_{[i]})}(t_n^-)$ for $i = 1, \dots, \lfloor \frac{k}{2} \rfloor$ and $i = r+1$ with $\ell_{[i]} = \min\{i-1, \lfloor \frac{k}{2} \rfloor\}$, we obtain

$$\begin{pmatrix} \bar{U}(t_n^-) \\ \vdots \\ \bar{U}(\lfloor \frac{k}{2} \rfloor)(t_n^-) \end{pmatrix} = \begin{pmatrix} I_{\lfloor k/2 \rfloor, r+1} \\ e_{r+1}^T \end{pmatrix} \begin{pmatrix} \vdots \\ \bar{g}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} = \begin{pmatrix} I_{\lfloor k/2 \rfloor, r+1} \\ e_{r+1}^T \end{pmatrix} (I_{r+1, r+1} - \frac{\tau_n}{2} A_n^{\text{VTD}} \lambda)^{-1} \begin{pmatrix} \vdots \\ \bar{U}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix}.$$

This identity already looks quite promising. However, for $\lambda \neq 0$ the various derivatives of \bar{U} at t_{n-1}^- on the right-hand side are coupled, whereas for the desired statement we need for each $i = 0, \dots, \lfloor \frac{k}{2} \rfloor$ an expression that links $\bar{U}^{(i)}(t_n^-)$ to $\bar{U}^{(i)}(t_{n-1}^-)$ only.

But, from the collocation conditions at t_{n-1}^- for $n > 1$, cf. (1.22b), or the definition of the discrete initial values by $\bar{U}^{(i)}(t_0^-) = \bar{u}^{(i)}(t_0^+)$, respectively, we have that

$$\bar{U}^{(i)}(t_{n-1}^-) = \lambda \bar{U}^{(i-1)}(t_{n-1}^-) \quad \text{for all } 1 \leq i \leq \lfloor \frac{k}{2} \rfloor.$$

Therefore, we find for $0 \leq i \leq \lfloor \frac{k}{2} \rfloor$ that

$$\begin{aligned} \bar{U}^{(i)}(t_n^-) &= e_{\sigma_{[i]}}^T (I_{r+1, r+1} - \frac{\tau_n}{2} A_n^{\text{VTD}} \lambda)^{-1} \left(\sum_{j=1}^{r+1} \bar{U}^{(\ell_{[j]})}(t_{n-1}^-) e_j \right) \\ &= e_{\sigma_{[i]}}^T (I_{r+1, r+1} - \frac{\tau_n}{2} A_n^{\text{VTD}} \lambda)^{-1} \left(\sum_{j=1}^i \bar{U}^{(j-1)}(t_{n-1}^-) e_j + \sum_{j=i+1}^{r+1} \lambda^{\ell_{[j]}-i} \bar{U}^{(i)}(t_{n-1}^-) e_j \right). \end{aligned}$$

Further, exploiting that $A_n^{\text{VTD}} = S_n^{-1} A^{\text{VTD}} S_n$ and taking advantage of the special structure of A^{VTD} , see (2.20), which is transferred to $(I_{r+1, r+1} - \frac{\tau_n}{2} A^{\text{VTD}} \lambda)^{-1}$, we gain

$$\begin{aligned} \bar{U}^{(i)}(t_n^-) &= e_{\sigma_{[i]}}^T S_n^{-1} (I_{r+1, r+1} - \frac{\tau_n}{2} A^{\text{VTD}} \lambda)^{-1} S_n \left(\sum_{j=i+1}^{r+1} \lambda^{\ell_{[j]}-i} \bar{U}^{(i)}(t_{n-1}^-) e_j \right) \\ &= e_{\sigma_{[i]}}^T (I_{r+1, r+1} - A^{\text{VTD}} (\frac{\tau_n}{2} \lambda))^{-1} \left(\sum_{j=i+1}^{r+1} (\frac{\tau_n}{2} \lambda)^{\ell_{[j]}-i} e_j \right) \bar{U}^{(i)}(t_{n-1}^-). \end{aligned}$$

From this, we easily complete the proof. \square

Remark 2.27

Starting with the generalized iteration equation (2.19) for $\mathbf{VTD}_k^r(0)$ applied to Dahlquist's stability problem (1.29) and using similar arguments as in the proof of Lemma 2.26, it also can be shown for $0 \leq i \leq \lfloor \frac{k}{2} \rfloor$ that

$$R(z) = 1 + z e_{\sigma_{[i]}}^T A^{\text{VTD}} (I_{r+1, r+1} - A^{\text{VTD}} z)^{-1} \left(\sum_{j=i+1}^{r+1} z^{\min\{j-1, \lfloor \frac{k}{2} \rfloor\}-i} e_j \right),$$

where the notation of Lemma 2.26 is reused.

This representation of the stability function is quite similar to that for Runge–Kutta methods, cf. [38, Proposition IV.3.1, p. 40]. Note that due to $B^{\text{VTD}} = \begin{pmatrix} I_{\lfloor k/2 \rfloor, r+1} \\ e_{r+1}^T \end{pmatrix} A^{\text{VTD}}$ we have that $e_{\sigma_{[i]}}^T A^{\text{VTD}} = e_{i+1}^T B^{\text{VTD}}$ is just the $(i+1)$ th row of B^{VTD} . Moreover, if $i = \lfloor \frac{k}{2} \rfloor = 0$, then $\sum_{j=i+1}^{r+1} z^{\min\{j-1, \lfloor \frac{k}{2} \rfloor\}-i} e_j$ simply is the all-ones vector in \mathbb{R}^{r+1} . \clubsuit

2.3 (Stiff) Error analysis

In this section, we derive error estimates for **VTD** methods also in the case of stiff systems of ordinary differential equations. To be exact, throughout the whole section, we consider the $\mathbf{VTD}_k^r(\overline{M}M^{-1}g)$ method with g fulfilling Assumption 2.1 where $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and where we suppose that f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable. The scheme is used as approximation for the initial value problem (2.8). Moreover, we let $\mu \geq \mu[-\overline{A}]$, i.e., μ is supposed to satisfy (2.9).

The presented error analysis strongly uses that the considered **VTD** methods can be reformulated in a Runge–Kutta-like way as it has been observed in the previous section, see Corollary 2.16. Therefore, also the notation introduced and specified in Section 2.2 is used further, especially the linear functionals (2.10) and their local versions (2.11).

Let the operator $\mathcal{P}_n^{\text{VTD}} : C^{\lfloor \frac{k}{2} \rfloor}(\overline{I}_n) \rightarrow P_r(I_n)$ be defined by

$$\mathcal{P}_n^{\text{VTD}}v = \sum_{i=1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell[i]} \tilde{N}_{i,n}(v)(\hat{B}_i \circ T_n^{-1}).$$

Since, in view of (2.12), this operator is a projection onto $P_r(I_n)$, it provides for $0 \leq i \leq r+1$ the following approximation error estimates

$$\sup_{t \in I_n} |(v - \mathcal{P}_n^{\text{VTD}}v)^{(i)}(t)| \leq C \left(\frac{\tau_n}{2}\right)^{r+1-i} \sup_{t \in I_n} |v^{(r+1)}(t)| \quad \forall v \in C^{r+1}(\overline{I}_n). \quad (2.23)$$

Moreover, it follows that

$$(\mathcal{P}_n^{\text{VTD}}v)^{(l)} = \sum_{i=l+1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell[i]-l} \tilde{N}_{i,n}(v)(\hat{B}_i^{(l)} \circ T_n^{-1})$$

for all $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$.

2.3.1 Recursion scheme for the global error

According to Corollary 2.16, also see (2.13), the solution \overline{U} of $\mathbf{VTD}_k^r(\overline{M}M^{-1}g)$ with g fulfilling Assumption 2.1 satisfies certain recursion schemes. In preparation for the error analysis we show now that a similar recursion scheme holds for the error too.

Similar to \check{U} (cf. (2.16)) the exact solution \overline{u} of (2.8) satisfies for $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$ and $t \in I_n$

$$\begin{aligned} \overline{u}^{(l)}(t) &= \overline{u}^{(l)}(t_{n-1}^+) + \int_{t_{n-1}}^t (\mathcal{P}_n^{\text{VTD}}\overline{u}')^{(l)}(\tilde{s}) d\tilde{s} + \int_{t_{n-1}}^t (\overline{u}' - \mathcal{P}_n^{\text{VTD}}\overline{u}')^{(l)}(\tilde{s}) d\tilde{s} \\ &= \overline{u}^{(l)}(t_{n-1}^+) + \frac{\tau_n}{2} \sum_{j=l+1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell[j]-l} \tilde{N}_{j,n}(\overline{u}'(\cdot)) \int_{-1}^{T_n^{-1}(t)} \hat{B}_j^{(l)}(\hat{s}) d\hat{s} \\ &\quad + \int_{t_{n-1}}^t (\overline{u}' - \mathcal{P}_n^{\text{VTD}}\overline{u}')^{(l)}(\tilde{s}) d\tilde{s}. \end{aligned} \quad (2.24)$$

Therefore, analogously to (2.17) for \check{U} , we find for \bar{u}

$$\begin{pmatrix} \vdots \\ \tilde{N}_{i,n}(\bar{u}(\cdot)) \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \bar{u}^{(\ell_{[i]})}(t_{n-1}^+) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(\bar{u}'(\cdot)) \\ \vdots \end{pmatrix} + \begin{pmatrix} \vdots \\ \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} \quad (2.25)$$

where $\bar{r}_{i,n}^{\text{VTD}} \in \mathbb{R}^d$, $i = 1, \dots, r+1$, is given by

$$\bar{r}_{i,n}^{\text{VTD}} := \tilde{N}_{i,n}^* \left(\int_{t_{n-1}}^{\cdot} (\bar{u}' - \mathcal{P}_n^{\text{VTD}} \bar{u}')^{(\ell_{[i]})}(\tilde{s}) \, d\tilde{s} \right). \quad (2.26)$$

Now, combining (2.17) and (2.25) as well as using that (2.8) and (2.18) hold, we find for the error $\bar{e} = \bar{u} - \bar{U}$ (and setting $\check{e} = \bar{u} - \check{U}$) that

$$\begin{aligned} \begin{pmatrix} \vdots \\ \tilde{N}_{i,n}(\bar{e}(\cdot)) \\ \vdots \end{pmatrix} &= \begin{pmatrix} \vdots \\ \check{e}^{(\ell_{[i]})}(t_{n-1}^+) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \tilde{N}_{j,n}(\check{e}'(\cdot)) \\ \vdots \end{pmatrix} + \begin{pmatrix} \vdots \\ \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} \\ &= \begin{pmatrix} \vdots \\ \bar{e}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \bar{M}M^{-1}\tilde{N}_{j,n}((f-g)(\cdot)) - \bar{A}\tilde{N}_{j,n}(\bar{e}(\cdot)) \\ \vdots \end{pmatrix} + \begin{pmatrix} \vdots \\ \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix}. \end{aligned}$$

Hence, rewriting this, it follows that

$$\begin{pmatrix} \vdots \\ \tilde{N}_{i,n}(\bar{e}(\cdot)) \\ \vdots \end{pmatrix} = ((I_{r+1,r+1} \otimes I_{d,d}) + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes \bar{A}))^{-1} \begin{pmatrix} \vdots \\ \left(\begin{pmatrix} \vdots \\ \bar{e}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \bar{M}M^{-1}\tilde{N}_{j,n}((f-g)(\cdot)) \\ \vdots \end{pmatrix} + \begin{pmatrix} \vdots \\ \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} \right) \\ \vdots \end{pmatrix}. \quad (2.27)$$

Recalling Remark 2.7, Corollary 2.9, and Corollary 2.23, in order to guarantee the existence of the inverse matrix on the right-hand side, we only need that τ_n is sufficiently small, and that only if $\mu > 0$.

Furthermore, with an argument similar to that in the proof of Corollary 2.16, especially recalling that $\tilde{N}_{i,n}(\bar{e}(\cdot)) = \bar{e}^{(\ell_{[i]})}(t_n^-)$ for $i = 1, \dots, \lfloor \frac{k}{2} \rfloor$ and $i = r+1$ with $\ell_{[i]} = \min\{i-1, \lfloor \frac{k}{2} \rfloor\}$, we find

$$\begin{pmatrix} \bar{e}(t_n^-) \\ \vdots \\ \bar{e}(\lfloor \frac{k}{2} \rfloor)(t_n^-) \end{pmatrix} = \left(\begin{pmatrix} I_{\lfloor \frac{k}{2} \rfloor, r+1} \\ e_{r+1}^T \end{pmatrix} \otimes I_{d,d} \right) ((I_{r+1,r+1} \otimes I_{d,d}) + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes \bar{A}))^{-1} \begin{pmatrix} \vdots \\ \left(\begin{pmatrix} \vdots \\ \bar{e}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix} + \frac{\tau_n}{2} (A_n^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \bar{M}M^{-1}\tilde{N}_{j,n}((f-g)(\cdot)) \\ \vdots \end{pmatrix} + \begin{pmatrix} \vdots \\ \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} \right) \\ \vdots \end{pmatrix}.$$

Using that $A_n^{\text{VTD}} = S_n^{-1} A^{\text{VTD}} S_n$, the right-hand side can be rewritten and split as follows

$$\begin{pmatrix} \bar{e}(t_n^-) \\ \vdots \\ \bar{e}(\lfloor \frac{k}{2} \rfloor)(t_n^-) \end{pmatrix} = \underbrace{\begin{pmatrix} \mathcal{E}_{(i),0} \\ \vdots \\ \mathcal{E}_{(i),\lfloor k/2 \rfloor} \end{pmatrix}}_{\mathcal{E}_{(i)}} + \underbrace{\begin{pmatrix} \mathcal{E}_{(ii),0} \\ \vdots \\ \mathcal{E}_{(ii),\lfloor k/2 \rfloor} \end{pmatrix}}_{\mathcal{E}_{(ii)}} + \underbrace{\begin{pmatrix} \mathcal{E}_{(iii),0} \\ \vdots \\ \mathcal{E}_{(iii),\lfloor k/2 \rfloor} \end{pmatrix}}_{\mathcal{E}_{(iii)}} \quad (2.28)$$

where, reusing the notation of Lemma 2.26, the block components $(l = 0, \dots, \lfloor \frac{k}{2} \rfloor)$ of the block vectors $\mathcal{E}_{(i)}$, $\mathcal{E}_{(ii)}$, and $\mathcal{E}_{(iii)}$ are given by

$$\begin{aligned} \mathcal{E}_{(i),l} &= \Xi_{\sigma_{[l]}} \begin{pmatrix} \vdots \\ (\frac{\tau_n}{2})^{\ell_{[i]}-l} \bar{e}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix}, \\ \mathcal{E}_{(ii),l} &= \Xi_{\sigma_{[l]}} (A^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ (\frac{\tau_n}{2})^{\ell_{[j]}+1-l} \overline{M} M^{-1} \tilde{N}_{j,n}((f-g)(\cdot)) \\ \vdots \end{pmatrix}, \end{aligned}$$

and

$$\mathcal{E}_{(iii),l} = \Xi_{\sigma_{[l]}} \begin{pmatrix} \vdots \\ (\frac{\tau_n}{2})^{\ell_{[i]}-l} \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix}$$

with

$$\Xi_{\sigma_{[l]}} := (e_{\sigma_{[l]}}^T \otimes I_{d,d}) ((I_{r+1,r+1} \otimes I_{d,d}) + \frac{\tau_n}{2} (A^{\text{VTD}} \otimes \bar{A}))^{-1}.$$

For further progress, we rewrite the term $\mathcal{E}_{(i)}$. First of all, from (2.8) and from the collocation conditions at t_{n-1}^- for $n > 1$, cf. (1.22b), or the definition of discrete initial values for $n = 1$, respectively, we gain that

$$\bar{e}^{(i)}(t_{n-1}^-) = \overline{M} M^{-1} (f-g)^{(i-1)}(t_{n-1}) - \bar{A} \bar{e}^{(i-1)}(t_{n-1}^-) \quad \text{for } 1 \leq i \leq \lfloor \frac{k}{2} \rfloor \text{ and } 1 \leq n \leq N.$$

Now, because of Assumption 2.1, the occurring differences $(f-g)^{(i-1)}(t_{n-1})$ always vanish. So, we actually get

$$\bar{e}^{(j)}(t_{n-1}^-) = (-\bar{A})^{j-i} \bar{e}^{(i)}(t_{n-1}^-) \quad \text{for all } 0 \leq i \leq j \leq \lfloor \frac{k}{2} \rfloor \text{ and } 1 \leq n \leq N. \quad (2.29)$$

Hence, adapting the argument used in the proof of Lemma 2.26, we conclude

$$\begin{aligned} \mathcal{E}_{(i),l} &= (e_{\sigma_{[l]}}^T \otimes I_{d,d}) ((I_{r+1,r+1} \otimes I_{d,d}) + \frac{\tau_n}{2} (A^{\text{VTD}} \otimes \bar{A}))^{-1} \left(\sum_{j=l+1}^{r+1} (\frac{\tau_n}{2})^{\ell_{[j]}-l} e_j \otimes \bar{e}^{(\ell_{[j]})}(t_{n-1}^-) \right) \\ &= (e_{\sigma_{[l]}}^T \otimes I_{d,d}) ((I_{r+1,r+1} \otimes I_{d,d}) - (A^{\text{VTD}} \otimes (-\frac{\tau_n}{2} \bar{A})))^{-1} \left(\sum_{j=l+1}^{r+1} e_j \otimes (-\frac{\tau_n}{2} \bar{A})^{\ell_{[j]}-l} \right) \bar{e}^{(l)}(t_{n-1}^-) \end{aligned}$$

for $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$. But this means that

$$\mathcal{E}_{(i)} = (I_{\lfloor k/2 \rfloor + 1, \lfloor k/2 \rfloor + 1} \otimes R(-\frac{\tau_n}{2} \bar{A})) \begin{pmatrix} \bar{e}(t_{n-1}^-) \\ \vdots \\ \bar{e}(\lfloor \frac{k}{2} \rfloor)(t_{n-1}^-) \end{pmatrix}, \quad (2.30)$$

where R is the (stability) function of Lemma 2.25, also cf. Lemma 2.26, associated to the respective **VTD** method.

2.3.2 Error estimates

Before the actual error estimate is addressed, we derive some bound on the inverse of the main part of the system matrix. In the proof the following technical result, known from [19, 36], is applied.

Lemma 2.28 (Cf. [19, Lemma 3.4] and [36, Theorem 4])

Let $\omega \in \mathbb{R}$ and let ϕ be a rational function without poles in $\{z \in \mathbb{C} : \operatorname{Re}(z) \leq \omega\}$. Suppose that $\Lambda \in \mathbb{R}^{s \times s}$, $s \in \mathbb{N}$, satisfies $(v, \Lambda v) \leq \omega \|v\|^2$ for all $v \in \mathbb{R}^s$, i.e., $\mu[\Lambda] \leq \omega$. Then, $\phi(\Lambda)$ exists and we have in the corresponding matrix norm, i.e., in the spectral norm, that

$$\|\phi(\Lambda)\| \leq \sup \{|\phi(z)| : z \in \mathbb{C}, \operatorname{Re}(z) \leq \omega\}.$$

Lemma 2.29

Let $\mu \geq \mu[-\bar{A}]$, i.e., μ is supposed to satisfy (2.9). Then, it holds

$$\left\| \left((I_{r+1, r+1} \otimes I_{d, d}) + \frac{\tau_n}{2} (A^{\text{VTD}} \otimes \bar{A}) \right)^{-1} \right\| \leq C \quad \text{for all } \tau_n \in (0, \bar{\tau}]$$

with sufficiently small $\bar{\tau} > 0$. Note that $\bar{\tau}$ can be chosen arbitrarily large if $\mu \leq 0$.

Proof. We reuse and slightly adapt the notation of Lemma 2.6, which is shortly recalled now. For $z \in \mathbb{C}$, let $V(z) = (v_{ij}(z)) = (I_{r+1, r+1} - A^{\text{VTD}} z)$ and $W(z) = (w_{ij}(z)) = V(z)^{-1}$ if $V(z)$ is regular. Then, according to the notation introduced for matrix-valued functions, the main part of the system matrix $((I_{r+1, r+1} \otimes I_{d, d}) + \frac{\tau_n}{2} (A^{\text{VTD}} \otimes \bar{A}))$ simply reads $V(-\frac{\tau_n}{2} \bar{A})$. Similarly $((I_{r+1, r+1} \otimes I_{d, d}) + \frac{\tau_n}{2} (A^{\text{VTD}} \otimes \bar{A}))^{-1}$ then can be shortly written as $W(-\frac{\tau_n}{2} \bar{A})$.

From Proposition 2.24 we have *ASI*-stability for the considered Runge–Kutta-like formulation and, thus, there exists an $\omega > 0$ such that $V(z)$ is regular and all entries $w_{ij}(z)$ of $W(z)$ are uniformly bounded for $\operatorname{Re}(z) \leq \omega$. Therefore, from Lemma 2.6 we know that $W(-\frac{\tau_n}{2} \bar{A})$ exists if additionally $\frac{\tau_n}{2} \mu \leq \omega$. Furthermore, because of (2.9), Lemma 2.28 yields that

$$\|w_{ij}(-\frac{\tau_n}{2} \bar{A})\| \leq \sup \{|w_{ij}(z)| : z \in \mathbb{C}, \operatorname{Re}(z) \leq \frac{\tau_n}{2} \mu\} \leq C$$

if $\frac{\tau_n}{2} \mu \leq \omega$. But this implies $\|W(-\frac{\tau_n}{2} \bar{A})\| \leq C$ for $\tau_n > 0$ sufficiently small, which is the desired statement. Note that for $\mu \leq 0$ no restriction on τ_n (from above) is necessary. \square

We now are well prepared for the derivation of error estimates. In the next theorem a bound for the error in the time mesh points is presented. Afterwards, also the pointwise error is estimated. For convenience, we here suppose that similar to (2.23) it holds for $0 \leq i \leq r+1$ and $1 \leq n \leq N$

$$\sup_{t \in I_n} \|(f - g)^{(i)}(t)\| \leq C \left(\frac{\tau_n}{2}\right)^{r+1-i} \sup_{t \in I_n} \|f^{(r+1)}(t)\| \quad (2.31)$$

when f is sufficiently smooth and its approximation g satisfies Assumption 2.1.

Theorem 2.30

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $\mu \geq \mu[-\bar{A}]$. Moreover, suppose that f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable. Denote by \bar{u} the solution of (2.8) and let $\bar{U} \in Y_r$ be the solution of $\mathbf{VTD}_k^r(\bar{M}M^{-1}g)$ with g fulfilling Assumption 2.1 and (2.31) where we assume, if $\mu > 0$, that $\tau_n \in (0, \bar{\tau}]$ for all n with $\bar{\tau}$ sufficiently small. Then, for all $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$ and $0 \leq n \leq N$, it holds

$$\|(\bar{u} - \bar{U})^{(l)}(t_n^-)\| \leq C \tau^{r+1-l} \left(\sup_{t \in (t_0, t_n]} \|\bar{M}M^{-1}f^{(r+1)}(t)\| + \sup_{t \in (t_0, t_n]} \|\bar{u}^{(r+2)}(t)\| \right),$$

where C is independent of $\tau = \max_{1 \leq n \leq N} \tau_n$ but exponentially depends on T . Moreover, C and $\bar{\tau}$ may depend on μ but are independent of the two-sided Lipschitz constant.

Proof. Starting from the splitting (2.28), each term on the right-hand side shall be analyzed separately.

For the block components of $\mathcal{E}_{(\text{iii})}$ we find from Lemma 2.29 for all $\tau_n \in (0, \bar{\tau}]$ with $\bar{\tau} > 0$ sufficiently small that

$$\|\mathcal{E}_{(\text{iii}),l}\| \leq \underbrace{\|(e_{\sigma[l]}^T \otimes I_{d,d})\|}_{=1} \underbrace{\|((I_{r+1,r+1} \otimes I_{d,d}) + \frac{\tau_n}{2}(A^{\text{VTD}} \otimes \bar{A}))^{-1}\|}_{\leq C} \left\| \begin{pmatrix} \vdots \\ (\frac{\tau_n}{2})^{\ell_{[i]}-l} \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} \right\|$$

for $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$. Moreover, (2.26) and (2.23) imply

$$\begin{aligned} \|\bar{r}_{i,n}^{\text{VTD}}\| &= \|\tilde{N}_{i,n}^* \left(\int_{t_{n-1}}^t (\bar{u}' - \mathcal{P}_n^{\text{VTD}} \bar{u}')^{(\ell_{[i]})}(\tilde{s}) d\tilde{s} \right)\| \leq \sup_{t \in I_n} \left\| \int_{t_{n-1}}^t (\bar{u}' - \mathcal{P}_n^{\text{VTD}} \bar{u}')^{(\ell_{[i]})}(\tilde{s}) d\tilde{s} \right\| \\ &\leq \tau_n \sup_{t \in I_n} \|(\bar{u}' - \mathcal{P}_n^{\text{VTD}} \bar{u}')^{(\ell_{[i]})}(t)\| \leq C \left(\frac{\tau_n}{2}\right)^{r+2-\ell_{[i]}} \sup_{t \in I_n} \|\bar{u}^{(r+2)}(t)\|. \end{aligned}$$

Combining both estimates, we gain

$$\|\mathcal{E}_{(\text{iii}),l}\| \leq C \left(\frac{\tau_n}{2}\right)^{r+2-l} \sup_{t \in I_n} \|\bar{u}^{(r+2)}(t)\|.$$

Similarly, again with Lemma 2.29, it follows for the block components of $\mathcal{E}_{(\text{ii})}$ that

$$\|\mathcal{E}_{(\text{ii}),l}\| \leq \underbrace{\|\Xi_{\sigma[l]}\|}_{\leq 1 \cdot C} \underbrace{\|(A^{\text{VTD}} \otimes I_{d,d})\|}_{\leq C} \left\| \begin{pmatrix} \vdots \\ (\frac{\tau_n}{2})^{\ell_{[j]}+1-l} \bar{M}M^{-1} \tilde{N}_{j,n}((f-g)(\cdot)) \\ \vdots \end{pmatrix} \right\|$$

for all $\tau_n \in (0, \bar{\tau}]$ with $\bar{\tau} > 0$ sufficiently small. From (2.31) we gain

$$\left\| \tilde{N}_{j,n}((f - g)(\cdot)) \right\| \leq \sup_{t \in I_n} \|(f - g)^{(\ell_{[j]})}(t)\| \leq C \left(\frac{\tau_n}{2}\right)^{r+1-\ell_{[j]}} \sup_{t \in I_n} \|f^{(r+1)}(t)\|.$$

Therefore, we get for $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$

$$\|\mathcal{E}_{(ii),l}\| \leq C \left(\frac{\tau_n}{2}\right)^{r+2-l} \sup_{t \in I_n} \|\overline{M}M^{-1}f^{(r+1)}(t)\|.$$

For the block components of $\mathcal{E}_{(i)}$, we get from (2.30) that

$$\|\mathcal{E}_{(i),l}\| \leq \|R(-\frac{\tau_n}{2}\overline{A})\| \|\bar{e}^{(l)}(t_{n-1}^-)\|$$

for $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$. Further, because of (2.9), Lemma 2.28, and Lemma 2.25, we conclude

$$\|R(-\frac{\tau_n}{2}\overline{A})\| \leq \sup \{|R(z)| : z \in \mathbb{C}, \operatorname{Re}(z) \leq \frac{\tau_n}{2}\mu\} \leq \begin{cases} 1, & \text{if } \mu \leq 0, \\ 1 + \tilde{C}\frac{\tau_n}{2}, & \text{if } \mu > 0, \end{cases}$$

for $\tau_n \in (0, \bar{\tau}]$ with $\bar{\tau} > 0$ sufficiently small (if $\mu > 0$, to ensure that R has no poles in the considered area).

Altogether, the above estimates result in

$$\begin{aligned} \|\bar{e}^{(l)}(t_n^-)\| &\leq \|\mathcal{E}_{(i),l}\| + \|\mathcal{E}_{(ii),l}\| + \|\mathcal{E}_{(iii),l}\| \\ &\leq (1 + \tilde{C}\frac{\tau_n}{2}) \|\bar{e}^{(l)}(t_{n-1}^-)\| + C \left(\frac{\tau_n}{2}\right)^{r+2-l} \left(\sup_{t \in I_n} \|\overline{M}M^{-1}f^{(r+1)}(t)\| + \sup_{t \in I_n} \|\bar{u}^{(r+2)}(t)\| \right) \end{aligned}$$

for $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$ and $1 \leq n \leq N$ if $\tau_n \in (0, \bar{\tau}]$ with $\bar{\tau} > 0$ sufficiently small. Again we note that $\bar{\tau}$ can be chosen arbitrarily large if $\mu \leq 0$. A discrete version of Gronwall's lemma, see Lemma A.1, then yields

$$\begin{aligned} \|\bar{e}^{(l)}(t_n^-)\| &\leq \exp\left(\frac{\tilde{C}}{2}(t_n - t_0)\right) \\ &\quad \left(\|\bar{e}^{(l)}(t_0^-)\| + \sum_{\nu=1}^n C \left(\frac{\tau_\nu}{2}\right)^{r+2-l} \left(\sup_{t \in I_\nu} \|\overline{M}M^{-1}f^{(r+1)}(t)\| + \sup_{t \in I_\nu} \|\bar{u}^{(r+2)}(t)\| \right) \right) \\ &\leq \exp\left(\frac{\tilde{C}}{2}(t_n - t_0)\right) \left(C\tau^{r+1-l} \left(\sup_{t \in (t_0, t_n]} \|\overline{M}M^{-1}f^{(r+1)}(t)\| + \sup_{t \in (t_0, t_n]} \|\bar{u}^{(r+2)}(t)\| \right) \left(\sum_{\nu=1}^n \frac{\tau_\nu}{2} \right) \right) \\ &\leq C(t_n - t_0) \exp\left(\frac{\tilde{C}}{2}(t_n - t_0)\right) \tau^{r+1-l} \left(\sup_{t \in (t_0, t_n]} \|\overline{M}M^{-1}f^{(r+1)}(t)\| + \sup_{t \in (t_0, t_n]} \|\bar{u}^{(r+2)}(t)\| \right), \end{aligned}$$

where we also used that $\bar{e}^{(l)}(t_0^-) = 0$. This is the desired statement. \square

Theorem 2.31

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $\mu \geq \mu[-\bar{A}]$. Moreover, suppose that f is globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable. Denote by \bar{u} the solution of (2.8) and let $\bar{U} \in Y_r$ be the solution of $\mathbf{VTD}_k^r(\bar{M}M^{-1}g)$ with g fulfilling Assumption 2.1 and (2.31) where we assume, if $\mu > 0$, that $\tau_n \in (0, \bar{\tau}]$ for all n with $\bar{\tau}$ sufficiently small. Then, for all $1 \leq n \leq N$, it holds

$$\begin{aligned} & \sup_{t \in I_n} \|(\bar{u} - \bar{U})(t)\| \\ & \leq C\tau^{r+1} \left(\sup_{t \in (t_0, t_n]} \|\bar{M}M^{-1}f^{(r+1)}(t)\| + \sup_{t \in (t_0, t_n]} \|\bar{u}^{(r+1)}(t)\| + \sup_{t \in (t_0, t_n]} \|\bar{u}^{(r+2)}(t)\| \right), \end{aligned}$$

where C is independent of $\tau = \max_{1 \leq n \leq N} \tau_n$ but exponentially depends on T . Moreover, C and $\bar{\tau}$ may depend on μ but are independent of the two-sided Lipschitz constant.

Proof. We start decomposing the error $\bar{e} = \bar{u} - \bar{U}$ as follows

$$\sup_{t \in I_n} \|(\bar{u} - \bar{U})(t)\| \leq \sup_{t \in I_n} \|(\bar{u} - \mathcal{P}_n^{\text{VTD}}\bar{u})(t)\| + \sup_{t \in I_n} \|(\mathcal{P}_n^{\text{VTD}}\bar{u} - \bar{U})(t)\|.$$

The first term on the right-hand side can be bounded by (2.23) to

$$\sup_{t \in I_n} \|(\bar{u} - \mathcal{P}_n^{\text{VTD}}\bar{u})(t)\| \leq C \left(\frac{\tau_n}{2}\right)^{r+1} \sup_{t \in I_n} \|\bar{u}^{(r+1)}(t)\|.$$

In order to estimate the second term, we use that $\mathcal{P}_n^{\text{VTD}}\bar{U}|_{I_n} = \bar{U}|_{I_n}$, which holds since $\bar{U}|_{I_n} \in P_r(I_n, \mathbb{R}^d)$, and exploit the definition of $\mathcal{P}_n^{\text{VTD}}$ to obtain

$$\begin{aligned} & \sup_{t \in I_n} \|(\mathcal{P}_n^{\text{VTD}}\bar{u} - \bar{U})(t)\| = \sup_{t \in I_n} \|\mathcal{P}_n^{\text{VTD}}\bar{e}(t)\| \\ & \leq \sum_{i=1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell_{[i]}} \|\tilde{N}_{i,n}(\bar{e}(\cdot))\| \sup_{t \in I_n} |(\hat{B}_i \circ T_n^{-1})(t)| = \sum_{i=1}^{r+1} \left(\frac{\tau_n}{2}\right)^{\ell_{[i]}} \|\tilde{N}_{i,n}(\bar{e}(\cdot))\| \overbrace{\sup_{\hat{t} \in (-1, 1]} |\hat{B}_i(\hat{t})|}^{\leq C}. \end{aligned}$$

Thus, it only remains to derive suitable bounds on $\left(\frac{\tau_n}{2}\right)^{\ell_{[i]}} \|\tilde{N}_{i,n}(\bar{e}(\cdot))\|$.

Now, from the identity (2.27) and using $A_n^{\text{VTD}} = S_n^{-1}A^{\text{VTD}}S_n$, it follows that

$$\left(\frac{\tau_n}{2}\right)^{\ell_{[l]}} \|\tilde{N}_{l,n}(\bar{e}(\cdot))\| \leq \|\tilde{\mathcal{E}}_{(i),l}\| + \|\tilde{\mathcal{E}}_{(ii),l}\| + \|\tilde{\mathcal{E}}_{(iii),l}\|$$

for $1 \leq l \leq r+1$ where the vectors $\tilde{\mathcal{E}}_{(i),l}$, $\tilde{\mathcal{E}}_{(ii),l}$, and $\tilde{\mathcal{E}}_{(iii),l}$ are given by

$$\begin{aligned} \tilde{\mathcal{E}}_{(i),l} &= \Xi_l \begin{pmatrix} \vdots \\ \left(\frac{\tau_n}{2}\right)^{\ell_{[i]}} \bar{e}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix}, \\ \tilde{\mathcal{E}}_{(ii),l} &= \Xi_l (A^{\text{VTD}} \otimes I_{d,d}) \begin{pmatrix} \vdots \\ \left(\frac{\tau_n}{2}\right)^{\ell_{[l]}+1} \bar{M}M^{-1} \tilde{N}_{j,n}((f-g)(\cdot)) \\ \vdots \end{pmatrix}, \\ \tilde{\mathcal{E}}_{(iii),l} &= \Xi_l \begin{pmatrix} \vdots \\ \left(\frac{\tau_n}{2}\right)^{\ell_{[i]}} \bar{r}_{i,n}^{\text{VTD}} \\ \vdots \end{pmatrix} \end{aligned}$$

with

$$\Xi_l := (e_l^T \otimes I_{d,d}) \left((I_{r+1,r+1} \otimes I_{d,d}) + \frac{\tau_n}{2} (A^{\text{VTD}} \otimes \bar{A}) \right)^{-1}.$$

Applying similar techniques as used to bound $\mathcal{E}_{(\text{ii}),l}$ and $\mathcal{E}_{(\text{iii}),l}$, see the proof of Theorem 2.30, we gain

$$\|\tilde{\mathcal{E}}_{(\text{ii}),l}\| + \|\tilde{\mathcal{E}}_{(\text{iii}),l}\| \leq C \left(\frac{\tau_n}{2} \right)^{r+2} \left(\sup_{t \in I_n} \|\bar{M} M^{-1} f^{(r+1)}(t)\| + \sup_{t \in I_n} \|\bar{u}^{(r+2)}(t)\| \right).$$

Further, Lemma 2.29 and Theorem 2.30 yield

$$\begin{aligned} \|\tilde{\mathcal{E}}_{(i),l}\| &\leq \underbrace{\|(e_l^T \otimes I_{d,d})\|}_{=1} \underbrace{\left\| \left((I_{r+1,r+1} \otimes I_{d,d}) + \frac{\tau_n}{2} (A^{\text{VTD}} \otimes \bar{A}) \right)^{-1} \right\|}_{\leq C} \left\| \begin{pmatrix} \vdots \\ (\frac{\tau_n}{2})^{\ell_{[i]}} \bar{e}^{(\ell_{[i]})}(t_{n-1}^-) \\ \vdots \end{pmatrix} \right\| \\ &\leq C \tau^{r+1} \left(\sup_{t \in (t_0, t_{n-1}]} \|\bar{M} M^{-1} f^{(r+1)}(t)\| + \sup_{t \in (t_0, t_{n-1}]} \|\bar{u}^{(r+2)}(t)\| \right) \end{aligned}$$

if $\tau_\nu \in (0, \bar{\tau}]$ for all ν with $\bar{\tau}$ sufficiently small. Of course, also here $\bar{\tau}$ can be chosen arbitrarily large if $\mu \leq 0$.

Combining the above estimates, we easily finish the proof. \square

Remark 2.32

Note that, because of $e = \bar{M}^{-1} \bar{e}$ and since \bar{M} is independent of t , we also gain analogous results for the error $e = u - U$. To this end, we only need to use that

$$\|(u - U)^{(l)}(t)\| = \|\bar{M}^{-1} (\bar{u} - \bar{U})^{(l)}(t)\| \leq \|\bar{M}^{-1}\| \|(\bar{u} - \bar{U})^{(l)}(t)\| \leq C \|(\bar{u} - \bar{U})^{(l)}(t)\|$$

for $0 \leq l \leq \lfloor \frac{k}{2} \rfloor$ and $t \in I_n$, $1 \leq n \leq N$.

Of course, C is independent of τ . However, for example for semi-discretizations in space of time-space problems, \bar{M} and so C may depend on the spatial mesh parameter h . Therefore, closer considerations would be needed to check whether or not we can also conclude h -uniform estimates for $u - U$ in this case. \clubsuit

2.3.3 Numerical results

In this subsection, we want to present some computational results in the case of stiff problems. To this end, we have a look on one of the standard problems in the study of numerical methods for stiff differential equations – the example of Prothero and Robinson, see [49, Example 1].

Example

We consider the initial value problem

$$\begin{aligned} u'(t) &= \tilde{g}'(t) + \lambda(u(t) - \tilde{g}(t)), \quad t \in (0, 10), \quad u(0) = \tilde{g}(0), \\ \tilde{g}(t) &= 10 - (10 + t)e^{-t}, \quad \lambda \in \mathbb{R}. \end{aligned} \tag{2.32}$$

For any λ the solution of the problem is given by $u(t) = \tilde{g}(t) = 10 - (10 + t)e^{-t}$.

This example has many advantages. On the one hand, it is quite simple and fits in the form (1.21). On the other hand, the stiffness is directly controllable via λ while the solution itself does not depend on λ . Moreover, there is a non-vanishing right-hand side $f = \tilde{g}' - \lambda \tilde{g}$ such that the test problem does not automatically force the case of cascadic interpolation.

All computational results given below were carried out with the software Julia [18], where we used the floating point data type `BigFloat` with 512 bits.

We are mainly interested in studying the influence of stiffness on the convergence behavior. Therefore, in Table 2.1 and Table 2.2 the errors of $Q_3^6\text{-VTD}_3^6$ in different norms and semi-norms are listed for $\lambda \in \{-10, -1000\}$ and $\lambda = -100000$, respectively. Here, of course, problem (2.32) can be assessed as non-stiff for $\lambda = -10$ while it is rather stiff in the case $\lambda = -100000$. Error results and associated experimental orders of convergence (eoc) are given for a wide range of equidistant meshes with $N = 2^i$, $i = 5, \dots, 13$, uniform subintervals.

We note that the pointwise errors $\|u - U\|_{L^\infty}$ are quite similar for all λ . Moreover, we clearly obtain an associated convergence order of $r + 1 = 7$ as expected from Theorem 2.31. Hence, the numerical results show that stiffness does not influence the L^∞ -order.

The situation is quite different for $\|u - U\|_{\ell^\infty}$. Although the error in the time mesh points is significantly smaller than the pointwise error for all λ , there are substantial differences in the obtained experimental orders of convergence. In the non-stiff case $\lambda = -10$, we clearly see the typical (non-stiff) superconvergence order $2r - k + 1 = 10$ over a wide range of meshes. For $\lambda = -1000$ we start for the coarse mesh with $N = 32$ subintervals with an eoc of about 6 and only reach an order just under 10 for the relatively fine mesh with $N = 4096$. For the rather stiff case $\lambda = -100000$ the experimental convergence orders are about 6 for all considered meshes, although they show an upward trend. Thus, a classical superconvergence behavior as in the non-stiff case cannot be expected for stiff problems. However, we again want to stress that the error in the time mesh points is very much smaller than the pointwise error also in the stiff case.

Table 2.1: Error of $Q_3^6\text{-VTD}_3^6$ in different (semi-)norms and associated experimental convergence orders

N	$\lambda = -10$				$\lambda = -1000$			
	$\ u - U\ _{L^\infty}$	eoc	$\ u - U\ _{\ell^\infty}$	eoc	$\ u - U\ _{L^\infty}$	eoc	$\ u - U\ _{\ell^\infty}$	eoc
32	7.376e-11	6.92	1.646e-13	9.78	7.409e-11	6.93	6.452e-15	6.08
64	6.090e-13	6.96	1.870e-16	9.94	6.091e-13	6.97	9.552e-17	6.24
128	4.884e-15	6.98	1.902e-19	9.98	4.881e-15	6.98	1.262e-18	6.51
256	3.864e-17	6.99	1.877e-22	10.00	3.862e-17	6.99	1.380e-20	6.99
512	3.038e-19	7.00	1.838e-25	10.00	3.036e-19	7.00	1.087e-22	7.74
1024	2.381e-21	7.00	1.797e-28	10.00	2.380e-21	7.00	5.077e-25	8.68
2048	1.863e-23	7.00	1.755e-31	10.00	1.863e-23	7.00	1.234e-27	9.46
4096	1.456e-25	7.00	1.714e-34	10.00	1.456e-25	7.00	1.754e-30	9.84
8192	1.138e-27		1.674e-37		1.138e-27		1.921e-33	

Table 2.2: Error of $Q_3^6\text{-VTD}_3^6$ in different (semi-)norms and associated experimental convergence orders

$\lambda = -100000$								
N	$\ u - U\ _{L^\infty}$	eoc	$\ u - U\ _{\ell^\infty}$	eoc	$\ (u - U)'\ _{L^\infty}$	eoc	$\ (u - U)'\ _{\ell^\infty}$	eoc
32	7.414e-11	6.93	7.094e-19	5.94	2.599e-09	5.92	7.094e-14	5.94
64	6.095e-13	6.96	1.154e-20	5.97	4.281e-11	5.96	1.154e-15	5.97
128	4.884e-15	6.98	1.836e-22	5.99	6.867e-13	5.98	1.836e-17	5.99
256	3.864e-17	6.99	2.886e-24	6.00	1.087e-14	5.99	2.886e-19	6.00
512	3.038e-19	7.00	4.497e-26	6.02	1.710e-16	6.00	4.497e-21	6.02
1024	2.381e-21	7.00	6.937e-28	6.04	2.680e-18	6.00	6.937e-23	6.04
2048	1.863e-23	7.00	1.052e-29	6.09	4.195e-20	6.00	1.052e-24	6.09
4096	1.456e-25	7.00	1.548e-31	6.17	6.560e-22	6.00	1.548e-26	6.17
8192	1.138e-27		2.143e-33		1.025e-23		2.143e-28	

The results presented in Table 2.2 show two conspicuous features. Firstly, the experimental ℓ^∞ -order is smaller than the L^∞ -order and, secondly, the errors $\|u - U\|_{\ell^\infty}$ and $\|(u - U)'\|_{\ell^\infty}$ only differ by a factor. Therefore, for further examination the results of $Q_k^6\text{-VTD}_k^6$, $k = 0, \dots, 6$, for $\lambda = -100000$ are summarized in Table 2.3. The given errors are those for $N \in \{256, 512\}$ and the experimental convergence orders are calculated from these values. However, we want to remark that not for all k the range of the experimental ℓ^∞ -orders (when considering meshes with $N = 2^i$, $i = 5, \dots, 13$) is as narrow as for $k = 3$.

First of all, all methods show an L^∞ -order of $r + 1 = 7$ and, thus, also confirm Theorem 2.31. Moreover, for $k \geq 2$ we have that $\|u - U\|_{\ell^\infty}$ and $\|(u - U)'\|_{\ell^\infty}$ only differ by factor $|\lambda| = 10^5$, while this is not the case for $k \in \{0, 1\}$. This behavior is in full accordance with (2.29). The unexpectedly lower experimental ℓ^∞ -order compared to the L^∞ -order, which was already seen for $k = 3$, also shows up for all $k \geq 2$. It seems that an experimental ℓ^∞ -order of about $r + 1 - \lfloor \frac{k}{2} \rfloor$ is obtained and, thus, just the order of the maximal derivative covered by Theorem 2.30. Nevertheless, because of $\|(u - U)^{(l)}\|_{\ell^\infty} \leq \|(u - U)^{(l)}\|_{L^\infty}$ and since we gain the expected L^∞ - and $W^{1,\infty}$ -orders of $r + 1 = 7$ and $r = 6$, respectively, this is not really a contradiction to the estimates of Theorem 2.30.

Table 2.3: Error of $Q_k^6\text{-VTD}_k^6$, $k = 0, \dots, 6$, in different (semi-)norms and associated experimental convergence orders

$\lambda = -100000$								
k	$\ u - U\ _{L^\infty}$	eoc	$\ u - U\ _{\ell^\infty}$	eoc	$\ (u - U)'\ _{L^\infty}$	eoc	$\ (u - U)'\ _{\ell^\infty}$	eoc
0	4.755e-17 3.738e-19	6.99	2.744e-22 2.128e-24	7.01	5.965e-14 9.378e-16	5.99	8.494e-15 1.337e-16	5.99
1	3.259e-21 2.551e-23	7.00	1.259e-24 5.716e-27	7.78	1.164e-17 1.822e-19	6.00	1.163e-17 1.821e-19	6.00
2	6.086e-21 4.764e-23	7.00	1.129e-27 1.713e-29	6.04	2.327e-17 3.643e-19	6.00	1.129e-22 1.713e-24	6.04
3	3.864e-17 3.038e-19	6.99	2.886e-24 4.497e-26	6.00	1.087e-14 1.710e-16	5.99	2.886e-19 4.497e-21	6.00
4	1.888e-20 1.478e-22	7.00	1.783e-29 5.490e-31	5.02	1.529e-17 2.394e-19	6.00	1.783e-24 5.490e-26	5.02
5	1.517e-16 1.193e-18	6.99	2.980e-25 9.906e-27	4.91	3.261e-14 5.129e-16	5.99	2.980e-20 9.906e-22	4.91
6	6.833e-16 5.377e-18	6.99	2.731e-29 1.709e-30	4.00	6.999e-14 1.101e-15	5.99	2.731e-24 1.709e-25	4.00

Part II

Variational Time Discretization Methods for Parabolic Problems

3 Introduction to Parabolic Problems

In the following, we want to study parabolic problems of the form

$$\partial_t u(t) + \mathcal{A}u(t) = f(t) \quad \text{in } \Omega, \quad t_0 < t < t_0 + T, \quad (3.1a)$$

$$\mathcal{B}u(t) = 0 \quad \text{on } \partial\Omega, \quad t_0 < t < t_0 + T, \quad (3.1b)$$

$$u(t_0) = u_0 \quad \text{in } \Omega, \quad (3.1c)$$

where $\Omega \subset \mathbb{R}^{d_\Omega}$, $d_\Omega \in \mathbb{N}$, is a bounded domain with boundary $\partial\Omega$ and $T > 0$ some time horizon. Here, \mathcal{A} is a uniformly elliptic linear differential operator independent of time t . In addition, \mathcal{B} is some linear operator (also independent of t) modeling the boundary conditions. Further assumptions on \mathcal{A} and \mathcal{B} will be stated later on. As before, we set $I = (t_0, t_0 + T]$ for brevity.

Most parts of our analysis will, however, consider parabolic problems in their weak formulation. Therefore, we provide an abstract setting for this generalized formulation at first. To this end, let $(H, (\cdot, \cdot))$ and $(V, (\cdot, \cdot)_V)$ denote two Hilbert spaces with V continuously embedded in H (for brevity, $V \hookrightarrow H$), i.e., $V \subset H$ and there is a positive constant $C_{\text{emb}} > 0$ such that $\|v\| \leq C_{\text{emb}} \|v\|_V$ for all $v \in V$. Moreover, suppose that V is dense in H . Then, identifying H with H' , we have that $V \subset H \equiv H' \subset V'$. Thereby the duality pairing $\langle \cdot, \cdot \rangle_{V', V}$ can be viewed as extension of (\cdot, \cdot) . Furthermore, let $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ be a continuous, V -elliptic bilinear form, i.e.,

$$\exists \alpha > 0 : \quad a(v, v) \geq \alpha \|v\|_V^2 \quad \forall v \in V, \quad (3.2a)$$

$$\exists C_a > 0 : \quad |a(v, w)| \leq C_a \|v\|_V \|w\|_V \quad \forall v, w \in V. \quad (3.2b)$$

Here, (3.2a) means the V -ellipticity and (3.2b) the continuity of the bilinear form $a(\cdot, \cdot)$.

The abstract generalized formulation then is given for $f \in L^2(I, V')$ and $u_0 \in H$ by

Find $u \in \mathcal{W}(V, V') := \{v \in L^2(I, V) : \partial_t v \in L^2(I, V')\}$ with $u(t_0) = u_0$ such that

$$\langle \partial_t u(t), v \rangle_{V', V} + a(u(t), v) = \langle f(t), v \rangle_{V', V} \quad \text{for a.e. } t \in I, \forall v \in V. \quad (3.3)$$

Note that it holds $\mathcal{W}(V, V') \subset C(\bar{I}, H)$. So, the initial condition $u(t_0) = u_0$ is meaningful. Here, usual notation for Bochner spaces is used, for details see e.g. [25, Subsection 6.1.1]. Especially, the definitions of the function spaces introduced in Part I are extended to Banach space-valued functions.

It is well known that under certain further assumptions on the data, problem (3.1) can be rewritten in the form (3.3). Of course, appropriate choices of the spaces H and V as well as of the bilinear form $a(\cdot, \cdot)$ then strongly depend on the nature of \mathcal{A} and \mathcal{B} . Note that in this way the requirements made on the involved spaces and bilinear form may implicitly cause (additional) assumptions on \mathcal{A} (and \mathcal{B}).

In addition to the already used notation for (Banach space-valued) square-integrable or continuously differentiable functions, in the following also standard notation for Sobolev and Bochner–Sobolev spaces is used. So, for arbitrary $p \in [1, \infty]$ and $m \in \mathbb{Z}$, $m \geq 0$, let $W^{m,p}(\Omega)$ denote the Sobolev space of $L^p(\Omega)$ functions whose weak derivatives up to order m are also in $L^p(\Omega)$. The associated norms are given by

$$\begin{aligned} \|v\|_{W^{m,p}(\Omega)} &= \left(\sum_{0 \leq |\alpha| \leq m} \|D^\alpha v\|_{L^p(\Omega)}^p \right)^{1/p} = \left(\sum_{0 \leq |\alpha| \leq m} \int_{\Omega} |D^\alpha v(x)|^p dx \right)^{1/p}, \quad \text{if } p \in [1, \infty), \\ \|v\|_{W^{m,\infty}(\Omega)} &= \max_{0 \leq |\alpha| \leq m} \|D^\alpha v\|_{L^\infty(\Omega)} = \max_{0 \leq |\alpha| \leq m} \operatorname{ess\,sup}_{x \in \Omega} |D^\alpha v(x)|, \quad \text{if } p = \infty. \end{aligned}$$

Here, we use the notation $D^\alpha = \partial_{x_1}^{\alpha_1} \cdots \partial_{x_{d_\Omega}}^{\alpha_{d_\Omega}}$ where $\alpha = (\alpha_1, \dots, \alpha_{d_\Omega}) \in \mathbb{N}_0^{d_\Omega}$ is a multi-index with $|\alpha| = \alpha_1 + \dots + \alpha_{d_\Omega}$. Usually, we simply write $L^p(\Omega)$ instead of $W^{0,p}(\Omega)$ and $H^m(\Omega)$ for $W^{m,2}(\Omega)$. Moreover, we write $H_0^1(\Omega) := \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega\}$ for the subspace of $H^1(\Omega)$ of functions having zero boundary traces and $H^{-1}(\Omega) = H_0^1(\Omega)'$ for its dual space.

In analogy to the definition of Sobolev spaces, for an arbitrary interval J and a Banach space X let $W^{m,p}(J, X)$ with $p \in [1, \infty]$ and $m \in \mathbb{Z}$, $m \geq 0$, denote the respective Bochner–Sobolev space of X -valued functions. Of course, also here $W^{0,p}(J, X) = L^p(J, X)$ and $H^m(J, X) = W^{m,2}(J, X)$. We have

$$\begin{aligned} \|v\|_{W^{m,p}(J,X)} &= \left(\sum_{0 \leq j \leq m} \|\partial_t^j v\|_{L^p(J,X)}^p \right)^{1/p} = \left(\sum_{0 \leq j \leq m} \int_J \|\partial_t^j v(t)\|_X^p dt \right)^{1/p}, \quad \text{if } p \in [1, \infty), \\ \|v\|_{W^{m,\infty}(J,X)} &= \max_{0 \leq j \leq m} \|\partial_t^j v\|_{L^\infty(J,X)} = \max_{0 \leq j \leq m} \operatorname{ess\,sup}_{t \in J} \|\partial_t^j v(t)\|_X, \quad \text{if } p = \infty. \end{aligned}$$

Moreover, for sufficiently smooth functions we also use the norm $\|v\|_{C(J,X)} = \sup_{t \in J} \|v(t)\|_X$.

Model problem

For simplicity in the following we mainly concentrate on the model problem

$$\partial_t u(t) - \operatorname{div}(\epsilon \nabla u(t)) + b \cdot \nabla u(t) + cu(t) = f(t) \quad \text{in } \Omega, \quad t_0 < t < t_0 + T, \quad (3.4a)$$

$$u(t) = 0 \quad \text{on } \partial\Omega, \quad t_0 < t < t_0 + T, \quad (3.4b)$$

$$u(t_0) = u_0 \quad \text{in } \Omega, \quad (3.4c)$$

with coefficients ϵ , b , and c defined over Ω and taking values in $\mathbb{R}^{d_\Omega \times d_\Omega}$, \mathbb{R}^{d_Ω} , and \mathbb{R} , respectively, where we additionally assume that there is a constant $\epsilon_0 > 0$ such that

$$z^T \epsilon(x) z \geq \epsilon_0 z^T z \quad \text{for a.e. } x \in \Omega \text{ and all } z \in \mathbb{R}^{d_\Omega}. \quad (3.4d)$$

This means that the operators \mathcal{A} and \mathcal{B} in (3.1) are specified by

$$\mathcal{A}v = -\operatorname{div}(\epsilon \nabla v) + b \cdot \nabla v + cv \quad \text{and} \quad \mathcal{B}v = v. \quad (3.5a)$$

Moreover, in this case we have that

$$a(v, w) = (\epsilon \nabla v, \nabla w) + (b \cdot \nabla v, w) + (cv, w) \quad (3.5b)$$

and the two occurring Hilbert spaces then are

$$\begin{aligned} H &= L^2(\Omega) \quad \text{with the inner product} \quad (v, w) = \int_{\Omega} vw \, dx \quad \text{and} \\ V &= H_0^1(\Omega) \quad \text{with the inner product} \quad (v, w)_V = (\nabla v, \nabla w) + (v, w). \end{aligned} \quad (3.5c)$$

Obviously, it holds $V \hookrightarrow H$ with $C_{\text{emb}} = 1$. Provided that $\epsilon \in [L^\infty(\Omega)]^{d_\Omega \times d_\Omega}$ satisfies (3.4d), $b \in [L^\infty(\Omega)]^{d_\Omega}$, $\text{div}(b) \in L^\infty(\Omega)$, as well as $c \in L^\infty(\Omega)$, the V -ellipticity and the continuity of $a(\cdot, \cdot)$ can be guaranteed if $c - \frac{1}{2}\text{div}(b) \geq 0$. For further details see e.g. [25, Theorem 3.8, pp. 115–116]. In the following, we assume that (3.2) holds for this model problem.

Note that the stationary problem associated to (3.4), i.e.,

$$\mathcal{A}u = \tilde{f} \quad \text{in } \Omega, \quad \mathcal{B}u = 0 \quad \text{on } \partial\Omega,$$

is uniformly elliptic and has homogeneous Dirichlet boundary conditions. We say that the stationary problem is H^2 -regular if for all $\tilde{f} \in L^2(\Omega)$ the adjoint variational problem

Find $u \in V$ such that

$$a(v, u) = (\tilde{f}, v) \quad \forall v \in V$$

has a unique solution $u \in V \cap H^2(\Omega)$ that satisfies the estimate $\|u\|_{H^2(\Omega)} \leq C \|\tilde{f}\|_{L^2(\Omega)}$. ♣

Remark 3.1

As we have already seen for the model problem, the space V and the bilinear form $a(\cdot, \cdot)$ are just those arising in the weak formulation of the associated stationary problem

$$\mathcal{A}u = \tilde{f} \quad \text{in } \Omega, \quad \mathcal{B}u = 0 \quad \text{on } \partial\Omega.$$

For linear elliptic differential equations of second order and the most typical (combinations of) boundary conditions the weak formulations and conditions for the V -ellipticity of the associated bilinear form $a(\cdot, \cdot)$ are derived in e.g. [25, Section 3.1]. Also see [25, Remark 6.10] where the time-dependent versions are broached.

An easily comprehensible overview on elliptic boundary value problems and their weak formulations that also handles higher order problems is given in [35]. For details, on how the (system of) boundary differential operator(s) \mathcal{B} could look like, especially see [35, Subsections 5.2.1 and 5.3.2] and the references provided there. For a discussion of some associated weak formulations especially see [35, Sections 7.2 and 7.4].

For a very detailed and general study of elliptic, parabolic, but also hyperbolic partial differential equations, we refer to [56]. So, for example, conditions on the equivalence of (3.1) and (3.3) are discussed in [56, Satz 27.6, pp. 403–404]. However, due to the very general setup, the notation used there is somewhat more difficult to understand. ♣

3.1 Regularity of solutions

The existence, uniqueness, and regularity of solutions to (3.1) or (3.3), respectively, have been studied in detail in the literature, see e.g. [56, Chapter 26 and 27] and references therein.

Theorem 3.2 (Cf. [56, Satz 26.1, p. 384, Satz 27.2, p. 393])

Let $j \in \mathbb{Z}$, $j \geq 0$, and suppose that (3.2) holds. Moreover, let

$$f \in H^j(I, V') \quad \text{and} \quad \partial_t^i u_0 \in V, \quad i = 0, \dots, j-1, \quad \partial_t^j u_0 \in H. \quad (3.6)$$

Then, the abstract problem (3.3) has a unique solution u satisfying

$$u \in H^j(I, V), \quad \partial_t^{j+1} u \in L^2(I, V'), \quad \partial_t^i u(t_0) = \partial_t^i u_0, \quad i = 0, \dots, j.$$

The quantities $\partial_t^i u_0$ occurring in the theorem are recursively defined via

$$\begin{aligned} \partial_t^0 u_0 &= u_0, \\ \langle \partial_t^i u_0, v \rangle_{V', V} &= \langle f^{(i-1)}(t_0), v \rangle_{V', V} - a(\partial_t^{i-1} u_0, v) \quad \forall v \in V, \quad i = 1, \dots, j. \end{aligned} \quad (3.7)$$

This nicely shows that (3.6) should not be misinterpreted as additional initial conditions but actually states certain compatibility conditions, i.e., the initial condition u_0 , $f^{(i)}(t_0)$ for $i = 0, \dots, j-1$, and the boundary conditions (given by V) should match at t_0 .

Some situations in which (3.6) is guaranteed are discussed in [56, pp. 396–397]. Note that the notation there is somewhat different since the linear operator from V to V' representing the bilinear form $a(\cdot, \cdot)$ is used to define $\partial_t^i u_0$.

If, as in our case, the abstract problem (3.3) originates from a weak formulation of a parabolic problem of the form (3.1), then also the interaction between time and space as well as the regularity of the solution with respect to the space variable is of interest. Appropriate results are given in [56, Subsections 27.2 and 27.3], especially see [56, Satz 27.5, pp. 402–403]. So, provided that the problem data satisfies suitable regularity and compatibility assumptions, then the solution can be guaranteed to be as smooth as desired in time and space. Note that in the literature often quite strong regularity assumptions on the domain Ω are supposed. However, some results may also hold for domains with nonsmooth boundary, also see the following remark.

Remark 3.3

For the solution u of the model problem, cf. (3.4), we also have on a convex domain that

$$u \in L^2(I, H_0^1(\Omega) \cap H^2(\Omega)), \quad \partial_t u \in L^2(I, L^2(\Omega))$$

if $f \in L^2(I, L^2(\Omega))$ and $u_0 \in H^1(\Omega)$, see for example [50, Proposition 11.12, p. 215]. Here additionally note that the assumptions on Ω stated in [50] can be further weakened since the elliptic H^2 -regularity, that was used to prove the H^2 -regularity, can also be guaranteed on convex domains, cf. [33, Theorem 3.2.1.2, p. 147] or [35, Theorem 9.24, p. 282].

Note that under certain assumptions the H^2 -regularity of the solution to the stationary problem can be proven for even more general domains, see [2] and [30]. ♣

3.2 Semi-discretization in space

There are several ways to approach the numerical approximation of problem (3.1) and (3.3), respectively. We shall follow the method of lines and first approximate the solution to (3.3)

in space only. This approach results in a coupled system of ordinary differential equations with respect to the time variable t . Later, in a second step, for example the methods known from Part I can be applied to obtain a fully discrete scheme.

We denote by V_h a finite dimensional subspace of V . Moreover, in order to keep things clear and simple, let $f \in C(\bar{I}, V')$. Then, consider the following semi-discretized problem

Find $u_h \in C^1(\bar{I}, V_h)$ with $u_h(t_0) = u_{h,0}$ such that

$$(\partial_t u_h(t), v_h) + a(u_h(t), v_h) = \langle f(t), v_h \rangle_{V', V} \quad \forall t \in \bar{I}, \forall v_h \in V_h, \quad (3.8)$$

where $u_{h,0} \in V_h$ is an approximation of the initial value u_0 .

This problem is called the semi-discretization in space of (3.3) and well-known from the literature, see e.g. [25, Subsection 6.1.4] or [34, Subsection 5.1.2]. If $u_0 \in V'$, then $u_{h,0} \in V_h$ may be determined via the projection $P_h : V' \rightarrow V_h$ given by

$$\langle P_h v, w \rangle_{V', V} = \langle v, w \rangle_{V', V} \quad \forall w \in V_h. \quad (3.9)$$

Note that in this definition the duality pairing $\langle \cdot, \cdot \rangle_{V', V}$ can be replaced by (\cdot, \cdot) if $v \in H$. Moreover, P_h is stable in $\|\cdot\|$, i.e., it holds $\|P_h v\| \leq \|v\|$ for all $v \in H$, which can be easily shown using the Cauchy–Schwarz inequality. However, as we shall see later, other choices for $u_{h,0}$ may be more appropriate. The concrete choices, of course, then strongly depend on the properties of the data and the properties desired from u_h .

3.2.1 Reformulation as ode system

In order to make the structure of (3.8) more clear, the problem is reformulated. Denoting by $\{\varphi_i\}_{i=1, \dots, \dim(V_h)}$ a basis of V_h , we can write $u_h \in C^1(\bar{I}, V_h)$ as

$$u_h(t) = \sum_{i=1}^{\dim(V_h)} U_{h,i}(t) \varphi_i$$

with $U_{h,i} \in C^1(\bar{I})$. Now, testing in (3.8) with φ_j , $j = 1, \dots, \dim(V_h)$, we get

$$(\partial_t u_h(t), \varphi_j) + a(u_h(t), \varphi_j) = \langle f(t), \varphi_j \rangle_{V', V} \quad \forall j = 1, \dots, \dim(V_h).$$

Then, defining the mass matrix M and the stiffness matrix A as usual by

$$M_{ij} = (\varphi_j, \varphi_i), \quad A_{ij} = a(\varphi_j, \varphi_i) \quad \forall i, j = 1, \dots, \dim(V_h), \quad (3.10)$$

the left-hand side can be rewritten as

$$\begin{aligned} (\partial_t u_h(t), \varphi_j) + a(u_h(t), \varphi_j) &= \sum_{i=1}^{\dim(V_h)} U'_{h,i}(t) (\varphi_i, \varphi_j) + \sum_{i=1}^{\dim(V_h)} U_{h,i}(t) a(\varphi_i, \varphi_j) \\ &= (MU'_h(t) + AU_h(t))_j. \end{aligned}$$

Therefore, setting $\tilde{F}_j(t) := \langle f(t), \varphi_j \rangle_{V', V}$, the basis representation U_h of the solution u_h of (3.8) satisfies the initial value problem

$$MU'_h(t) + AU_h(t) = \tilde{F}(t) \quad \forall t \in I, \quad U_h(t_0) = U_{h,0}. \quad (3.11)$$

Here, $U_{h,0} \in \mathbb{R}^{\dim(V_h)}$ denotes the basis representation of $u_{h,0}$, i.e., $u_{h,0} = \sum_{i=1}^{\dim(V_h)} (U_{h,0})_i \varphi_i$.

Remark 3.4

Looking on $P_h f(t) \in V_h$ in its basis representation, i.e., $P_h f(t) = \sum_{i=1}^{\dim(V_h)} F_i^{P_h}(t) \varphi_i$, we get from the definition of the orthogonal projection P_h that

$$\begin{aligned} \tilde{F}_j(t) &= \langle f(t), \varphi_j \rangle_{V',V} = \langle P_h f(t), \varphi_j \rangle_{V',V} \\ &= \sum_{i=1}^{\dim(V_h)} F_i^{P_h}(t) \langle \varphi_i, \varphi_j \rangle_{V',V} = \sum_{i=1}^{\dim(V_h)} F_i^{P_h}(t) (\varphi_i, \varphi_j) = (M F^{P_h}(t))_j. \end{aligned}$$

Hence, it holds $\tilde{F}(t) = M F^{P_h}(t)$. ♣

Since (3.11) is a system of coupled odes, standard ode theory can be applied to answer questions on solvability and regularity. However, we should know somewhat more about the involved matrices M and A as well as about the right-hand side \tilde{F} .

Lemma 3.5

The mass matrix $M \in \mathbb{R}^{\dim(V_h) \times \dim(V_h)}$ is symmetric and positive definite.

Proof. The symmetry of M follows easily from its definition due to the symmetry of the inner product (\cdot, \cdot) .

Now, we study the positive definiteness. To this end, let $Z \in \mathbb{R}^{\dim(V_h)} \setminus \{0\}$ and associated $z_h = \sum_{i=1}^{\dim(V_h)} Z_i \varphi_i \in V_h \setminus \{0\}$ be given. Then, it holds

$$Z^T M Z = \sum_{i,j=1}^{\dim(V_h)} Z_j (\varphi_j, \varphi_i) Z_i = \left(\sum_{j=1}^{\dim(V_h)} Z_j \varphi_j, \sum_{i=1}^{\dim(V_h)} Z_i \varphi_i \right) = (z_h, z_h) = \|z_h\|^2 > 0$$

and we are done. □

We now know that M is symmetric and positive definite, which also implies existence, symmetry, and positive definiteness of $M^{1/2}$. It is appropriate to define

$$\overline{M} := M^{1/2} \quad \text{and} \quad \overline{A} := \overline{M} M^{-1} A \overline{M}^{-1} = M^{-1/2} A M^{-1/2}.$$

Then, setting $\overline{U}_h = \overline{M} U_h = M^{1/2} U_h$ and $\overline{U}_{h,0} = \overline{M} U_{h,0} = M^{1/2} U_{h,0}$, in addition to (3.11) it also holds

$$\overline{U}_h'(t) + \overline{A} \overline{U}_h(t) = \overline{M} M^{-1} \tilde{F}(t) \quad \forall t \in I, \quad \overline{U}_h(t_0) = \overline{U}_{h,0}. \quad (3.12)$$

Since this is a finite linear system of ordinary differential equations with constant coefficients in standard form, we have that

$$\overline{U}_h(t) = e^{-(t-t_0)\overline{A}} \overline{U}_h(t_0) + \int_{t_0}^t e^{-(t-s)\overline{A}} \overline{M} M^{-1} \tilde{F}(s) ds. \quad (3.13)$$

Thus, the regularity of \overline{U}_h only depends on the smoothness of the right-hand side \tilde{F} .

In case of the standard application, where V_h is a conforming finite element space, it is well-known that the system gets stiffer if the spatial mesh gets finer. In fact, on shape-regular,

quasi-uniform meshes the two-sided Lipschitz constant associated to the semi-discretization of model problem (3.4) then is proportional to h^{-2} with h denoting the spatial mesh parameter, also see Remark 3.8. Therefore, we ask and check whether at least a uniform one-sided Lipschitz condition is satisfied for problem (3.12). To this end, we have to verify that (2.9) holds with μ independent of V_h .

In the following, the notations $\|\cdot\|$ and (\cdot, \cdot) are also used for the Euclidean norm and inner product. From the context, however, it will always be easy to understand what meaning is meant.

Lemma 3.6

For every $\bar{Z} \in \mathbb{R}^{\dim(V_h)}$ it holds

$$(-\bar{A}\bar{Z}, \bar{Z}) \leq \mu \|\bar{Z}\|^2$$

with $\mu = -\alpha C_{\text{emb}}^{-2} < 0$, where $\alpha > 0$ is the V -ellipticity constant of $a(\cdot, \cdot)$.

Proof. Let $Z \in \mathbb{R}^{\dim(V_h)}$ be arbitrarily chosen and $z_h = \sum_{i=1}^{\dim(V_h)} Z_i \varphi_i \in V_h$. Then,

$$(AZ, Z) = Z^T AZ = \sum_{i,j=1}^{\dim(V_h)} Z_j a(\varphi_j, \varphi_i) Z_i = a\left(\sum_{j=1}^{\dim(V_h)} Z_j \varphi_j, \sum_{i=1}^{\dim(V_h)} Z_i \varphi_i\right) = a(z_h, z_h).$$

Now, due to the V -ellipticity of $a(\cdot, \cdot)$ and $V \hookrightarrow H$, we have

$$a(z_h, z_h) \geq \alpha \|z_h\|_V^2 \geq \alpha C_{\text{emb}}^{-2} \|z_h\|^2.$$

Recalling the identity used in the proof of Lemma 3.5, the norm on the right-hand side can further be rewritten as

$$\|z_h\|^2 = Z^T M Z = Z^T M^{1/2} M^{1/2} Z = (M^{1/2} Z)^T (M^{1/2} Z) = \|M^{1/2} Z\|^2,$$

where we used that $M^{1/2}$ exists and is symmetric since M is symmetric and positive definite. Hence, it follows

$$(AZ, Z) \geq \alpha C_{\text{emb}}^{-2} \|M^{1/2} Z\|^2.$$

Multiplying this identity by -1 , setting $Z = M^{-1/2} \bar{Z}$, and recalling the definition of \bar{A} , the desired statement follows easily. Here, also note that $M^{-1/2}$ exists and is symmetric and positive definite. \square

Remark 3.7

Note that within the proof of Lemma 3.6 we have made the following observations. Let $z_h \in V_h$ be represented by the coefficient vector $Z \in \mathbb{R}^{\dim(V_h)}$, i.e., $z_h = \sum_{i=1}^{\dim(V_h)} Z_i \varphi_i$, and define $\bar{Z} \in \mathbb{R}^{\dim(V_h)}$ by $\bar{Z} = \bar{M} Z = M^{1/2} Z$. Then, it holds

$$\|\bar{Z}\| = \|z_h\| \quad \text{and} \quad (\bar{A} \bar{Z}, \bar{Z}) = a(z_h, z_h).$$

Therefore, stability and error results obtained for the coefficient vectors immediately also yield results for the represented functions in an appropriate norm and vice versa. Moreover, this suggests that estimates for ode systems as those of Section 2.3 can nicely be interpreted if the ode system results from a spatial semi-discretization of a time-space problem. \clubsuit

Remark 3.8

Adapting the arguments used in the proof of Lemma 3.6, we obtain for functions $y_h, z_h \in V_h$ and their associated basis representation vectors $Y, Z \in \mathbb{R}^{\dim(V_h)}$, i.e., $y_h = \sum_{j=1}^{\dim(V_h)} Y_j \varphi_j$ and $z_h = \sum_{i=1}^{\dim(V_h)} Z_i \varphi_i$, that

$$a(y_h, z_h) = (AY, Z) = (\overline{A} \overline{Y}, \overline{Z})$$

where $\overline{Y} = \overline{M}Y = M^{1/2}Y$ and $\overline{Z} = \overline{M}Z = M^{1/2}Z$.

Thus, inspired by the proof of [25, Theorem 9.11, pp. 388–389], we find for the spectral norm of \overline{A} that

$$\|\overline{A}\| = \sup_{\overline{Y} \in \mathbb{R}^{\dim(V_h)} \setminus \{0\}} \frac{\|\overline{A} \overline{Y}\|}{\|\overline{Y}\|} = \sup_{\overline{Y}, \overline{Z} \in \mathbb{R}^{\dim(V_h)} \setminus \{0\}} \frac{(\overline{A} \overline{Y}, \overline{Z})}{\|\overline{Y}\| \|\overline{Z}\|} = \sup_{y_h, z_h \in V_h \setminus \{0\}} \frac{a(y_h, z_h)}{\|y_h\| \|z_h\|},$$

where we also exploited the observations of Remark 3.7. Using the V -ellipticity and the continuity of $a(\cdot, \cdot)$, we further conclude

$$\alpha \left(\sup_{z_h \in V_h \setminus \{0\}} \frac{\|z_h\|_V}{\|z_h\|} \right)^2 \leq \|\overline{A}\| = \sup_{y_h, z_h \in V_h \setminus \{0\}} \frac{a(y_h, z_h)}{\|y_h\| \|z_h\|} \leq C_a \left(\sup_{z_h \in V_h \setminus \{0\}} \frac{\|z_h\|_V}{\|z_h\|} \right)^2.$$

In the setting of model problem (3.4) with (3.5) and considering a conforming finite element space V_h on a shape-regular, quasi-uniform mesh, we have that $\sup_{z_h \in V_h \setminus \{0\}} \frac{\|z_h\|_V}{\|z_h\|}$ is proportional to h^{-1} with h denoting the spatial mesh parameter. Here, an upper bound follows from an inverse inequality and an appropriate lower bound follows from choosing any non-zero function in V_h whose support has a diameter of order h , also see [25, (9.12) and (9.15), pp. 388 and 390]. Hence, $\|\overline{A}\|$ is proportional to h^{-2} then. \clubsuit

Lemma 3.9

The solution \overline{U}_h of (3.12) satisfies the following stability estimate

$$\|\overline{U}_h(t)\| \leq e^{-\alpha C_{\text{emb}}^{-2}(t-t_0)} \|\overline{U}_h(t_0)\| + \int_{t_0}^t e^{-\alpha C_{\text{emb}}^{-2}(t-s)} \|\overline{M} M^{-1} \tilde{F}(s)\| \, ds.$$

Proof. Scalar multiplying (3.12) by $\overline{U}_h(t)$, we get

$$(\overline{U}_h'(t), \overline{U}_h(t)) + (\overline{A} \overline{U}_h(t), \overline{U}_h(t)) = (\overline{M} M^{-1} \tilde{F}(t), \overline{U}_h(t)).$$

Using the result of Lemma 3.6 and the Cauchy–Schwarz inequality, we therefore gain

$$\frac{1}{2} \partial_t \|\overline{U}_h(t)\|^2 + \alpha C_{\text{emb}}^{-2} \|\overline{U}_h(t)\|^2 \leq \|\overline{M} M^{-1} \tilde{F}(t)\| \|\overline{U}_h(t)\|.$$

From this, we conclude that

$$\partial_t \|\overline{U}_h(t)\| + \alpha C_{\text{emb}}^{-2} \|\overline{U}_h(t)\| \leq \|\overline{M} M^{-1} \tilde{F}(t)\|.$$

Further, multiplying by $e^{\alpha C_{\text{emb}}^{-2} t}$, we find that

$$\partial_t (e^{\alpha C_{\text{emb}}^{-2} t} \|\overline{U}_h(t)\|) = e^{\alpha C_{\text{emb}}^{-2} t} \partial_t \|\overline{U}_h(t)\| + \alpha C_{\text{emb}}^{-2} e^{\alpha C_{\text{emb}}^{-2} t} \|\overline{U}_h(t)\| \leq e^{\alpha C_{\text{emb}}^{-2} t} \|\overline{M} M^{-1} \tilde{F}(t)\|.$$

So, replacing t by s and integrating over s from t_0 to t , we obtain the desired estimate. \square

Recalling Remark 3.4 and Remark 3.7, we immediately get from Lemma 3.9 that

$$\|u_h(t)\| \leq e^{-\alpha C_{\text{emb}}^{-2}(t-t_0)} \|u_h(t_0)\| + \int_{t_0}^t e^{-\alpha C_{\text{emb}}^{-2}(t-s)} \|f(s)\| \, ds. \quad (3.14)$$

However, under weaker assumptions on f still the following stability results can be shown.

Lemma 3.10

The solution u_h of (3.8) satisfies the following stability estimates

$$\|u_h(t)\|^2 + \alpha \int_{t_0}^t \|u_h(s)\|_V^2 \, ds \leq \|u_h(t_0)\|^2 + \frac{1}{\alpha} \int_{t_0}^t \|f(s)\|_{V_h'}^2 \, ds$$

and

$$\|u_h(t)\|^2 \leq e^{-\alpha C_{\text{emb}}^{-2}(t-t_0)} \|u_h(t_0)\|^2 + \frac{1}{\alpha} \int_{t_0}^t e^{-\alpha C_{\text{emb}}^{-2}(t-s)} \|f(s)\|_{V_h'}^2 \, ds.$$

Proof. Suitably adapt the proof of [25, Theorem 6.7, pp. 283–284]. \square

3.2.2 Differentiability with respect to time

Next, we study the differentiability of the semi-discrete solution with respect to time. Obviously, from (3.13) we have that \bar{U}_h is $(j+1)$ -times continuously differentiable, if \tilde{F} is j -times continuously differentiable with respect to t on \bar{I} . The connection to the regularity of the right-hand side f is shown in the following lemma.

Lemma 3.11

Let $j \in \mathbb{Z}$, $j \geq 0$, and suppose that $f \in C^j(\bar{I}, V')$. Then, $\tilde{F} \in C^j(\bar{I}, \mathbb{R}^{\dim(V_h)})$.

Proof. By definition we have $\tilde{F}_i(\cdot) = \langle f(\cdot), \varphi_i \rangle_{V', V}$ for all $i = 1, \dots, \dim(V_h)$. The statement now is proven for each component separately. So, consider an arbitrary $i = 1, \dots, \dim(V_h)$. Obviously, $\langle \cdot, \varphi_i \rangle_{V', V}$ defines a linear functional on V' . Therefore, $f \in C^j(\bar{I}, V')$ implies that

$$\tilde{F}_i(\cdot) = \langle f(\cdot), \varphi_i \rangle_{V', V} \in C^j(\bar{I}, \mathbb{R}),$$

also see [57, beginning of the proof of Proposition 3.6, p. 77]. \square

Now, if $\bar{U}_h \in C^{j+1}(\bar{I}, \mathbb{R}^{\dim(V_h)})$, also the differential equation (3.12) can be differentiated with respect to t and we obtain

$$\bar{U}_h^{(i+1)}(t) + \bar{A} \bar{U}_h^{(i)}(t) = \bar{M} M^{-1} \tilde{F}^{(i)}(t) \quad \forall t \in I, \quad \bar{U}_h^{(i)}(t_0) = \bar{U}_h^{(i)}(t_0^+),$$

for $i = 0, \dots, j$, where we used that \bar{A} and \bar{M} are independent of time t . Analogously, in function representation, we have that $u_h \in C^{j+1}(\bar{I}, V_h)$ with $u_h(t_0) = u_{h,0}$ satisfies

$$\begin{aligned} (u_h^{(i+1)}(t), v_h) + a(u_h^{(i)}(t), v_h) &= \langle f^{(i)}(t), v_h \rangle_{V', V} \quad \forall t \in \bar{I}, \forall v_h \in V_h, \\ u_h^{(i)}(t_0) &= u_h^{(i)}(t_0^+), \end{aligned} \quad (3.15)$$

for $i = 0, \dots, j$.

Since $\bar{U}_h^{(i)}$ and $u_h^{(i)}$ satisfy quite similar initial value problems as \bar{U}_h and u_h , there also hold analog stability estimates, cf. Lemma 3.9 or Lemma 3.10, respectively.

Corollary 3.12

Let $j \in \mathbb{Z}$, $j \geq 0$, and suppose that $f \in C^j(\bar{I}, V')$. Then, the solution \bar{U}_h of (3.12) and the solution u_h of (3.8) satisfy for $i = 0, \dots, j$ the stability estimates

$$\|\bar{U}_h^{(i)}(t)\| \leq e^{-\alpha C_{\text{emb}}^{-2}(t-t_0)} \|\bar{U}_h^{(i)}(t_0^+)\| + \int_{t_0}^t e^{-\alpha C_{\text{emb}}^{-2}(t-s)} \|\bar{M} M^{-1} \tilde{F}^{(i)}(s)\| \, ds$$

or (in function representation)

$$\|u_h^{(i)}(t)\|^2 \leq e^{-\alpha C_{\text{emb}}^{-2}(t-t_0)} \|u_h^{(i)}(t_0^+)\|^2 + \frac{1}{\alpha} \int_{t_0}^t e^{-\alpha C_{\text{emb}}^{-2}(t-s)} \|f^{(i)}(s)\|_{V_h'}^2 \, ds,$$

respectively.

While the latter (integral) terms in the stability estimates can always be bounded by terms of the given data independent of h , appropriate uniform bounds for the initial value(s) can only be guaranteed if $u_{h,0}$ (and so $\bar{U}_{h,0}$) is properly chosen. For more details on this topic we refer to Subsection 4.2.3.

3.2.3 Error estimates for the semi-discrete approximation

The error analysis for the semi-discretization in space is well understood, see e.g. [25, Theorem 6.14, pp. 287–288] or [34, pp. 324–326]. Therefore, we shall only sketch the derivation of error estimates and concentrate on the results.

It is convenient to introduce another spatial projection operator. As before, let $V_h \subset V$ be a finite dimensional subspace of V . We define $R_h : V \rightarrow V_h$ to be the Ritz projection operator given by

$$a(R_h v, w) = a(v, w) \quad \forall w \in V_h.$$

Note that R_h is stable in $\|\cdot\|_V$, i.e., it holds $\|R_h v\|_V \leq C \|v\|_V$ for all $v \in V$. This can be easily derived from the V -ellipticity and continuity of $a(\cdot, \cdot)$. Indeed, from (3.2) we get

$$\alpha \|R_h v\|_V^2 \leq a(R_h v, R_h v) = a(v, R_h v) \leq C_a \|R_h v\|_V \|v\|_V,$$

which yields $\|R_h v\|_V \leq \frac{C_a}{\alpha} \|v\|_V$ for all $v \in V$.

The Ritz projection can be extended to functions of space and time in the L^2 -sense by setting $(R_h v)(t) := R_h(v(t))$ for all $v \in L^2(J, V)$. Then, $R_h : L^2(J, V) \rightarrow L^2(J, V_h)$ satisfies

$$\int_J a(R_h v, w) \, dt = \int_J a(v, w) \, dt \quad \forall w \in L^2(J, V_h).$$

Similarly, the projection operator $P_h : V' \rightarrow V_h$ of (3.9) can also be extended by setting $(P_h v)(t) := P_h(v(t))$ for all $v \in L^2(J, V')$ such that $P_h : L^2(J, V') \rightarrow L^2(J, V_h)$ then satisfies

$$\int_J \langle P_h v, w \rangle_{V', V} \, dt = \int_J \langle v, w \rangle_{V', V} \, dt \quad \forall w \in L^2(J, V_h).$$

These extended projections will be needed in the later error analysis.

Standard spatial discretization

In the setting of model problem (3.4) the discrete space $V_h \subset V$ is often chosen as a finite element space of continuous piecewise polynomials of a certain order, say $\kappa \geq 1$. This space is based on a triangulation \mathcal{T}_h of Ω , for example in simplices. Here, h is not only used as abstract parameter for notation purposes but also denotes the maximum among the diameters of mesh cells contained in the triangulation \mathcal{T}_h .

Under some standard assumptions on the triangulation it is well-known that R_h has the following approximation properties, see e.g. [21, Theorem 3.2.2, p. 134, Theorem 3.2.5, pp. 138–139]. If $v \in H_0^1(\Omega) \cap H^q(\Omega)$, then

$$\|v - R_h v\|_{H^1(\Omega)} \leq Ch^{q-1} \|v\|_{H^q(\Omega)} \quad (3.16a)$$

for $1 \leq q \leq \kappa + 1$. If in addition the associated stationary problem is H^2 -regular, we also have

$$\|v - R_h v\|_{L^2(\Omega)} \leq Ch^q \|v\|_{H^q(\Omega)}, \quad (3.16b)$$

which, compared to (3.16a), provides an improved L^2 -norm error estimate. \clubsuit

The following (abstract) error estimates can be shown.

Theorem 3.13

Provided that u and f are sufficiently smooth, it holds

$$\begin{aligned} \|u^{(i)}(t) - u_h^{(i)}(t)\| &\leq \|u^{(i)}(t) - R_h u^{(i)}(t)\| + \|R_h u^{(i)}(t_0^+) - u_h^{(i)}(t_0^+)\| e^{-\alpha C_{\text{emb}}^{-2}(t-t_0)/2} \\ &\quad + \frac{1}{\sqrt{\alpha}} \left(\int_{t_0}^t e^{-\alpha C_{\text{emb}}^{-2}(t-s)} \|\partial_t(u^{(i)} - R_h u^{(i)})(s)\|_{V'}^2 ds \right)^{1/2} \end{aligned}$$

and

$$\begin{aligned} \|u^{(i)}(t) - u_h^{(i)}(t)\| &\leq \|u^{(i)}(t) - R_h u^{(i)}(t)\| + \|R_h u^{(i)}(t_0^+) - u_h^{(i)}(t_0^+)\| e^{-\alpha C_{\text{emb}}^{-2}(t-t_0)} \\ &\quad + \int_{t_0}^t e^{-\alpha C_{\text{emb}}^{-2}(t-s)} \|\partial_t(u^{(i)} - R_h u^{(i)})(s)\| ds. \end{aligned}$$

Proof. Since the arguments are quite analog, we give a detailed proof for $i = 0$ only.

For estimation the error is split as follows $u - u_h = (u - R_h u) + (R_h u - u_h)$. Now, using the definition of R_h and subtracting (3.8) from (3.3), we find

$$\begin{aligned} \langle \partial_t(R_h u - u_h)(t), v_h \rangle_{V', V} + a((R_h u - u_h)(t), v_h) &= -\langle \partial_t(u - R_h u)(t), v_h \rangle_{V', V} \\ &\quad \forall t \in \bar{I}, \forall v_h \in V_h. \end{aligned}$$

Therefore, the stability estimates, cf. (3.14) and Lemma 3.10, also give bounds for $R_h u - u_h$ (in certain norms) when f is replaced by $-\partial_t(u - R_h u)$. Because of

$$\|(u - u_h)(t)\| \leq \|(u - R_h u)(t)\| + \|(R_h u - u_h)(t)\|,$$

the desired results are proven easily.

If u is sufficiently smooth, similar arguments can also be used to prove the desired statement for $i \geq 1$. Here, note that an identity similar to (3.3) also holds for $u^{(i)}$ with f replaced by $f^{(i)}$ and, moreover, that (3.15) can be applied instead of (3.8). \square

3.3 Full discretization in space and time

After semi-discretizing (3.3) in space according to Section 3.2, we are still faced with differential equations. Therefore, in order to obtain a fully computable discrete scheme, further discretization of the remaining system of coupled odes is needed. Here, it should be noted that the system of ordinary differential equations becomes larger and also stiffer if the spatial discretization gets finer. For this reason, careful consideration should be given to the choice of the temporal discretization.

In the following, we shall apply and analyze the variational time discretization methods presented in Part I. Our previous findings suggest that these methods are well-suited in this context since they provide suitable stability properties (at least A -stability) and enable a proper error analysis also in the case of stiff problems.

3.3.1 Formulation of the methods

First of all, the variational time discretization (**VTD**) methods of higher smoothness as introduced in Chapter 1 are formulated also in the setting of parabolic problems. To this end, we again use a time mesh

$$t_0 < t_1 < \cdots < t_{N-1} < t_N = t_0 + T.$$

Also recall the associated notation, e.g., we write I_n for time mesh intervals and τ_n for the time mesh interval lengths, see p. 7 for details.

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, be given. Then, the local problem on I_n , $1 \leq n \leq N$, reads:

Find $u_{\tau h}|_{I_n} \in P_r(I_n, V_h)$ such that

$$u_{\tau h}(t_{n-1}^+) = u_{\tau h}(t_{n-1}^-), \quad \text{if } k \geq 1, \quad (3.17a)$$

$$(\partial_t^{i+1} u_{\tau h}(t_n^-), v_h) + a(\partial_t^i u_{\tau h}(t_n^-), v_h) = \langle g^{(i)}(t_n^-), v_h \rangle_{V', V} \quad \forall v_h \in V_h, \quad (3.17b)$$

$$\begin{aligned} & \text{if } k \geq 2, i = 0, \dots, \left\lfloor \frac{k}{2} \right\rfloor - 1, \\ (\partial_t^{i+1} u_{\tau h}(t_{n-1}^+), v_h) + a(\partial_t^i u_{\tau h}(t_{n-1}^+), v_h) &= \langle g^{(i)}(t_{n-1}^+), v_h \rangle_{V', V} \quad \forall v_h \in V_h, \quad (3.17c) \\ & \text{if } k \geq 3, i = 0, \dots, \left\lfloor \frac{k-1}{2} \right\rfloor - 1, \end{aligned}$$

and

$$\begin{aligned} \int_{I_n} (\partial_t u_{\tau h}, v_{\tau h}) + a(u_{\tau h}, v_{\tau h}) dt + \delta_{0,k}([u_{\tau h}]_{n-1}, v_{\tau h}(t_{n-1}^+)) &= \int_{I_n} \langle g, v_{\tau h} \rangle_{V', V} dt \\ &\forall v_{\tau h} \in P_{r-k}(I_n, V_h), \quad (3.17d) \end{aligned}$$

where the initial value $u_{\tau h}(t_0^-) \in V_h$ should be a suitable approximation of $u(t_0) = u_0$. Moreover, g is some approximation of f .

In the following, we mainly consider $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\}$ and, if $k \geq 2$, also $g = \mathcal{I}_{k-2,*}^r f$. By these choices we are already able to model the exactly integrated version, the numerically integrated version (with quadrature rule Q_k^r), and the version with cascadic interpolated right-hand side of the \mathbf{VTD}_k^r method. Moreover, for $k \geq 2$ we can consider the situation after a postprocessing of $Q_{k-2}^{r-1} \mathbf{VTD}_{k-2}^{r-1}(f)$. Also cf. Remark 1.43 and Remark 2.14.

Note that the integral on the left-hand side of (3.17d) can always be replaced by any quadrature formula that is exact for polynomials of maximal degree $2r - k$. The same applies to the integral on the right-hand side if $g \in P_r(I_n, V')$.

3.3.2 Reformulation and solvability

The fully discrete method can be rewritten using the same ideas as in Subsection 3.2.1. So, recalling that $\{\boldsymbol{\varphi}_i\}_{i=1, \dots, \dim(V_h)}$ denotes a basis of V_h , for any $n = 1, \dots, N$ we can write $u_{\tau h}|_{I_n} \in P_r(I_n, V_h)$ as

$$u_{\tau h}(t) = \sum_{i=1}^{\dim(V_h)} U_{\tau h, i}(t) \boldsymbol{\varphi}_i \quad \forall t \in I_n$$

with $U_{\tau h, i} \in P_r(I_n)$. Then, the initial condition (3.17a) and the collocation conditions (3.17b) and (3.17c) can be reformulated as

$$\begin{aligned} U_{\tau h}(t_{n-1}^+) &= U_{\tau h}(t_{n-1}^-), & \text{if } k \geq 1, \\ MU_{\tau h}^{(i+1)}(t_n^-) + AU_{\tau h}^{(i)}(t_n^-) &= \tilde{G}^{(i)}(t_n^-), & \text{if } k \geq 2, i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1, \\ MU_{\tau h}^{(i+1)}(t_{n-1}^+) + AU_{\tau h}^{(i)}(t_{n-1}^+) &= \tilde{G}^{(i)}(t_{n-1}^+), & \text{if } k \geq 3, i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor - 1, \end{aligned}$$

with mass matrix M and stiffness matrix A given by (3.10) and right-hand side term \tilde{G} determined by $\tilde{G}_j(t) := \langle g(t), \boldsymbol{\varphi}_j \rangle_{V', V}$ for all $j = 1, \dots, \dim(V_h)$.

Similarly, the variational condition (3.17d) alternatively reads

$$\int_{I_n} (MU'_{\tau h} + AU_{\tau h}, V_{\tau h}) dt + \delta_{0, k}(M[U_{\tau h}]_{n-1}, V_{\tau h}(t_{n-1}^+)) = \int_{I_n} (\tilde{G}, V_{\tau h}) dt$$

$$\forall V_{\tau h} \in P_{r-k}(I_n, \mathbb{R}^{\dim(V_h)}).$$

Here, we also used that $v_{\tau h} \in P_{r-k}(I_n, V_h)$ can be represented by $V_{\tau h} \in P_{r-k}(I_n, \mathbb{R}^{\dim(V_h)})$ via

$$v_{\tau h}(t) = \sum_{j=1}^{\dim(V_h)} V_{\tau h, j}(t) \boldsymbol{\varphi}_j.$$

In this reformulated representation, it becomes obvious that the full discretization (3.17) of (3.3) can be viewed as $\mathbf{VTD}_k^r(\tilde{G})$ approximation (in the style of (1.22)) to the semi-discrete problem (3.11). Therefore, especially the findings of Part I on the solvability can be easily transferred. More concrete, from Proposition 2.24 and due to Lemma 3.6 we have that the fully discrete problem (3.17) is uniquely solvable, where no restriction on the time step length τ_n is needed.

4 Error Analysis for VTD Methods

In this chapter, we want to present an error analysis for variational time discretization (**VTD**) methods of higher smoothness as introduced in Chapter 1 in the setting of parabolic problems. To this end, we combine variational techniques as usually used in the error analysis of discontinuous Galerkin (dG) and continuous Galerkin–Petrov (cGP) methods and techniques that are known from the stiff error analysis of Runge–Kutta(-like) methods (cf. Section 2.3). As byproduct we give a variational error analysis capturing both cGP and dG time stepping methods. In the following, we assume that u as well as f and g are smooth enough to guarantee that the occurring terms are well-defined.

Let $u_{\tau h}$ denote the solution of the **VTD** $_k^r(g)$ method as given in (3.17) where $r, k \in \mathbb{Z}$ with $0 \leq k \leq r$. For our analysis we assume that g is at least globally $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable. Moreover, we choose and include the initial condition in a very special manner. More detailed, the initial values $\partial_t^i u_{\tau h}(t_0^-) \in V_h$, $i = \lfloor \frac{k}{2} \rfloor, \dots, 0$, are determined by

$$\begin{aligned} \partial_t^{\lfloor \frac{k}{2} \rfloor} u_{\tau h}(t_0^-) &:= \tilde{P}_h^0 \partial_t^{\lfloor \frac{k}{2} \rfloor} u_0, \\ \partial_t^i u_{\tau h}(t_0^-) &\in V_h \text{ with } i = \lfloor \frac{k}{2} \rfloor - 1, \dots, 0 : \\ a(\partial_t^i u_{\tau h}(t_0^-), v_h) &= \langle g^{(i)}(t_0^+), v_h \rangle_{V', V} - (\partial_t^{i+1} u_{\tau h}(t_0^-), v_h) \quad \forall v_h \in V_h, \end{aligned} \quad (4.1)$$

with $\tilde{P}_h^0 \in \{R_h, P_h\}$. Here, R_h is as before the Ritz projection and P_h is the projection of (3.9), which is some generalization of the (global) L^2 -projection onto V_h . For a definition of $\partial_t^i u_0$, $i \geq 0$, see (3.7).

In view of Subsection 1.1.1 the assumptions on the smoothness of g and the choice of the initial value(s) ensure that (3.17a) and (3.17c) could be replaced by

$$\partial_t^i u_{\tau h}(t_{n-1}^+) = \partial_t^i u_{\tau h}(t_{n-1}^-) \quad \forall i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor. \quad (4.2)$$

It follows that $u_{\tau h}$ is globally $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable for $k \geq 1$ and, in general, discontinuous at the time (mesh) points for $k = 0$. Therefore, we are interested whether the results of Subsection 1.4.4 can be transferred to the present setting.

Lemma 4.1

Let $0 \leq j \leq \lfloor \frac{k}{2} \rfloor$ and assume that g is $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable on \bar{I} . Moreover, suppose that $u_{\tau h}$ satisfies (3.17) with initial value determined by (4.1). Then, it holds for $1 \leq n \leq N$

$$\begin{aligned} \int_{I_n} (\partial_t u_{\tau h}^{(j)}, v_{\tau h}) + a(u_{\tau h}^{(j)}, v_{\tau h}) \, dt + \delta_{0, k-2j}([u_{\tau h}^{(j)}]_{n-1}, v_{\tau h}(t_{n-1}^+)) &= \int_{I_n} \langle g^{(j)}, v_{\tau h} \rangle_{V', V} \, dt \\ &\quad \forall v_{\tau h} \in P_{r-k+j}(I_n, V_h). \end{aligned}$$

Proof. For $j = 0$ the statement is obvious. So we only consider the case $j \geq 1$, which also directly implies that $k \geq 2$ and so there is no jump term in (3.17d).

Let $v_{\tau h} \in P_{r-k+j}(I_n, V_h)$. Integrating by parts j times in time, we obtain

$$\begin{aligned} \int_{I_n} (\partial_t u_{\tau h}^{(j)}, v_{\tau h}) + a(u_{\tau h}^{(j)}, v_{\tau h}) dt &= \int_{I_n} (\partial_t^{j+1} u_{\tau h}, v_{\tau h}) + a(\partial_t^j u_{\tau h}, v_{\tau h}) dt \\ &= - \int_{I_n} (\partial_t^j u_{\tau h}, v'_{\tau h}) + a(\partial_t^{j-1} u_{\tau h}, v'_{\tau h}) dt + [(\partial_t^j u_{\tau h}, v_{\tau h}) + a(\partial_t^{j-1} u_{\tau h}, v_{\tau h})] \Big|_{t_{n-1}^+}^{t_n^-} \\ &= \dots \\ &= (-1)^j \int_{I_n} (\partial_t u_{\tau h}, v_{\tau h}^{(j)}) + a(u_{\tau h}, v_{\tau h}^{(j)}) dt \\ &\quad + \sum_{l=0}^{j-1} (-1)^l [(\partial_t^{j-l} u_{\tau h}, v_{\tau h}^{(l)}) + a(\partial_t^{j-1-l} u_{\tau h}, v_{\tau h}^{(l)})] \Big|_{t_{n-1}^+}^{t_n^-}. \end{aligned}$$

Because of (3.17b), (3.17c), and (3.17d), we gain

$$\begin{aligned} \int_{I_n} (\partial_t u_{\tau h}^{(j)}, v_{\tau h}) + a(u_{\tau h}^{(j)}, v_{\tau h}) dt &= (-1)^j \int_{I_n} \langle g, v_{\tau h}^{(j)} \rangle_{V', V} dt + \sum_{l=0}^{j-1} (-1)^l [\langle g^{(j-1-l)}, v_{\tau h}^{(l)} \rangle_{V', V}] \Big|_{t_{n-1}^+}^{t_n^-} \\ &\quad + \langle g^{(j-1)}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+) \rangle_{V', V} - (\partial_t^j u_{\tau h}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+)) - a(\partial_t^{j-1} u_{\tau h}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+)). \end{aligned} \quad (4.3)$$

Note that we did not rewrite the term for $l = 0$ at t_{n-1}^+ but only added the auxiliary term $\langle g^{(j-1)}(t_{n-1}^+) - g^{(j-1)}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+) \rangle_{V', V} = 0$ since (3.17c) does not apply in the case $j = \lfloor \frac{k}{2} \rfloor > \lfloor \frac{k-1}{2} \rfloor$. Now, again using integration by parts j times, the first line of the right-hand side of (4.3) can be rewritten as

$$(-1)^j \int_{I_n} \langle g, v_{\tau h}^{(j)} \rangle_{V', V} dt + \sum_{l=0}^{j-1} (-1)^l [\langle g^{(j-1-l)}, v_{\tau h}^{(l)} \rangle_{V', V}] \Big|_{t_{n-1}^+}^{t_n^-} = \int_{I_n} \langle g^{(j)}, v_{\tau h} \rangle_{V', V} dt.$$

It remains to study the second line of the right-hand side of (4.3). Since $g^{(j-1)}$ is globally continuous, we get from (3.17b) (if $n \geq 1$) or the definition of the initial values (4.1) (if $n = 0$) that

$$\begin{aligned} \langle g^{(j-1)}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+) \rangle_{V', V} &= \langle g^{(j-1)}(t_{n-1}^-), v_{\tau h}(t_{n-1}^+) \rangle_{V', V} \\ &= (\partial_t^j u_{\tau h}(t_{n-1}^-), v_{\tau h}(t_{n-1}^+)) + a(\partial_t^{j-1} u_{\tau h}(t_{n-1}^-), v_{\tau h}(t_{n-1}^+)). \end{aligned}$$

Because of (4.2), we have that $u_{\tau h}$ is globally $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable and, thus, especially $u_{\tau h}^{(j-1)}$ is globally continuous. Therefore, we conclude

$$\begin{aligned} \langle g^{(j-1)}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+) \rangle_{V', V} - (\partial_t^j u_{\tau h}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+)) - a(\partial_t^{j-1} u_{\tau h}(t_{n-1}^+), v_{\tau h}(t_{n-1}^+)) \\ = -\delta_{0, k-2j}([u_{\tau h}^{(j)}]_{n-1}, v_{\tau h}(t_{n-1}^+)). \end{aligned}$$

Combining the above identities, we are done. \square

In order to guarantee that Lemma 4.1 is always applicable and that we do not need to know about g when defining the discrete initial values, cf. (4.1), we suppose that from now on the following assumption holds true.

Assumption

We assume that f and g are $(\lfloor \frac{k}{2} \rfloor - 1)$ -times continuously differentiable on \bar{I} . Moreover, we suppose that

$$g^{(i)}(t_0^+) = f^{(i)}(t_0^+) \quad \text{for all } i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1.$$

In the following, we always set $\ell := \lfloor \frac{k}{2} \rfloor$. Using the preceding lemma and provided sufficiently smooth data, we see that the ℓ th derivative $w_{\tau h} = u_{\tau h}^{(\ell)}$ of $u_{\tau h}$ solves on I_n , $1 \leq n \leq N$, the local problem:

Find $w_{\tau h}|_{I_n} \in P_{r-\ell}(I_n, V_h)$ such that

$$w_{\tau h}(t_{n-1}^+) = w_{\tau h}(t_{n-1}^-), \quad \text{if } k \text{ is odd,} \quad (4.4a)$$

and

$$\begin{aligned} \int_{I_n} (\partial_t w_{\tau h}, v_{\tau h}) + a(w_{\tau h}, v_{\tau h}) \, dt + \delta_{0, k-2\ell}([w_{\tau h}]_{n-1}, v_{\tau h}(t_{n-1}^+)) &= \int_{I_n} \langle g^{(\ell)}, v_{\tau h} \rangle_{V', V} \, dt \\ \forall v_{\tau h} \in P_{r-k+\ell}(I_n, V_h), \end{aligned} \quad (4.4b)$$

where $w_{\tau h}(t_0^-)$ is given by

$$w_{\tau h}(t_0^-) = \tilde{P}_h^0 \partial_t^\ell u_0.$$

When also u is sufficiently smooth (e.g., $u \in C^{\ell+1}([t_0, t_0 + T], V)$), then the function $w = u^{(\ell)}$ satisfies

$$w(t_{n-1}^+) = w(t_{n-1}^-),$$

and

$$\int_{I_n} (\partial_t w, v) + a(w, v) \, dt = \int_{I_n} \langle f^{(\ell)}, v \rangle_{V', V} \, dt \quad \forall v \in L^2(I_n, V),$$

where $w(t_0^-)$ is given by

$$w(t_0^-) = \partial_t^\ell u_0.$$

Comparing the initial values for the continuous and the discrete problem, we see that

$$w_{\tau h}(t_0^-) = \tilde{P}_h^0 \partial_t^\ell u_0 = \tilde{P}_h^0 w(t_0^-),$$

where $\tilde{P}_h^0 \in \{R_h, P_h\}$.

4.1 Error estimates for the ℓ th derivative

Recalling the above observations and that $\ell = \lfloor \frac{k}{2} \rfloor$, we conclude that $w_{\tau h} = u_{\tau h}^{(\ell)}$ is the solution of a $\mathbf{VTD}_{k-2\ell}^{r-\ell}$ method with adapted initial value applied to the modified problem (cf. (3.1))

$$\begin{aligned} \partial_t w(t) + \mathcal{A} w(t) &= f^{(\ell)}(t) && \text{in } \Omega, \quad t_0 < t < t_0 + T, \\ \mathcal{B} w(t) &= 0 && \text{on } \partial\Omega, \quad t_0 < t < t_0 + T, \\ w(t_0) &= \partial_t^\ell u_0 && \text{in } \Omega, \end{aligned}$$

which is solved by $w = u^{(\ell)}$. If $k \geq 1$ is odd, it holds $k - 2\ell = 1$ and so the discrete problem is that of a cGP method, whereas for $k \geq 0$ even it holds $k - 2\ell = 0$, which implies that $u_{\tau h}^{(\ell)}$ is solution of a dG method.

Therefore, for the derivation of error estimates for the ℓ th derivative of \mathbf{VTD}_k^r methods we can build on the broad knowledge for the analysis of dG and cGP methods. We, however, shall present a unified analysis for the global L^2 -error in the H -norm. Moreover, global L^2 -error estimates in the V -norm, pointwise error estimates in the H -norm, and some supercloseness results are derived. Because of the usage of g as approximation of f on the right-hand side of (3.17), we can easily study various variants of the method in one. Furthermore, in order to gain even more flexibility, (especially) for the analysis, we consider an integrator \mathcal{I}_n that satisfies the following assumption.

Assumption

We assume that the integrator \mathcal{I}_n either represents the exact integral over I_n , i.e., $\mathcal{I}_n = \int_{I_n}$, or the application of a quadrature formula based on function values of the integrand in \bar{I}_n that is exact for polynomials of maximal degree $2r - k$ and has positive weights only.

Quadrature formulas that fulfill this assumption are, for example, the Gauss–Legendre, the Gauss–Radau, or the Gauss–Lobatto quadrature rules with sufficiently high number of quadrature points. We will typically use $\mathcal{I}_n = \int_{I_n}$ or $\mathcal{I}_n = Q_{k-2\ell,n}^{r-\ell}$.

By assumption the integrator \mathcal{I}_n is exact for polynomials up to degree $2r - k$, i.e.,

$$\mathcal{I}_n[v] = \int_{I_n} v(t) \, dt \quad \forall v \in P_{2r-k}(I_n, \mathbb{R}). \quad (4.5a)$$

Moreover, because of the positive weights in case of quadrature, we have that

$$\mathcal{I}_n[v] \leq \mathcal{I}_n[w] \quad \forall v, w : \bar{I}_n \rightarrow \mathbb{R} \quad \text{with} \quad v(t) \leq w(t) \, \forall t \in \bar{I}_n, \quad (4.5b)$$

which also implies $|\mathcal{I}_n[v]| \leq \mathcal{I}_n[|v|]$, and that the Cauchy–Schwarz inequality also holds for \mathcal{I}_n , i.e.,

$$\mathcal{I}_n[vw] \leq (\mathcal{I}_n[v^2])^{1/2} (\mathcal{I}_n[w^2])^{1/2} \quad \forall v, w : \bar{I}_n \rightarrow \mathbb{R}. \quad (4.5c)$$

Here, in (4.5b) and (4.5c) we tacitly assume that for v and w all occurring expressions are well-defined.

Of course, depending on the concrete choice of \mathcal{J}_n , the integrands need to satisfy different conditions. Therefore, similar to Part I, we set $k_{\mathcal{J}} = 0$ if \mathcal{J}_n represents a quadrature formula based on function values and so requires integrands that are continuous on \bar{I}_n . For the case $\mathcal{J}_n = \int_{I_n}$, which requires integrable integrands only, we set $k_{\mathcal{J}} = -1$.

The integrator \mathcal{J}_n also can be well interpreted for Banach space-valued functions. Indeed, if $\mathcal{J}_n = \int_{I_n}$, the integral is read in Bochner sense. Otherwise, if \mathcal{J}_n is a quadrature formula, we just have a weighted sum of function values, which also makes sense in Banach spaces. So, denoting by X a Banach space over \mathbb{R} , we have that \mathcal{J}_n is a bounded linear operator from $L^1(I_n, X)$ to X if $k_{\mathcal{J}} = -1$ or from $C(\bar{I}_n, X)$ to X if $k_{\mathcal{J}} = 0$, respectively.

Note that the integral in (4.4b) can be replaced by an integrator \mathcal{J}_n satisfying (4.5a) if $g^{(\ell)}|_{I_n} \in P_{r-\ell}(I_n, V')$ for all $n = 1, \dots, N$. The latter can always be achieved since in (3.17) we can use $\Pi_k^r g$, cf. (1.28), instead of g without changing the discrete solution.

For sufficiently smooth functions v and w define a bilinear form by

$$B_n^{\mathcal{J}}(v, w) := \mathcal{J}_n[(\partial_t v, w) + a(v, w)] + \delta_{0, k-2\ell}([v]_{n-1}, w(t_{n-1}^+)).$$

Then, for all $n = 1, \dots, N$ we have that $u_{\tau h}^{(\ell)}|_{I_n} \in P_{r-\ell}(I_n, V_h)$ satisfies

$$u_{\tau h}^{(\ell)}(t_{n-1}^+) = u_{\tau h}^{(\ell)}(t_{n-1}^-) \in V_h, \quad \text{if } k - 2\ell = 1 \ (\Leftrightarrow k \text{ is odd}), \quad (4.6a)$$

and

$$B_n^{\mathcal{J}}(u_{\tau h}^{(\ell)}, v_{\tau h}) = \mathcal{J}_n[\langle g^{(\ell)}, v_{\tau h} \rangle_{V', V}] \quad \forall v_{\tau h} \in P_{r-k+\ell}(I_n, V_h), \quad (4.6b)$$

where $u_{\tau h}^{(\ell)}(t_0^-) := \tilde{P}_h^0 \partial_t^\ell u_0$ with $\tilde{P}_h^0 \in \{R_h, P_h\}$.

4.1.1 Projection operators

To prepare the error analysis, we need to define some projection operators with respect to time. For generality the projections are defined for X -valued functions where X denotes some Banach space over \mathbb{R} . Note that we directly give the (local) operator definitions on I_n for the concrete polynomial degrees (depending on r and k) that are actually needed in the later argumentation. For stand-alone definitions of the operators and the study of their well-definedness see Appendix C.2.

First, for $v \in L^2(I_n, X)$ let $\Pi_{r-k+\ell} v \in P_{r-k+\ell}(I_n, X)$ denote the (local) L^2 -projection onto polynomials of maximal degree $r - k + \ell$, i.e.,

$$\int_{I_n} (v - \Pi_{r-k+\ell} v) w \, dt = 0 \quad \forall w \in P_{r-k+\ell}(I_n),$$

cf. Definition C.4. The integral here needs to be understood in Bochner sense and the 0 on the right-hand side should be read as the zero element in X .

For the case where \mathcal{J}_n is not just the integral over I_n , we also define an analog projection with respect to the integrator \mathcal{J}_n , i.e., for $v \in C^{k_{\mathcal{J}}}(\bar{I}_n, X)$ let $\Pi_{r-k+\ell}^{\mathcal{J}} v \in P_{r-k+\ell}(I_n, X)$ be determined by

$$\mathcal{J}_n[(v - \Pi_{r-k+\ell}^{\mathcal{J}} v) w] = 0 \quad \forall w \in P_{r-k+\ell}(I_n),$$

cf. Definition C.9. Here, we use that the integrator \mathcal{J}_n can be well interpreted also for X -valued functions. Recall that $C^{-1}(\bar{I}_n, X)$ is interpreted as $L^2(\bar{I}_n, X)$.

Finally, there is another projection which is essentially used in the following analysis. For $v \in H^1(I_n, X) \cap C^{k_{\mathcal{J}}+1}(\bar{I}_n, X)$ we define $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v \in P_{r-\ell}(I_n, X)$ by

$$\begin{aligned} (v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)(t_{n-1}^+) &= 0, & \text{if } k - 2\ell = 1, \\ \mathcal{J}_n[\partial_t(v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)w] + \delta_{0, k-2\ell}(v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)(t_{n-1}^+)w(t_{n-1}^+) &= 0 & \forall w \in P_{r-k+\ell}(I_n), \end{aligned}$$

cf. Definition C.10. Note that from the definition of $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ with $w \equiv 1$, we conclude

$$\begin{aligned} (v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)(t_n^-) &= \int_{I_n} \partial_t(v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v) dt + (v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)(t_{n-1}^+) \\ &= \int_{I_n} \partial_t v dt - \mathcal{J}_n[\partial_t v] + \mathcal{J}_n[\partial_t(v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)] + \delta_{0, k-2\ell}(v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)(t_{n-1}^+) \\ &= \int_{I_n} \partial_t v dt - \mathcal{J}_n[\partial_t v] =: \omega_n^{\mathcal{J}}(v), \end{aligned} \tag{4.7}$$

where also the fundamental theorem of calculus and the properties of \mathcal{J}_n were used. Thus, $\omega_n^{\mathcal{J}}(v)$ is an integrator error. For convenience, we set $\omega_0^{\mathcal{J}}(v) := 0$.

Composing the approximations locally defined by $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$, we can define a global approximation. For simplicity, the associated global approximation operator is also denoted by $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$. More concrete, for $v \in \{w \in H^1(I, X) : w|_{I_n} \in C^{k_{\mathcal{J}}+1}(\bar{I}_n, X), n = 1, \dots, N\}$ we set

$$\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v(t_0^-) := v(t_0^-), \quad (\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v)|_{I_n} = \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}(v|_{I_n}), \quad n = 1, \dots, N.$$

Of course, this global approximation operator strongly depends on the time mesh. Note that $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v(t_0^-)$ needs to be defined as $v(t_0^-)$ in order to be consistent with $\omega_0^{\mathcal{J}}(v) = 0$.

Remark 4.2

For $v \in C(\bar{I}_n, X)$ a projection $\tilde{\Pi}_{k-2\ell}^{r-\ell} v \in P_{r-\ell}(I_n, X)$ could also be defined by

$$\begin{aligned} (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v)(t_{n-1}^+) &= 0, & \text{if } k - 2\ell = 1, \\ (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v)(t_n^-) &= 0, \\ \int_{I_n} (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v)w dt &= 0 & \forall w \in P_{r-k+\ell-1}(I_n), \end{aligned}$$

cf. Definition C.6. For $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd) this is the projection that is typically used in the analysis of the Galerkin–Petrov time stepping, see for example [4, (4.12)], [11, (2.7)], and [26, (70.19), p. 202]. If $k - 2\ell = 0$ ($\Leftrightarrow k$ is even), the projection is the standard one in the context of the discontinuous Galerkin time stepping method, see for example [5, (3.1)], [26, (69.26), p. 186], and [52, (12.9), p. 207].

Note that from integration by parts the definition of $\tilde{\Pi}_{k-2\ell}^{r-\ell}$ also implies that for functions

$v \in H^1(I_n, X) \subset C(\bar{I}_n, X)$ it holds

$$\begin{aligned} & \int_{I_n} \partial_t (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v) w \, dt + \delta_{0,k-2\ell} (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v)(t_{n-1}^+) w(t_{n-1}^+) \\ &= - \int_{I_n} (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v) \partial_t w \, dt + (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v) w|_{t_{n-1}^+}^{t_n^-} + \delta_{0,k-2\ell} (v - \tilde{\Pi}_{k-2\ell}^{r-\ell} v)(t_{n-1}^+) w(t_{n-1}^+) \\ &= 0 \end{aligned} \quad \forall w \in P_{r-k+\ell}(I_n). \quad (4.8)$$

So, the projection operator $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ can be viewed as a generalization of $\tilde{\Pi}_{k-2\ell}^{r-\ell}$ for the case where \mathcal{J}_n not simply represents the integration over I_n . \clubsuit

Since $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ preserves polynomials up to degree $r - \ell$, by a standard approach, see Lemma B.9 or also cf. [21, Theorem 3.1.4, p. 121] or [25, Theorem 1.103, p. 59] (where the special case $X = \mathbb{R}$ is handled), we get for all $\max\{0, k_{\mathcal{J}}\} + 2 \leq q \leq r - \ell + 1$ and $0 \leq m \leq q$ that

$$\begin{aligned} |v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v|_{H^m(I_n, X)} &\leq C \tau_n^{q-m} |v|_{H^q(I_n, X)} & \forall v \in H^q(I_n, X), \\ |v - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v|_{W^{m, \infty}(I_n, X)} &\leq C \tau_n^{q-m} |v|_{W^{q, \infty}(I_n, X)} & \forall v \in W^{q, \infty}(I_n, X). \end{aligned} \quad (4.9)$$

In the case of exact integration, i.e., $\mathcal{J}_n = \int_{I_n}$, some of these estimates are already known from the literature, see e.g. [52, (12.10), p. 208] or [26, (69.27), p. 187, (70.20), p. 202].

In the case that $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd) we also need some norm equivalence in finite dimensional spaces for the further analysis. The following lemma is proven later in a more general setting, see Lemma D.2, where we here use that $r - \ell - 1 = r - k + \ell$ if k is odd. Note that, since $V_h \subset V \subset H$ is finite dimensional, both $(V_h, \|\cdot\|_V)$ and $(V_h, \|\cdot\|)$ are (finite dimensional) Hilbert spaces.

Lemma 4.3

Let $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd) and $\|\cdot\|_W \in \{\|\cdot\|_V, \|\cdot\|\}$. Then, the mappings

$$v \mapsto \left(\int_{I_n} \|v(t)\|_W^2 \, dt \right)^{1/2} \quad \text{and} \quad v \mapsto \left(\int_{I_n} \|\Pi_{r-k+\ell} v(t)\|_W^2 \, dt + \left(\frac{\tau_n}{2} \right) \|v(t_n)\|_W^2 \right)^{1/2}$$

define equivalent norms on $P_{r-\ell}(I_n, V_h)$ where the equivalence constants are independent of τ_n and of V_h .

4.1.2 Global L^2 -error in the H -norm

At first, the global L^2 -error in the H -norm of the ℓ th derivative is studied. To this end, we apply standard variational arguments as they are typically used in and known from the analysis of cGP and dG methods. What makes it special, however, is that we study both types of methods in one error analysis. This nicely shows, on the one hand, how similar the arguments are and, on the other hand, where the differences lie.

Moreover, this subsection provides the basis for the following error analysis in different norms since many results can and will be reused as well as many techniques that are used in the proofs can and will be adapted later.

We start to show a quite useful property of the bilinear form $B_n^{\mathcal{J}}(\cdot, \cdot)$, which will allow us to control certain parts of the fully discrete solution and the fully discrete error.

Lemma 4.4

Let $v_\tau \in P_{r-\ell}(I_n, V)$ and $v_\tau(t_{n-1}^-) \in H$ be given. Then,

$$\begin{aligned} & B_n^{\mathcal{J}}(v_\tau, \Pi_{r-k+\ell} v_\tau) + \delta_{1,k-2\ell}([v_\tau]_{n-1}, v_\tau(t_{n-1}^+)) \\ &= \int_{I_n} (\partial_t v_\tau, \Pi_{r-k+\ell} v_\tau) + a(v_\tau, \Pi_{r-k+\ell} v_\tau) dt + ([v_\tau]_{n-1}, v_\tau(t_{n-1}^+)) \\ &\geq \frac{1}{2} \|v_\tau(t_n^-)\|^2 - \frac{1}{2} \|v_\tau(t_{n-1}^-)\|^2 + \frac{1}{2} \|[v_\tau]_{n-1}\|^2 + \alpha \int_{I_n} \|\Pi_{r-k+\ell} v_\tau\|_V^2 dt. \end{aligned}$$

Proof. First of all, from (4.5a) we have that all integral terms in $B_n^{\mathcal{J}}(v_\tau, \Pi_{r-k+\ell} v_\tau)$ are integrated exactly by \mathcal{J}_n . Thus, \mathcal{J}_n can be replaced by the integral over I_n . Moreover, the Kronecker delta term in $B_n^{\mathcal{J}}(\cdot, \cdot)$ only appears if $k = 2\ell$ in which case $\Pi_{r-k+\ell} v_\tau = \Pi_{r-\ell} v_\tau = v_\tau$ in I_n . This shows the desired identity.

In order to derive the lower bound, we note that $\partial_t v_\tau \in P_{r-\ell-1}(I_n, V)$ is a feasible test function for the L^2 -projection $\Pi_{r-k+\ell}$ due to $r - k + \ell = r - \lfloor \frac{k-1}{2} \rfloor - 1 \geq r - \ell - 1$, also cf. Corollary C.14. Therefore, we get by the fundamental theorem of calculus that

$$\int_{I_n} (\partial_t v_\tau, \Pi_{r-k+\ell} v_\tau) dt = \int_{I_n} (\partial_t v_\tau, v_\tau) dt = \frac{1}{2} \int_{I_n} \partial_t \|v_\tau\|^2 dt = \frac{1}{2} (\|v_\tau(t_n^-)\|^2 - \|v_\tau(t_{n-1}^+)\|^2).$$

Further, because of

$$\|v_\tau(t_{n-1}^-)\|^2 = \|v_\tau(t_{n-1}^+) - [v_\tau]_{n-1}\|^2 = \|v_\tau(t_{n-1}^+)\|^2 - 2([v_\tau]_{n-1}, v_\tau(t_{n-1}^+)) + \|[v_\tau]_{n-1}\|^2,$$

we find that

$$-\frac{1}{2} \|v_\tau(t_{n-1}^+)\|^2 + ([v_\tau]_{n-1}, v_\tau(t_{n-1}^+)) = -\frac{1}{2} \|v_\tau(t_{n-1}^-)\|^2 + \frac{1}{2} \|[v_\tau]_{n-1}\|^2.$$

Finally, using that $\Pi_{r-k+\ell} v_\tau \in P_{r-k+\ell}(I_n, V)$ is a feasible test function for $\Pi_{r-k+\ell}$, again also cf. Corollary C.14, and involving the V -ellipticity of $a(\cdot, \cdot)$, we obtain

$$\int_{I_n} a(v_\tau, \Pi_{r-k+\ell} v_\tau) dt = \int_{I_n} a(\Pi_{r-k+\ell} v_\tau, \Pi_{r-k+\ell} v_\tau) dt \geq \alpha \int_{I_n} \|\Pi_{r-k+\ell} v_\tau\|_V^2 dt.$$

Combining the above identities and estimates, we easily gain the desired statement. \square

In order to study the ℓ th derivative of the error $e(t) = u(t) - u_{\tau h}(t)$, we use the following splitting

$$e^{(\ell)}(t) = (u^{(\ell)}(t) - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}(t)) + e_{\tau h, \ell}^{\mathcal{J}}(t) \quad \text{with} \quad e_{\tau h, \ell}^{\mathcal{J}}(t) := R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}(t) - u_{\tau h}^{(\ell)}(t).$$

The identity of the next lemma shows how in the a priori error analysis we can get rid of the fully discrete solution. Moreover, we see that the fully discrete error $e_{\tau h, \ell}^{\mathcal{J}}$ is connected to certain projection and approximation errors.

Lemma 4.5

Let $1 \leq n \leq N$, then for all $v_{\tau h} \in P_{r-k+\ell}(I_n, V_h)$ it holds

$$\begin{aligned} B_n^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, v_{\tau h}) &= -\mathcal{J}_n[(u^{(\ell+1)} - R_h u^{(\ell+1)}, v_{\tau h})] - \mathcal{J}_n[a(u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h})] \\ &\quad + \mathcal{J}_n[\langle f^{(\ell)} - g^{(\ell)}, v_{\tau h} \rangle_{V', V}] + \delta_{0, k-2\ell}(\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}), v_{\tau h}(t_{n-1}^+)) \end{aligned}$$

with $\omega_{n-1}^{\mathcal{J}}(\cdot)$ as defined in (4.7). Moreover, we have that

$$[e_{\tau h, \ell}^{\mathcal{J}}]_{n-1} = \omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}), \quad \text{if } k - 2\ell = 1 \ (\Leftrightarrow k \text{ is odd}).$$

Proof. According to (4.6b), it holds

$$B_n^{\mathcal{J}}(u_{\tau h}^{(\ell)}, v_{\tau h}) = \mathcal{J}_n[\langle g^{(\ell)}, v_{\tau h} \rangle_{V', V}] \quad \forall v_{\tau h} \in P_{r-k+\ell}(I_n, V_h).$$

At the same time, assuming that the exact solution u and the problem data are sufficiently smooth, especially $u^{(\ell)}$ globally continuous, we similarly have

$$B_n^{\mathcal{J}}(u^{(\ell)}, v_{\tau h}) = \mathcal{J}_n[\langle f^{(\ell)}, v_{\tau h} \rangle_{V', V}] \quad \forall v_{\tau h} \in P_{r-k+\ell}(I_n, V_h).$$

Altogether, this implies

$$B_n^{\mathcal{J}}(u^{(\ell)} - u_{\tau h}^{(\ell)}, v_{\tau h}) = \mathcal{J}_n[\langle f^{(\ell)} - g^{(\ell)}, v_{\tau h} \rangle_{V', V}] \quad \forall v_{\tau h} \in P_{r-k+\ell}(I_n, V_h)$$

and, thus, we get

$$\begin{aligned} B_n^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, v_{\tau h}) &= B_n^{\mathcal{J}}(R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}, v_{\tau h}) \\ &= B_n^{\mathcal{J}}(R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u^{(\ell)}, v_{\tau h}) + \mathcal{J}_n[\langle f^{(\ell)} - g^{(\ell)}, v_{\tau h} \rangle_{V', V}] \quad \forall v_{\tau h} \in P_{r-k+\ell}(I_n, V_h). \end{aligned}$$

We now rewrite the first term on the right-hand side. To this end, we first note that because of the (assumed) global continuity of $u^{(\ell)}$ and (4.7) it holds

$$\begin{aligned} [u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}]_{n-1} &= -[R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}]_{n-1} = [R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}]_{n-1} \\ &= (R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})(t_{n-1}^+) - \omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}). \end{aligned} \quad (4.10)$$

From this and using that the spatial projection R_h commutes with the temporal projection $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$, cf. Corollary B.5 and Remark B.6, we find that

$$\begin{aligned} B_n^{\mathcal{J}}(u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h}) &= \mathcal{J}_n[(\partial_t(u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}), v_{\tau h})] + \delta_{0, k-2\ell}([u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}]_{n-1}, v_{\tau h}(t_{n-1}^+)) \\ &\quad + \mathcal{J}_n[a(u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h})] \\ &= \mathcal{J}_n[(\partial_t(u^{(\ell)} - R_h u^{(\ell)}), v_{\tau h})] - \delta_{0, k-2\ell}(\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}), v_{\tau h}(t_{n-1}^+)) \\ &\quad + \mathcal{J}_n[(\partial_t(R_h u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} R_h u^{(\ell)}), v_{\tau h})] + \delta_{0, k-2\ell}((R_h u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} R_h u^{(\ell)})(t_{n-1}^+), v_{\tau h}(t_{n-1}^+)) \\ &\quad + \mathcal{J}_n[a(u^{(\ell)} - R_h u^{(\ell)}, v_{\tau h})] + \mathcal{J}_n[a(R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h})] \end{aligned}$$

for all $v_{\tau h} \in P_{r-k+\ell}(I_n, V_h)$. From the definitions of $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ and R_h the penultimate line as well as the second to last term vanish. Furthermore, it holds

$$\mathcal{J}_n \left[a(R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h}) \right] = \mathcal{J}_n \left[a(u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h}) \right].$$

Hence, collecting and combining the above identities as well as using that the time derivative commutes with the spatial projection R_h , cf. [26, Lemma 64.34, p. 118], the first statement is easily shown.

The second statement follows quite analogously to (4.10). Indeed, if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd), it follows from (4.6a) and (4.7) that

$$[e_{\tau h, \ell}^{\mathcal{J}}]_{n-1} = [R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}]_{n-1} = \overbrace{(R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - R_h u^{(\ell)})(t_{n-1}^+)}^{=0} + \omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}),$$

where the first term on the right-hand side vanishes by definition of $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$. \square

Corollary 4.6

Let $1 \leq n \leq N$, then for all $v_{\tau h} \in P_{r-k+\ell}(I_n, V_h)$ it holds

$$\begin{aligned} B_n^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, v_{\tau h}) &\leq \left[C_{\text{emb}} \left(\mathcal{J}_n \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right)^{1/2} + C_a \left(\mathcal{J}_n \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \right)^{1/2} \right. \\ &\quad \left. + \left(\mathcal{J}_n \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right)^{1/2} \right] \left(\int_{I_n} \|v_{\tau h}\|_V^2 dt \right)^{1/2} + \delta_{0, k-2\ell} \|\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|v_{\tau h}(t_{n-1}^+)\| \end{aligned}$$

with $\omega_{n-1}^{\mathcal{J}}(\cdot)$ as defined in (4.7). Moreover, if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd), we have for all $w \in H$

$$([e_{\tau h, \ell}^{\mathcal{J}}]_{n-1}, w) \leq \|\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|w\|.$$

Proof. From Lemma 4.5 and the definition of $\Pi_{r-k+\ell}^{\mathcal{J}}$ we get for all $v_{\tau h} \in P_{r-k+\ell}(I_n, V_h)$ that

$$\begin{aligned} B_n^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, v_{\tau h}) &= -\mathcal{J}_n \left[(u^{(\ell+1)} - R_h u^{(\ell+1)}), v_{\tau h} \right] - \mathcal{J}_n \left[a(u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h}) \right] \\ &\quad + \mathcal{J}_n \left[\langle \Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)}), v_{\tau h} \rangle_{V', V} \right] + \delta_{0, k-2\ell} (\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}), v_{\tau h}(t_{n-1}^+)) \\ &\leq \mathcal{J}_n \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\| \|v_{\tau h}\| \right] + C_a \mathcal{J}_n \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V \|v_{\tau h}\|_V \right] \\ &\quad + \mathcal{J}_n \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'} \|v_{\tau h}\|_V \right] + \delta_{0, k-2\ell} \|\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|v_{\tau h}(t_{n-1}^+)\|, \end{aligned}$$

where we also used the properties (4.5) of \mathcal{J}_n , the Cauchy–Schwarz inequality, the continuity of $a(\cdot, \cdot)$, and the definition of the norm in V' . Because of $V \hookrightarrow H$, we furthermore have $\|v_{\tau h}\| \leq C_{\text{emb}} \|v_{\tau h}\|_V$. So, applying the Cauchy–Schwarz-type inequality for \mathcal{J}_n , we conclude

$$\begin{aligned} \mathcal{J}_n \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\| \|v_{\tau h}\| \right] &\leq C_{\text{emb}} \mathcal{J}_n \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\| \|v_{\tau h}\|_V \right] \\ &\leq C_{\text{emb}} \left(\mathcal{J}_n \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right)^{1/2} \left(\mathcal{J}_n \left[\|v_{\tau h}\|_V^2 \right] \right)^{1/2}, \\ \mathcal{J}_n \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V \|v_{\tau h}\|_V \right] &\leq \left(\mathcal{J}_n \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \right)^{1/2} \left(\mathcal{J}_n \left[\|v_{\tau h}\|_V^2 \right] \right)^{1/2}, \end{aligned}$$

and

$$\mathcal{J}_n \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'} \|v_{\tau h}\|_V \right] \leq \left(\mathcal{J}_n \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right)^{1/2} \left(\mathcal{J}_n \left[\|v_{\tau h}\|_V^2 \right] \right)^{1/2}.$$

Summarizing and using that \mathcal{J}_n is exact for $\|v_{\tau h}\|_V^2 = (v_{\tau h}, v_{\tau h})_V \in P_{2(r-k+\ell)}(I_n)$ due to $2(r-k+\ell) = 2r-k-(k-2\ell) \leq 2r-k$, we easily get the first desired estimate.

Since according to Lemma 4.5 it holds $[e_{\tau h, \ell}^{\mathcal{J}}]_{n-1} = \omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)})$ if $k-2\ell = 1$ (k is odd), the second estimate of the corollary simply follows from the Cauchy–Schwarz inequality. \square

Remark 4.7

Note that the projection operator $\Pi_{r-k+\ell}^{\mathcal{J}}$ in the above estimate could be dropped. However, because of Lemma C.15, it holds

$$\mathcal{J}_n \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \leq C \mathcal{J}_n \left[\|f^{(\ell)} - g^{(\ell)}\|_{V'}^2 \right]$$

anyway. But note that the left-hand side term vanishes in some relevant situations where the right-hand side term does not. \clubsuit

Next, an estimate for the fully discrete error $e_{\tau h, \ell}^{\mathcal{J}}$ is derived and presented. For brevity, we use the notation $\mathcal{J}_{[1, n]}[\cdot] = \sum_{\nu=1}^n \mathcal{J}_{\nu}[\cdot]$ in the following.

Lemma 4.8

For all $n = 1, \dots, N$ it holds

$$\begin{aligned} & \|e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)\|^2 + \frac{1}{2} \sum_{\nu=1}^n \|[e_{\tau h, \ell}^{\mathcal{J}}]_{\nu-1}\|^2 + \alpha \int_{t_0}^{t_n} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\ & \leq \exp(t_{n-1} - t_0) \\ & \quad \left[\frac{3}{\alpha} \left(C_{\text{emb}}^2 \mathcal{J}_{[1, n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_a^2 \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \right. \right. \\ & \quad \left. \left. + \mathcal{J}_{[1, n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right) + \|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 + \sum_{\nu=1}^{n-1} (2 + \tau_{\nu}^{-1}) \|\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)})\|^2 \right] \end{aligned}$$

with $\omega_{\nu}^{\mathcal{J}}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Remark 4.9

As already noted in the statement of the lemma, the exponential factor $\exp(t_{n-1} - t_0)$ can be dropped in the above estimate if $\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$. This holds, for example, if for all $\nu = 1, \dots, n-1$ the integrator \mathcal{J}_{ν} is the integral over I_{ν} . Here, recall that in this section \mathcal{J}_{ν} is not fixed by the concrete method but was introduced to enable more flexibility in the error analysis. So, the choice $\mathcal{J}_{\nu} = \int_{I_{\nu}}$ always is possible. \clubsuit

Proof. Combining Lemma 4.4 with $v_{\tau} := e_{\tau h, \ell}^{\mathcal{J}}$ and Corollary 4.6 with $v_{\tau h} = \Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}$ and

$w = e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^+)$, we gain

$$\begin{aligned}
 & \frac{1}{2} \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-)\|^2 - \frac{1}{2} \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^-)\|^2 + \frac{1}{2} \|[e_{\tau h, \ell}^{\mathcal{J}}]_{\nu-1}\|^2 + \alpha \int_{I_{\nu}} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\
 & \leq B_{\nu}^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, \Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}) + \delta_{1, k-2\ell}([e_{\tau h, \ell}^{\mathcal{J}}]_{\nu-1}, e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^+)) \\
 & \leq \left[C_{\text{emb}} \left(\mathcal{J}_{\nu} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right)^{1/2} + C_a \left(\mathcal{J}_{\nu} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \right)^{1/2} \right. \\
 & \quad \left. + \left(\mathcal{J}_{\nu} \left[\|\Pi_{r-k+\ell} (f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right)^{1/2} \right] \left(\int_{I_{\nu}} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \right)^{1/2} \\
 & \quad + \|\omega_{\nu-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^+)\| \\
 & \leq \frac{3}{2\alpha} \left(C_{\text{emb}}^2 \mathcal{J}_{\nu} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_a^2 \mathcal{J}_{\nu} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \right. \\
 & \quad \left. + \mathcal{J}_{\nu} \left[\|\Pi_{r-k+\ell} (f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right) + \frac{\alpha}{2} \int_{I_{\nu}} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\
 & \quad + \|\omega_{\nu-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^+)\|,
 \end{aligned}$$

where for the last step Young's inequality was used. From this it easily follows

$$\begin{aligned}
 & \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-)\|^2 - \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^-)\|^2 + \|[e_{\tau h, \ell}^{\mathcal{J}}]_{\nu-1}\|^2 + \alpha \int_{I_{\nu}} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\
 & \leq \frac{3}{\alpha} \left(C_{\text{emb}}^2 \mathcal{J}_{\nu} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_a^2 \mathcal{J}_{\nu} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \right. \\
 & \quad \left. + \mathcal{J}_{\nu} \left[\|\Pi_{r-k+\ell} (f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right) + 2\|\omega_{\nu-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^+)\|.
 \end{aligned}$$

Furthermore, the triangle inequality and again Young's inequality yield for $\nu > 1$

$$\begin{aligned}
 & 2\|\omega_{\nu-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^+)\| \leq 2\|\omega_{\nu-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^-)\| + \|[e_{\tau h, \ell}^{\mathcal{J}}]_{\nu-1}\| \right) \\
 & \leq (2 + \tau_{\nu-1}^{-1}) \|\omega_{\nu-1}^{\mathcal{J}}(R_h u^{(\ell)})\|^2 + \tau_{\nu-1} \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^-)\|^2 + \frac{1}{2} \|[e_{\tau h, \ell}^{\mathcal{J}}]_{\nu-1}\|^2.
 \end{aligned}$$

So, recalling that $\omega_0^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ and re-sorting the terms, we obtain (setting $\tau_0 = 1$)

$$\begin{aligned}
 & \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-)\|^2 + \frac{1}{2} \|[e_{\tau h, \ell}^{\mathcal{J}}]_{\nu-1}\|^2 + \alpha \int_{I_{\nu}} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\
 & \leq \frac{3}{\alpha} \left(C_{\text{emb}}^2 \mathcal{J}_{\nu} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_a^2 \mathcal{J}_{\nu} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \right. \\
 & \quad \left. + \mathcal{J}_{\nu} \left[\|\Pi_{r-k+\ell} (f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right) \\
 & \quad + (1 + (1 - \delta_{1, \nu}) \tau_{\nu-1}) \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu-1}^-)\|^2 + (1 - \delta_{1, \nu}) (2 + \tau_{\nu-1}^{-1}) \|\omega_{\nu-1}^{\mathcal{J}}(R_h u^{(\ell)})\|^2.
 \end{aligned}$$

Applying a discrete version of Gronwall's lemma, see Lemma A.1, we easily conclude the desired statement. \square

We now are ready to give an abstract L^2 -error estimate in the H -norm in terms of certain projection and approximation errors.

Lemma 4.10

Let $\|\cdot\|_W \in \{\|\cdot\|, \|\cdot\|_V\}$ if $k - 2\ell = 0$ ($\Leftrightarrow k$ is even) and $\|\cdot\|_W = \|\cdot\|$ if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd), respectively. Then, for all $n = 1, \dots, N$ it holds

$$\begin{aligned} & \int_{t_0}^{t_n} \|u^{(\ell)} - u_{\tau h}^{(\ell)}\|_W^2 dt \\ & \leq C \left(\int_{t_0}^{t_n} \|u^{(\ell)} - R_h u^{(\ell)}\|_W^2 dt + \int_{t_0}^{t_n} \|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 dt + \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_W^2 dt \right) \end{aligned}$$

where

$$\begin{aligned} & \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_W^2 dt \leq C(1 + \delta_{1, k-2\ell}(t_n - t_0)) \exp(t_{n-1} - t_0) \\ & \quad \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 + \sum_{\nu=1}^{n-1} (2 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|^2 + \mathcal{J}_{[1, n]}[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2] \right. \\ & \quad \left. + \mathcal{J}_{[1, n]}[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2] + \mathcal{J}_{[1, n]}[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'}^2] \right) \end{aligned}$$

with $\omega_\nu^{\mathcal{J}}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Proof. In order to estimate the error, we use the splitting

$$u^{(\ell)} - u_{\tau h}^{(\ell)} = (u^{(\ell)} - R_h u^{(\ell)}) + (R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}) + e_{\tau h, \ell}^{\mathcal{J}}, \quad e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}.$$

For the second summand the stability of R_h in $\|\cdot\|_V$ yields

$$\begin{aligned} & \int_{t_0}^{t_n} \|R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_W^2 dt \\ & \leq \int_{t_0}^{t_n} \max\{1, C_{\text{emb}}^2\} \|R_h(u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|_V^2 dt \leq C \int_{t_0}^{t_n} \|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 dt. \end{aligned}$$

So, the first statement follows easily by the triangle inequality.

It remains to study the third summand $e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}$. If $k - 2\ell = 0$ ($\Leftrightarrow k$ is even), it holds $e_{\tau h, \ell}^{\mathcal{J}} = \Pi_{r-k+\ell}^{\mathcal{J}} e_{\tau h, \ell}^{\mathcal{J}}$ and, thus,

$$\int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_W^2 dt = \int_{t_0}^{t_n} \|\Pi_{r-k+\ell}^{\mathcal{J}} e_{\tau h, \ell}^{\mathcal{J}}\|_W^2 dt \leq \int_{t_0}^{t_n} \max\{1, C_{\text{emb}}^2\} \|\Pi_{r-k+\ell}^{\mathcal{J}} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt.$$

Otherwise, if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd), we use the norm equivalence of Lemma 4.3 to obtain

$$\begin{aligned} & \int_{I_\nu} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \leq C \left(\int_{I_\nu} \|\Pi_{r-k+\ell}^{\mathcal{J}} e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt + \left(\frac{\tau_\nu}{2}\right) \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|^2 \right) \\ & \leq C \left(\int_{I_\nu} C_{\text{emb}}^2 \|\Pi_{r-k+\ell}^{\mathcal{J}} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt + \left(\frac{\tau_\nu}{2}\right) \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|^2 \right) \end{aligned}$$

from which it follows

$$\int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt = \sum_{\nu=1}^n \int_{I_\nu} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \leq C \int_{t_0}^{t_n} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt + C \sum_{\nu=1}^n \left(\frac{\tau_\nu}{2}\right) \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|^2.$$

Therefore, in both cases, from Lemma 4.8 we conclude the desired estimate, where we also used that $2 \sum_{\nu=1}^n \left(\frac{\tau_\nu}{2}\right) \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|^2 \leq (t_n - t_0) \max_{\nu=1, \dots, n} \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|^2$. \square

In conclusion, we have a look on the resulting convergence orders for a concrete setting. In order to easily consider different variants of the \mathbf{VTD}_k^r method, we use the short and clear notation

$$S_1 \cup_{\text{cond.}} S_2 := \begin{cases} S_1 \cup S_2, & \text{“cond.” is fulfilled,} \\ S_1, & \text{otherwise,} \end{cases}$$

where S_1 and S_2 are sets and “cond.” is a Boolean condition.

Theorem 4.11

Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} & \int_{t_0}^{t_n} \|u^{(\ell)} - u_{\tau h}^{(\ell)}\|^2 dt + \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \\ & \leq C(1 + (t_n - t_0)) \left[h^{2(\kappa+\sigma)} \|u^{(\ell)}\|_{H^1((t_0, t_n), H^{\kappa+1}(\Omega))}^2 \right. \\ & \quad \left. + \tau^{2(r-\ell+1)} \left(\|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^1(\Omega))}^2 + \|f\|_{H^{r+1}((t_0, t_n), H^{-1}(\Omega))}^2 \right) \right], \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Using Lemma 4.10 with choice $\mathcal{J}_\nu = \mathcal{I}_\nu$, we only need to bound the projection errors. From (3.16) we get

$$\int_{t_0}^{t_n} \|u^{(\ell)} - R_h u^{(\ell)}\|^2 dt \leq C h^{2(\kappa+\sigma)} \|u^{(\ell)}\|_{L^2((t_0, t_n), H^{\kappa+1}(\Omega))}^2,$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise. Similarly it follows

$$\int_{t_0}^{t_n} \|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 dt \leq C h^{2(\kappa+\sigma)} \|u^{(\ell+1)}\|_{L^2((t_0, t_n), H^{\kappa+1}(\Omega))}^2.$$

Furthermore, the projection error estimate (4.9) gives

$$\int_{t_0}^{t_n} \|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 dt \leq C \tau^{2(r-\ell+1)} \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^1(\Omega))}^2.$$

Recalling Remark 4.7 and due to $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$, standard interpolation/projection error estimates, cf. Lemma B.9, moreover imply

$$\int_{t_0}^{t_n} \|\Pi_{r-k+\ell}^{\mathcal{J}} (f^{(\ell)} - g^{(\ell)})\|_{V'}^2 dt \leq C \int_{t_0}^{t_n} \|f^{(\ell)} - g^{(\ell)}\|_{V'}^2 dt \leq C \tau^{2(r-\ell+1)} \|f\|_{H^{r+1}((t_0, t_n), H^{-1}(\Omega))}^2.$$

Finally, the special choice $u_{\tau h}^{(\ell)}(t_0^-) = \tilde{P}_h^0 \partial_t^\ell u_0$ of the initial value for the discrete problem enables

$$e_{\tau h, \ell}^{\mathcal{J}}(t_0^-) = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}(t_0^-) - u_{\tau h}^{(\ell)}(t_0^-) = R_h u^{(\ell)}(t_0^-) - \tilde{P}_h^0 u^{(\ell)}(t_0^-),$$

where we also have exploited that $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ preserves the point value in t_0^- . So, if $\tilde{P}_h^0 = R_h$, it follows $e_{\tau h, \ell}^{\mathcal{J}}(t_0^-) = 0$ and we are done. Otherwise, if $\tilde{P}_h^0 = P_h$, we conclude, using the definition of the projection P_h , that for $u^{(\ell)}(t_0^-) \in H$

$$\begin{aligned} \|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 &= (R_h u^{(\ell)}(t_0^-) - P_h u^{(\ell)}(t_0^-), e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)) = (R_h u^{(\ell)}(t_0^-) - u^{(\ell)}(t_0^-), e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)) \\ &\leq \|R_h u^{(\ell)}(t_0^-) - u^{(\ell)}(t_0^-)\| \|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|. \end{aligned}$$

Hence, then the projection error estimates for R_h , cf. (3.16), yield

$$\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 \leq \|R_h u^{(\ell)}(t_0^-) - u^{(\ell)}(t_0^-)\|^2 \leq Ch^{2(\kappa+\sigma)} \|u^{(\ell)}(t_0)\|_{H^{\kappa+1}(\Omega)}^2$$

with σ as above, which completes the proof. \square

Remark 4.12

By construction $u_{\tau h}^{(\ell)}$ is an approximation of $u^{(\ell)}$ that locally on I_ν lies in $P_{r-\ell}(I_\nu, V_h)$. The convergence orders in time and space, obtained in Theorem 4.11, thus are as expected. \clubsuit

4.1.3 Global L^2 -error in the V -norm

Inspecting the statements of Lemma 4.8 and Lemma 4.10, we see that, if $k - 2\ell = 0$ ($\Leftrightarrow k$ is even), we even have control over $\int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt$ and $\int_{t_0}^{t_n} \|u^{(\ell)} - u_{\tau h}^{(\ell)}\|_V^2 dt$, respectively. This immediately enables us to derive error estimates for dG-like methods also in the V -norm. In detail, suitably adapting the proof of Theorem 4.11, we gain the following result.

Corollary 4.13

Let $k - 2\ell = 0$ ($\Leftrightarrow k$ is even). Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} &\int_{t_0}^{t_n} \|u^{(\ell)} - u_{\tau h}^{(\ell)}\|_V^2 dt + \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\ &\leq C(1 + (t_n - t_0)) \left[h^{2\kappa} \|u^{(\ell)}\|_{H^1((t_0, t_n), H^{\kappa+1}(\Omega))}^2 \right. \\ &\quad \left. + \tau^{2(r-\ell+1)} \left(\|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^1(\Omega))}^2 + \|f\|_{H^{r+1}((t_0, t_n), H^{-1}(\Omega))}^2 \right) \right]. \end{aligned}$$

Of course, we now ask whether or not a similar estimate also can be shown for cGP-like methods, i.e., if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd). Inspecting the proof of Lemma 4.10, this could be done if we would have adequate control on $e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)$ also in the V -norm. Therefore, to gain such control, we suitably adapt the ideas used in Lemma 4.4. For the presented proof, however, some more assumptions are needed.

Assumption 4.1

We assume that the bilinear form $a(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ can be split such that

$$a(v, w) = a_0(v, w) + a_1(v, w)$$

where $a_0(\cdot, \cdot) : V \times V \rightarrow \mathbb{R}$ is a symmetric, V -elliptic, continuous bilinear form, i.e., $a_0(v, w) = a_0(w, v)$ for all $v, w \in V$,

$$\begin{aligned} \exists \alpha_0 > 0 : \quad & a_0(v, v) \geq \alpha_0 \|v\|_V^2 & \forall v \in V, \\ \exists C_{a_0} > 0 : \quad & |a_0(v, w)| \leq C_{a_0} \|v\|_V \|w\|_V & \forall v, w \in V, \end{aligned}$$

and $a_1(\cdot, \cdot) : V \times H \rightarrow \mathbb{R}$ is a continuous bilinear form, i.e.,

$$\exists C_{a_1} > 0 : \quad |a_1(v, w)| \leq C_{a_1} \|v\|_V \|w\| \quad \forall v \in V, w \in H.$$

The bilinear forms $a(\cdot, \cdot)$, $a_0(\cdot, \cdot)$, and $a_1(\cdot, \cdot)$ are all assumed to be independent of time t .

Remark 4.14

In the setting of model problem (3.4) the bilinear form $a_0(\cdot, \cdot)$ could be given by

$$a_0(v, w) = (\epsilon \nabla v, \nabla w) + (\tilde{c}v, w)$$

with $\tilde{c} \geq 0$ independent of t . Then, of course,

$$a_1(v, w) = a(v, w) - a_0(v, w) = (b \cdot \nabla v, w) + ((c - \tilde{c})v, w).$$

In the case that $c \geq 0$ one can choose $\tilde{c} = c$. Alternatively, setting $\tilde{c} \geq 1$ always guarantees that $a_0(v, v) \geq \|v\|^2$, i.e., control in the H -norm, independent of ϵ and without additional assumptions on c . ♣

Lemma 4.15

Let $1 \leq n \leq N$, $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd), and suppose that Assumption 4.1 holds. Then, for all $v_\tau \in P_{r-\ell}(I_n, V)$ we have

$$a_0(v_\tau, v_\tau) \Big|_{t_{n-1}^+}^{t_n^-} + \int_{I_n} \|\partial_t v_\tau\|^2 dt \leq 2B_n^{\mathcal{J}}(v_\tau, \partial_t v_\tau) + C_{a_1}^2 \int_{I_n} \|\Pi_{r-\ell-1} v_\tau\|_V^2 dt.$$

Proof. First of all, due to $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd), the jump term in $B_n^{\mathcal{J}}(\cdot, \cdot)$ drops out. So, exploiting the exactness of \mathcal{J}_n for polynomials of maximal degree $2r - k$ and the splitting of Assumption 4.1, we note that

$$\begin{aligned} B_n^{\mathcal{J}}(v_\tau, \partial_t v_\tau) &= \int_{I_n} (\partial_t v_\tau, \partial_t v_\tau) dt + \int_{I_n} a(v_\tau, \partial_t v_\tau) dt \\ &= \int_{I_n} \|\partial_t v_\tau\|^2 dt + \int_{I_n} a_0(v_\tau, \partial_t v_\tau) dt + \int_{I_n} a_1(v_\tau, \partial_t v_\tau) dt. \end{aligned}$$

Now, on the one hand, the symmetry of $a_0(\cdot, \cdot)$ enables the identity

$$\int_{I_n} a_0(v_\tau, \partial_t v_\tau) dt = \frac{1}{2} \int_{I_n} \partial_t a_0(v_\tau, v_\tau) dt = \frac{1}{2} a_0(v_\tau, v_\tau) \Big|_{t_{n-1}^+}^{t_n^-}.$$

On the other hand, $\partial_t v_\tau$ is a feasible test function for the L^2 -projection in time $\Pi_{r-\ell-1}$, also cf. Corollary C.14. Using this together with the continuity of $a_1(\cdot, \cdot)$, we obtain

$$\begin{aligned} \left| \int_{I_n} a_1(v_\tau, \partial_t v_\tau) dt \right| &= \left| \int_{I_n} a_1(\Pi_{r-\ell-1} v_\tau, \partial_t v_\tau) dt \right| \\ &\leq C_{a_1} \int_{I_n} \|\Pi_{r-\ell-1} v_\tau\|_V \|\partial_t v_\tau\| dt \leq \frac{C_{a_1}^2}{2} \int_{I_n} \|\Pi_{r-\ell-1} v_\tau\|_V^2 dt + \frac{1}{2} \int_{I_n} \|\partial_t v_\tau\|^2 dt, \end{aligned}$$

where we applied Young's inequality in the last step.

Altogether the above estimates yield

$$\begin{aligned} \frac{1}{2} a_0(v_\tau, v_\tau) \Big|_{t_{n-1}^+}^{t_n^-} + \int_{I_n} \|\partial_t v_\tau\|^2 dt &= B_n^{\mathcal{J}}(v_\tau, \partial_t v_\tau) - \int_{I_n} a_1(v_\tau, \partial_t v_\tau) dt \\ &\leq B_n^{\mathcal{J}}(v_\tau, \partial_t v_\tau) + \frac{C_{a_1}^2}{2} \int_{I_n} \|\Pi_{r-\ell-1} v_\tau\|_V^2 dt + \frac{1}{2} \int_{I_n} \|\partial_t v_\tau\|^2 dt. \end{aligned}$$

From this we can easily complete the proof. \square

Since, in contrast to Lemma 4.4, in the inequality of Lemma 4.15 the second argument of $B_n^{\mathcal{J}}(\cdot, \cdot)$ does not appear in the $L^2(V)$ -norm on the left-hand side but only in the $L^2(H)$ -norm, we also need to show a variant of Corollary 4.6 where instead of $(\int_{I_n} \|v_{\tau h}\|_V^2 dt)^{1/2}$ on the right-hand side it only appears $(\int_{I_n} \|v_{\tau h}\|^2 dt)^{1/2}$. For this purpose, we assume the following.

Assumption 4.2

We assume that there is a Hilbert space \tilde{V} satisfying $\tilde{V} \hookrightarrow V \hookrightarrow H$ and a bilinear form $\tilde{a}(\cdot, \cdot) : \tilde{V} \times H \rightarrow \mathbb{R}$ such that for all $v \in \tilde{V}$, $w \in V$ it holds

$$\tilde{a}(v, w) = a(v, w).$$

We furthermore assume that $\tilde{a}(\cdot, \cdot)$ is continuous, i.e.,

$$\exists C_{\tilde{a}} > 0 : \quad |\tilde{a}(v, w)| \leq C_{\tilde{a}} \|v\|_{\tilde{V}} \|w\| \quad \forall v \in \tilde{V}, w \in H.$$

Remark 4.16

In the setting of model problem (3.4) the space \tilde{V} could be chosen as $\tilde{V} = H^2(\Omega) \cap H_0^1(\Omega)$ where the bilinear form $\tilde{a}(\cdot, \cdot)$ is given by

$$\tilde{a}(v, w) = (\mathcal{A}v, w) = -(\operatorname{div}(\epsilon \nabla u), w) + (b \cdot \nabla v, w) + (cv, w).$$

As norm in \tilde{V} we then use $\|\cdot\|_{\tilde{V}} = \|\cdot\|_{H^2(\Omega)}$. \clubsuit

Under the additional Assumption 4.2 we can conclude from Lemma 4.5 the following statement.

Corollary 4.17

Let $1 \leq n \leq N$ and suppose that Assumption 4.2 holds true. Moreover, assume that $\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})(t) \in H$ for all $t \in \bar{I}$. Then, for all $v_{\tau h} \in P_{r-k+\ell}(I_n, V_h)$ it holds

$$\begin{aligned} B_n^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, v_{\tau h}) &\leq \left[\left(\mathcal{J}_n \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right)^{1/2} + C_{\tilde{a}} \left(\mathcal{J}_n \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] \right)^{1/2} \right. \\ &\quad \left. + \left(\mathcal{J}_n \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right)^{1/2} \right] \left(\int_{I_n} \|v_{\tau h}\|^2 dt \right)^{1/2} + \delta_{0, k-2\ell} \|\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|v_{\tau h}(t_{n-1}^+)\| \end{aligned}$$

with $\omega_{n-1}^{\mathcal{J}}(\cdot)$ as defined in (4.7).

Proof. Similar to the proof of Corollary 4.6, from Lemma 4.5, the definition of $\Pi_{r-k+\ell}^{\mathcal{J}}$, and Assumption 4.2 we get for all $v_{\tau h} \in P_{r-k+\ell}(I_n, V_h)$ that

$$\begin{aligned} B_n^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, v_{\tau h}) &= -\mathcal{J}_n[(u^{(\ell+1)} - R_h u^{(\ell+1)}, v_{\tau h})] - \mathcal{J}_n[a(u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h})] \\ &\quad + \mathcal{J}_n[\langle \Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)}), v_{\tau h} \rangle_{V', V}] + \delta_{0, k-2\ell} (\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}), v_{\tau h}(t_{n-1}^+)) \\ &= -\mathcal{J}_n[(u^{(\ell+1)} - R_h u^{(\ell+1)}, v_{\tau h})] - \mathcal{J}_n[\tilde{a}(u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}, v_{\tau h})] \\ &\quad + \mathcal{J}_n[(\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)}), v_{\tau h})] + \delta_{0, k-2\ell} (\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)}), v_{\tau h}(t_{n-1}^+)) \\ &\leq \mathcal{J}_n[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\| \|v_{\tau h}\|] + C_{\tilde{a}} \mathcal{J}_n[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}} \|v_{\tau h}\|] \\ &\quad + \mathcal{J}_n[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\| \|v_{\tau h}\|] + \delta_{0, k-2\ell} \|\omega_{n-1}^{\mathcal{J}}(R_h u^{(\ell)})\| \|v_{\tau h}(t_{n-1}^+)\|. \end{aligned}$$

Applying the Cauchy–Schwarz-type inequality (4.5c) and using the exactness of \mathcal{J}_n for polynomials up to degree $2r - k$, see (4.5a), we easily finish the proof. \square

We now get another estimate for the fully discrete error $e_{\tau h, \ell}^{\mathcal{J}}$ similar to Lemma 4.8.

Lemma 4.18

Let $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd) and suppose that Assumptions 4.1 and 4.2 hold. Moreover, assume that $\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})(t) \in H$ for all $t \in \bar{I}$. Then, for all $n = 1, \dots, N$ we have

$$\begin{aligned} \alpha_0 \|e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)\|_V^2 &+ \frac{1}{2} \int_{t_0}^{t_n} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \\ &\leq \exp(2(t_{n-1} - t_0)) \\ &\quad \left(6 \left[1 + \frac{C_{a_1}^2 C_{\text{emb}}^2}{2\alpha^2} \right] \left(\mathcal{J}_{[1, n]}[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2] + \mathcal{J}_{[1, n]}[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2] \right) \right. \\ &\quad + 6C_{\tilde{a}}^2 \mathcal{J}_{[1, n]}[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2] + 3 \frac{C_{a_1}^2 C_{\tilde{a}}^2}{\alpha^2} \mathcal{J}_{[1, n]}[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2] \\ &\quad \left. + \left[C_{a_0} + \frac{C_{a_1}^2 C_{\text{emb}}^2}{\alpha} \right] \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V^2 + \sum_{\nu=1}^{n-1} (2 + \tau_{\nu}^{-1}) \|\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)})\|_V^2 \right) \right) \end{aligned}$$

with $\omega_\nu^\mathcal{J}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_\nu^\mathcal{J}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Proof. Paying heed to $k - 2\ell = 1$, an application of Lemma 4.15 with $v_\tau = e_{\tau h, \ell}^\mathcal{J}$ and of Corollary 4.17 with $v_{\tau h} = \partial_t e_{\tau h, \ell}^\mathcal{J}$ yields

$$\begin{aligned} a_0(e_{\tau h, \ell}^\mathcal{J}, e_{\tau h, \ell}^\mathcal{J})|_{t_{\nu-1}^+}^{t_\nu^-} + \int_{I_\nu} \|\partial_t e_{\tau h, \ell}^\mathcal{J}\|^2 dt &\leq 2B_\nu^\mathcal{J}(e_{\tau h, \ell}^\mathcal{J}, \partial_t e_{\tau h, \ell}^\mathcal{J}) + C_{a_1}^2 \int_{I_\nu} \|\Pi_{r-\ell-1} e_{\tau h, \ell}^\mathcal{J}\|_V^2 dt \\ &\leq 2 \left[\left(\mathcal{J}_\nu \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right)^{1/2} + C_{\tilde{a}} \left(\mathcal{J}_\nu \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] \right)^{1/2} \right. \\ &\quad \left. + \left(\mathcal{J}_\nu \left[\|\Pi_{r-k+\ell}^\mathcal{J}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right)^{1/2} \right] \left(\int_{I_\nu} \|\partial_t e_{\tau h, \ell}^\mathcal{J}\|^2 dt \right)^{1/2} + C_{a_1}^2 \int_{I_\nu} \|\Pi_{r-\ell-1} e_{\tau h, \ell}^\mathcal{J}\|_V^2 dt \\ &\leq 6 \left(\mathcal{J}_\nu \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_{\tilde{a}}^2 \mathcal{J}_\nu \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] + \mathcal{J}_\nu \left[\|\Pi_{r-k+\ell}^\mathcal{J}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right) \\ &\quad + \frac{1}{2} \int_{I_\nu} \|\partial_t e_{\tau h, \ell}^\mathcal{J}\|^2 dt + C_{a_1}^2 \int_{I_\nu} \|\Pi_{r-\ell-1} e_{\tau h, \ell}^\mathcal{J}\|_V^2 dt, \end{aligned}$$

where we also used Young's inequality. Hence, we have

$$\begin{aligned} a_0(e_{\tau h, \ell}^\mathcal{J}(t_\nu^-), e_{\tau h, \ell}^\mathcal{J}(t_\nu^-)) + \frac{1}{2} \int_{I_\nu} \|\partial_t e_{\tau h, \ell}^\mathcal{J}\|^2 dt &\tag{4.11} \\ &\leq 6 \left(\mathcal{J}_\nu \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_{\tilde{a}}^2 \mathcal{J}_\nu \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] + \mathcal{J}_\nu \left[\|\Pi_{r-k+\ell}^\mathcal{J}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right) \\ &\quad + C_{a_1}^2 \int_{I_\nu} \|\Pi_{r-\ell-1} e_{\tau h, \ell}^\mathcal{J}\|_V^2 dt + a_0(e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^+), e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^+)). \end{aligned}$$

Recalling the second statement of Lemma 4.5, we get

$$e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^+) = e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-) + [e_{\tau h, \ell}^\mathcal{J}]_{\nu-1} = e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-) + \omega_{\nu-1}^\mathcal{J}(R_h u^{(\ell)}). \tag{4.12}$$

Therefore, since $a_0(\cdot, \cdot)$ is an inner product on V , it follows for $\nu > 1$

$$\begin{aligned} a_0(e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^+), e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^+)) &= a_0(e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-) + \omega_{\nu-1}^\mathcal{J}(R_h u^{(\ell)}), e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-) + \omega_{\nu-1}^\mathcal{J}(R_h u^{(\ell)})) \\ &= a_0(e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-), e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-)) \\ &\quad + 2a_0(e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-), \omega_{\nu-1}^\mathcal{J}(R_h u^{(\ell)})) + a_0(\omega_{\nu-1}^\mathcal{J}(R_h u^{(\ell)}), \omega_{\nu-1}^\mathcal{J}(R_h u^{(\ell)})) \\ &\leq (1 + \tau_{\nu-1}) a_0(e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-), e_{\tau h, \ell}^\mathcal{J}(t_{\nu-1}^-)) + C_{a_0} (1 + \tau_{\nu-1}^{-1}) \|\omega_{\nu-1}^\mathcal{J}(R_h u^{(\ell)})\|_V^2, \end{aligned} \tag{4.13}$$

where we applied the Cauchy–Schwarz inequality, Young's inequality, and the continuity of $a_0(\cdot, \cdot)$. Moreover, for $\nu = 1$ we find

$$a_0(e_{\tau h, \ell}^\mathcal{J}(t_0^+), e_{\tau h, \ell}^\mathcal{J}(t_0^+)) \leq C_{a_0} \|e_{\tau h, \ell}^\mathcal{J}(t_0^-)\|_V^2 \tag{4.14}$$

because of $\omega_0^\mathcal{J}(\cdot) = 0$.

Now, combining (4.11) with (4.13) for $\nu > 1$ or (4.14) for $\nu = 1$ and applying to the resulting inequalities a discrete version of Gronwall's lemma, see Lemma A.1, it follows

$$\begin{aligned}
 & a_0(e_{\tau h, \ell}^{\mathcal{J}}(t_n^-), e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)) + \frac{1}{2} \int_{t_0}^{t_n} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \\
 & \leq \exp(t_{n-1} - t_0) \left[6\mathcal{J}_{[1, n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + 6C_a^2 \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] \right. \\
 & \quad + 6\mathcal{J}_{[1, n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] + C_{a_0} \|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V^2 \\
 & \quad \left. + C_{a_0} \sum_{\nu=1}^{n-1} (1 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|_V^2 + C_{a_1}^2 \int_{t_0}^{t_n} \|\Pi_{r-\ell-1} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \right].
 \end{aligned}$$

Altogether, using the V -ellipticity of $a_0(\cdot, \cdot)$, Lemma 4.8 (note that here $r - \ell - 1 = r - k + \ell$), and taking into account that $\|w\| \leq C_{\text{emb}} \|w\|_V$ for all $w \in V$, which also implies the estimate $\|w\|_{V'} \leq C_{\text{emb}} \|w\|$ for all $w \in H \subset V'$, we obtain

$$\begin{aligned}
 & \alpha_0 \|e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)\|_V^2 + \frac{1}{2} \int_{t_0}^{t_n} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \\
 & \leq \exp(2(t_{n-1} - t_0)) \\
 & \quad \left(6 \left[1 + \frac{C_{a_1}^2 C_{\text{emb}}^2}{2\alpha^2} \right] \left(\mathcal{J}_{[1, n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + \mathcal{J}_{[1, n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right) \right. \\
 & \quad + 6C_a^2 \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] + 3 \frac{C_{a_1}^2 C_a^2}{\alpha^2} \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \\
 & \quad \left. + \left[C_{a_0} + \frac{C_{a_1}^2 C_{\text{emb}}^2}{\alpha} \right] \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V^2 + \sum_{\nu=1}^{n-1} (2 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|_V^2 \right) \right).
 \end{aligned}$$

Thus, we are done. \square

Since Lemma 4.18 provides the previously missing control on $\|e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)\|_V$, we obtain an abstract estimate for the L^2 -error in the V -norm also for cGP-like methods, i.e., if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd).

Lemma 4.19

Let $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd) and suppose that Assumptions 4.1 and 4.2 hold. Moreover, assume that $\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})(t) \in H$ for all $t \in \bar{I}$. Then, for all $n = 1, \dots, N$ it holds

$$\begin{aligned}
 & \int_{t_0}^{t_n} \|u^{(\ell)} - u_{\tau h}^{(\ell)}\|_V^2 dt \\
 & \leq C \left(\int_{t_0}^{t_n} \|u^{(\ell)} - R_h u^{(\ell)}\|_V^2 dt + \int_{t_0}^{t_n} \|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 dt + \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \right)
 \end{aligned}$$

where

$$\begin{aligned} \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt &\leq C(1 + (t_n - t_0)) \exp(2(t_{n-1} - t_0)) \\ &\quad \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V^2 + \sum_{\nu=1}^{n-1} (2 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|_V^2 + \mathcal{J}_{[1, n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right. \\ &\quad + \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] + \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \\ &\quad \left. + \mathcal{J}_{[1, n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right) \end{aligned}$$

with $\omega_\nu^{\mathcal{J}}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Proof. The arguments are quite analog to those used in the proof of Lemma 4.10. We therefore only consider some of the details for the derivation of the second statement.

The norm equivalence of Lemma 4.3 gives

$$\int_{I_\nu} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \leq C \left(\int_{I_\nu} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt + \left(\frac{\tau_\nu}{2}\right) \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|_V^2 \right).$$

A summation over $\nu = 1, \dots, n$ yields

$$\int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt = \sum_{\nu=1}^n \int_{I_\nu} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \leq C \int_{t_0}^{t_n} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt + C \sum_{\nu=1}^n \left(\frac{\tau_\nu}{2}\right) \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|_V^2.$$

Then, an application of Lemma 4.8 and Lemma 4.18 gives the desired second estimate. \square

Concrete convergence orders for the model problem are given in the next theorem.

Theorem 4.20

Let $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd). Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} &\int_{t_0}^{t_n} \|u^{(\ell)} - u_{\tau h}^{(\ell)}\|_V^2 dt + \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\ &\leq C(1 + (t_n - t_0)) \left[h^{2\kappa} \|u^{(\ell)}\|_{H^1((t_0, t_n), H^{\kappa+1}(\Omega))}^2 \right. \\ &\quad \left. + \tau^{2(r-\ell+1)} \left(\|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^2(\Omega))}^2 + \|f\|_{H^{r+1}((t_0, t_n), L^2(\Omega))}^2 \right) \right]. \end{aligned}$$

Proof. First of all, note that in the setting of model problem (3.4) the Assumptions 4.1 and 4.2 usually are fulfilled when the problem data is sufficiently smooth, see Remarks 4.14 and 4.16, respectively.

So, because of Lemma 4.19, used with $\mathcal{J}_\nu = \int_{I_\nu}$, it only remains to bound certain projection errors. These error terms can be estimated similar to the terms in the proof of Theorem 4.11. From (3.16a) we gain

$$\int_{t_0}^{t_n} \|u^{(\ell)} - R_h u^{(\ell)}\|_V^2 dt \leq C h^{2\kappa} \|u^{(\ell)}\|_{L^2((t_0, t_n), H^{\kappa+1}(\Omega))}^2$$

and

$$\int_{t_0}^{t_n} \|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 dt \leq Ch^{2\kappa} \|u^{(\ell+1)}\|_{L^2((t_0, t_n), H^{\kappa+1}(\Omega))}^2.$$

The error estimate (4.9) for $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ yields

$$\int_{t_0}^{t_n} \|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 dt + \int_{t_0}^{t_n} \|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 dt \leq C\tau^{2(r-\ell+1)} \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^2(\Omega))}^2$$

and, since $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$, standard interpolation/projection error estimates, cf. Lemma B.9, give

$$\int_{t_0}^{t_n} \|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 dt \leq C \int_{t_0}^{t_n} \|f^{(\ell)} - g^{(\ell)}\|^2 dt \leq C\tau^{2(r-\ell+1)} \|f\|_{H^{r+1}((t_0, t_n), L^2(\Omega))}^2.$$

Moreover, as seen in the proof of Theorem 4.11, we have $e_{\tau h, \ell}^{\mathcal{J}}(t_0^-) = R_h u^{(\ell)}(t_0^-) - \tilde{P}_h^0 u^{(\ell)}(t_0^-)$. Thus, $\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V = 0$ if $\tilde{P}_h^0 = R_h$. Otherwise, for $\tilde{P}_h^0 = P_h$, we conclude from the V -ellipticity and the continuity of $a(\cdot, \cdot)$ as well as the definition of R_h that

$$\begin{aligned} \alpha \|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V^2 &\leq a(R_h u^{(\ell)}(t_0^-) - P_h u^{(\ell)}(t_0^-), e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)) = a(u^{(\ell)}(t_0^-) - P_h u^{(\ell)}(t_0^-), e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)) \\ &\leq C_a \|u^{(\ell)}(t_0^-) - P_h u^{(\ell)}(t_0^-)\|_V \|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V. \end{aligned}$$

Using standard arguments to bound the error of P_h , we gain

$$\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V^2 \leq \alpha^{-2} C_a^2 \|u^{(\ell)}(t_0^-) - P_h u^{(\ell)}(t_0^-)\|_V^2 \leq Ch^{2\kappa} \|u^{(\ell)}(t_0^-)\|_{H^{\kappa+1}(\Omega)}^2.$$

Summarizing the above estimates gives the desired statement. \square

Remark 4.21

In this subsection we have looked at the stronger V -norm in space and not the H -norm anymore. This is also the reason why Corollary 4.13 and Theorem 4.20 show a slightly lower spatial order of convergence than Theorem 4.11. The proven convergence orders exactly match our expectations. \clubsuit

Remark 4.22

Similar estimates to those of Corollary 4.13 and Theorem 4.20 are well known from the literature for the discontinuous Galerkin method ($k = 0$), even in a more general setting, see e.g. [26, Theorem 69.18, p. 188]. However, for the continuous Galerkin–Petrov method ($k = 1$) typically only certain components of the error are estimated in the $L^2(V)$ -norm, see e.g. [26, Theorem 70.11, p. 203, note (70.17), p. 201]. This is since, in contrast to the dG methods, estimates in $L^2(V)$ are not directly obtained for cGP methods, also cf. [5, Remark 5.1]. Thus, for odd k such estimates may be new. \clubsuit

4.1.4 Global (locally weighted) L^2 -error of the time derivative in the H -norm

For $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd) Lemma 4.18 provides a bound for $\int_{t_0}^{t_n} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt$. This gives rise to error estimates for the $L^2(H)$ -norm of the time derivative of $u^{(\ell)} - u_{\tau h}^{(\ell)}$ for cGP-like methods.

Lemma 4.23

Let $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd) and suppose that Assumptions 4.1 and 4.2 hold. Moreover, assume that $\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})(t) \in H$ for all $t \in \bar{I}$. Then, for all $n = 1, \dots, N$ it holds

$$\begin{aligned} & \int_{t_0}^{t_n} \|\partial_t u^{(\ell)} - \partial_t u_{\tau h}^{(\ell)}\|^2 dt \\ & \leq C \left(\int_{t_0}^{t_n} \|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 dt + \int_{t_0}^{t_n} \|\partial_t (u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|_V^2 dt + \int_{t_0}^{t_n} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \right) \end{aligned}$$

where

$$\begin{aligned} & \int_{t_0}^{t_n} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \leq C \exp(2(t_{n-1} - t_0)) \\ & \quad \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|_V^2 + \sum_{\nu=1}^{n-1} (2 + \tau_{\nu}^{-1}) \|\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)})\|_V^2 + \mathcal{J}_{[1, n]}[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2] \right. \\ & \quad + \mathcal{J}_{[1, n]}[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2] + \mathcal{J}_{[1, n]}[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2] \\ & \quad \left. + \mathcal{J}_{[1, n]}[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2] \right) \end{aligned}$$

with $\omega_{\nu}^{\mathcal{J}}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Proof. A similar splitting as in the proof of Lemma 4.10 gives

$$\begin{aligned} & \partial_t u^{(\ell)} - \partial_t u_{\tau h}^{(\ell)} \\ & = (u^{(\ell+1)} - R_h u^{(\ell+1)}) + (R_h(\partial_t u^{(\ell)}) - R_h(\partial_t \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})) + (\partial_t R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - \partial_t u_{\tau h}^{(\ell)}), \end{aligned}$$

where we used that the time derivative commutes with the spatial operator R_h , see also [26, Lemma 64.34, p. 118]. The second summand can be estimated exploiting the stability of R_h in the V -norm as follows

$$\begin{aligned} & \int_{t_0}^{t_n} \|R_h(\partial_t u^{(\ell)}) - R_h(\partial_t \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|^2 dt \leq \int_{t_0}^{t_n} C_{\text{emb}}^2 \|R_h \partial_t (u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|_V^2 dt \\ & \leq C \int_{t_0}^{t_n} \|\partial_t (u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|_V^2 dt. \end{aligned}$$

A bound for the third summand $\partial_t R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - \partial_t u_{\tau h}^{(\ell)} = \partial_t e_{\tau h, \ell}^{\mathcal{J}}$ was already presented in Lemma 4.18. So, summarizing the above ideas and bounds, the proof is easily completed. \square

Theorem 4.24

Let $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd). Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} \int_{t_0}^{t_n} \|\partial_t u^{(\ell)} - \partial_t u_{\tau h}^{(\ell)}\|^2 dt &\leq C \left[h^{2(\kappa+\tilde{\sigma})} \|u^{(\ell)}\|_{H^1((t_0,t_n),H^{\kappa+1}(\Omega))}^2 + \tau^{2(r-\ell)} \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0,t_n),H^1(\Omega))}^2 \right. \\ &\quad \left. + \tau^{2(r-\ell+1)} \left(\|u^{(\ell)}\|_{H^{r-\ell+1}((t_0,t_n),H^2(\Omega))}^2 + \|f\|_{H^{r+1}((t_0,t_n),L^2(\Omega))}^2 \right) \right], \end{aligned}$$

where $\tilde{\sigma} = 1$ if the associated stationary problem is H^2 -regular as well as $\tilde{P}_h^0 = R_h$ and $\tilde{\sigma} = 0$ otherwise.

Proof. Analogously to the proof of Theorem 4.20 we (can) suppose that Assumptions 4.1 and 4.2 hold.

We bound the terms on the right-hand side of the estimate in Lemma 4.23 with $\mathcal{J}_\nu = \int_{I_\nu}$. By (3.16a) and (3.16b) we gain

$$\int_{t_0}^{t_n} \|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 dt \leq C h^{2(\kappa+\sigma)} \|u^{(\ell+1)}\|_{L^2((t_0,t_n),H^{\kappa+1}(\Omega))}^2,$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise. Moreover, on the basis of (4.9) we get

$$\int_{t_0}^{t_n} \|\partial_t (u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}} u^{(\ell)})\|_V^2 dt \leq C \tau^{2(r-\ell)} \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0,t_n),H^1(\Omega))}^2.$$

The remaining terms have already been estimated in the proof of Theorem 4.20. Especially, recall that

$$\|e_{\tau h,\ell}^{\mathcal{J}}(t_0^-)\|_V^2 \leq \begin{cases} 0, & \tilde{P}_h^0 = R_h, \\ C h^{2\kappa} \|u^{(\ell)}(t_0^-)\|_{H^{\kappa+1}(\Omega)}^2, & \tilde{P}_h^0 = P_h, \end{cases}$$

which must be reflected in the definition of $\tilde{\sigma}$. □

Remark 4.25

Compared to Theorem 4.11, we consider in Theorem 4.24 a time derivative of the error increased by one. Therefore, as expected, the temporal order of convergence is decreased by one. However, having a closer look at Lemma 4.23 and the proof of Theorem 4.24, we observe that for $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd)

$$\int_{t_0}^{t_n} \|\partial_t e_{\tau h,\ell}^{\mathcal{J}}\|^2 dt \leq C(f, u) (h^{2(\kappa+\tilde{\sigma})} + \tau^{2(r-\ell+1)}).$$

So, the decrease does not occur for the fully discrete error with $\mathcal{J}_\nu = \int_{I_\nu}$, which means that $\partial_t R_h \tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}} u^{(\ell)}$ is superclose to $\partial_t u_{\tau h}^{(\ell)}$.

Note that the spatial convergence order in Theorem 4.24 depends on the concrete choice of the projection operator \tilde{P}_h^0 used for the spatial approximation of the initial value. It is not yet clear whether this dependence is only due to the proof technique or whether it actually exists. ♣

A similar estimate, but in a locally weighted norm, shall now also be derived for $k - 2\ell = 0$ ($\Leftrightarrow k$ is even). We start with showing one further property of the bilinear form $B_n^{\mathcal{J}}(\cdot, \cdot)$ that can be used to provide another kind of control on the fully discrete error.

Lemma 4.26

Let $1 \leq n \leq N$, $k - 2\ell = 0$ ($\Leftrightarrow k$ is even), and suppose that Assumption 4.1 holds. Then, for all $v_\tau \in P_{r-\ell}(I_n, V)$ we have

$$\begin{aligned} \tau_n a_0(v_\tau(t_n^-), v_\tau(t_n^-)) + \int_{I_n} \|\partial_t v_\tau\|^2 (t - t_{n-1}) \, dt \\ \leq 2B_n^{\mathcal{J}}(v_\tau, (t - t_{n-1})\partial_t v_\tau) + (C_{a_0} + \tau_n C_{a_1}^2) \int_{I_n} \|v_\tau\|_V^2 \, dt. \end{aligned}$$

Proof. Since the test function $(t - t_{n-1})\partial_t v_\tau$ is zero at t_{n-1}^+ , the jump term in $B_n^{\mathcal{J}}(\cdot, \cdot)$ vanishes. So, under Assumption 4.1 and using the exactness of \mathcal{J}_n for polynomials of maximal degree $2r - k$, we get that (only here $k - 2\ell = 0$ is needed)

$$\begin{aligned} B_n^{\mathcal{J}}(v_\tau, (t - t_{n-1})\partial_t v_\tau) \\ = \int_{I_n} (\partial_t v_\tau, (t - t_{n-1})\partial_t v_\tau) \, dt + \int_{I_n} a(v_\tau, (t - t_{n-1})\partial_t v_\tau) \, dt \\ = \int_{I_n} \|\partial_t v_\tau\|^2 (t - t_{n-1}) \, dt + \int_{I_n} a_0(v_\tau, (t - t_{n-1})\partial_t v_\tau) \, dt + \int_{I_n} a_1(v_\tau, (t - t_{n-1})\partial_t v_\tau) \, dt. \end{aligned}$$

Because of the symmetry of $a_0(\cdot, \cdot)$, it follows

$$\begin{aligned} \frac{\tau_n}{2} a_0(v_\tau(t_n^-), v_\tau(t_n^-)) &= \frac{1}{2} \int_{I_n} \partial_t (a_0(v_\tau, v_\tau)(t - t_{n-1})) \, dt \\ &= \int_{I_n} a_0(v_\tau, (t - t_{n-1})\partial_t v_\tau) \, dt + \frac{1}{2} \int_{I_n} a_0(v_\tau, v_\tau) \, dt. \end{aligned}$$

Therefore, we obtain

$$\begin{aligned} \int_{I_n} \|\partial_t v_\tau\|^2 (t - t_{n-1}) \, dt + \frac{\tau_n}{2} a_0(v_\tau(t_n^-), v_\tau(t_n^-)) \\ = B_n^{\mathcal{J}}(v_\tau, (t - t_{n-1})\partial_t v_\tau) + \frac{1}{2} \int_{I_n} a_0(v_\tau, v_\tau) \, dt - \int_{I_n} a_1(v_\tau, (t - t_{n-1})\partial_t v_\tau) \, dt. \end{aligned}$$

The continuity of $a_0(\cdot, \cdot)$ and $a_1(\cdot, \cdot)$ as well as Young's inequality yield

$$\begin{aligned} \frac{1}{2} \int_{I_n} a_0(v_\tau, v_\tau) \, dt + \left| \int_{I_n} a_1(v_\tau, (t - t_{n-1})\partial_t v_\tau) \, dt \right| \\ \leq \frac{C_{a_0}}{2} \int_{I_n} \|v_\tau\|_V^2 \, dt + C_{a_1} \int_{I_n} \|v_\tau\|_V \|(t - t_{n-1})\partial_t v_\tau\| \, dt \\ \leq \frac{1}{2} (C_{a_0} + \tau_n C_{a_1}^2) \int_{I_n} \|v_\tau\|_V^2 \, dt + \frac{1}{2} \int_{I_n} \|\partial_t v_\tau\|^2 (t - t_{n-1}) \, dt. \end{aligned}$$

Combining this estimate with the above identity and re-sorting the terms, we easily finish the proof. \square

On the basis of Lemma 4.26 and using other already known estimates for the fully discrete error $e_{\tau h, \ell}^{\mathcal{J}}$, we can bound $\partial_t e_{\tau h, \ell}^{\mathcal{J}}$ in a locally weighted norm.

Lemma 4.27

Let $k - 2\ell = 0$ ($\Leftrightarrow k$ is even) and suppose that Assumptions 4.1 and 4.2 hold. Moreover, assume that $\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})(t) \in H$ for all $t \in \bar{I}$. Then, for all $n = 1, \dots, N$ we have

$$\begin{aligned} & \alpha_0 \sum_{\nu=1}^n \tau_{\nu} \|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^{-})\|_V^2 + \frac{1}{2} \sum_{\nu=1}^n \int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2(t - t_{\nu-1}) dt \\ & \leq \exp(t_{n-1} - t_0) \\ & \quad \left(6 \left[\tau + \frac{(C_{a_0} + \tau C_{a_1}^2) C_{\text{emb}}^2}{2\alpha^2} \right] \left(\mathcal{J}_{[1,n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + \mathcal{J}_{[1,n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right) \right. \\ & \quad + 6C_a^2 \tau \mathcal{J}_{[1,n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] + 3 \frac{(C_{a_0} + \tau C_{a_1}^2) C_a^2}{\alpha^2} \mathcal{J}_{[1,n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \\ & \quad \left. + \frac{C_{a_0} + \tau C_{a_1}^2}{\alpha} \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^{-})\|^2 + \sum_{\nu=1}^{n-1} (2 + \tau_{\nu}^{-1}) \|\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)})\|^2 \right) \right) \end{aligned}$$

with $\omega_{\nu}^{\mathcal{J}}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Proof. From Lemma 4.26 with $v_{\tau} = e_{\tau h, \ell}^{\mathcal{J}}$ and Corollary 4.17 with $v_{\tau h} = (t - t_{n-1}) \partial_t e_{\tau h, \ell}^{\mathcal{J}}$ we gain (noting that $v_{\tau h}(t_{n-1}^{+}) = 0$ by choice of $v_{\tau h}$)

$$\begin{aligned} & \tau_{\nu} a_0(e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^{-}), e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^{-})) + \int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2(t - t_{\nu-1}) dt \\ & \leq 2B_{\nu}^{\mathcal{J}}(e_{\tau h, \ell}^{\mathcal{J}}, (t - t_{\nu-1}) \partial_t e_{\tau h, \ell}^{\mathcal{J}}) + (C_{a_0} + \tau_{\nu} C_{a_1}^2) \int_{I_{\nu}} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\ & \leq 2 \left[\left(\mathcal{J}_{\nu} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right)^{1/2} + C_{\tilde{a}} \left(\mathcal{J}_{\nu} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] \right)^{1/2} \right. \\ & \quad \left. + \left(\mathcal{J}_{\nu} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right)^{1/2} \right] \left(\int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2(t - t_{\nu-1})^2 dt \right)^{1/2} \\ & \quad + (C_{a_0} + \tau_{\nu} C_{a_1}^2) \int_{I_{\nu}} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\ & \leq 6\tau_{\nu} \left(\mathcal{J}_{\nu} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_{\tilde{a}}^2 \mathcal{J}_{\nu} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] \right. \\ & \quad \left. + \mathcal{J}_{\nu} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right) + \frac{1}{2} \int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2(t - t_{\nu-1}) dt \\ & \quad + (C_{a_0} + \tau_{\nu} C_{a_1}^2) \int_{I_{\nu}} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt, \end{aligned}$$

where Young's inequality was used. Therefore, also considering the V -ellipticity of $a_0(\cdot, \cdot)$,

a summation over $\nu = 1, \dots, n$ yields

$$\begin{aligned} & \alpha_0 \sum_{\nu=1}^n \tau_\nu \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|_V^2 + \frac{1}{2} \sum_{\nu=1}^n \int_{I_\nu} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2(t - t_{\nu-1}) dt \\ & \leq 6\tau \left(\mathcal{J}_{[1, n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + C_{\tilde{a}}^2 \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] \right. \\ & \quad \left. + \mathcal{J}_{[1, n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right) + (C_{a_0} + \tau C_{a_1}^2) \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt. \end{aligned}$$

Estimating the last term according to Lemma 4.8 and noting that $\|w\| \leq C_{\text{emb}} \|w\|_V$ for all $w \in V$, which also implies $\|w\|_{V'} \leq C_{\text{emb}} \|w\|$ for all $w \in H \subset V'$, the desired statement follows easily. \square

The previous lemma leads to the following abstract and concrete estimates on the $(\ell+1)$ th derivative of the error $u - u_{\tau h}$ for dG-like methods.

Lemma 4.28

Let $k - 2\ell = 0$ ($\Leftrightarrow k$ is even) and suppose that Assumptions 4.1 and 4.2 hold. Moreover, assume that $\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})(t) \in H$ for all $t \in \bar{I}$. Then, for all $n = 1, \dots, N$ it holds

$$\begin{aligned} & \sum_{\nu=1}^n \int_{I_\nu} \|\partial_t u^{(\ell)} - \partial_t u_{\tau h}^{(\ell)}\|^2(t - t_{\nu-1}) dt \\ & \leq C(1 + \tau) \exp(t_{n-1} - t_0) \\ & \quad \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 + \sum_{\nu=1}^{n-1} (2 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|^2 + \mathcal{J}_{[1, n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|^2 \right] \right. \\ & \quad + \int_{t_0}^{t_n} \|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 dt + \mathcal{J}_{[1, n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] + \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \\ & \quad \left. + \tau \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_{\tilde{V}}^2 \right] + \tau \int_{t_0}^{t_n} \|\partial_t (u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|_V^2 dt \right) \end{aligned}$$

with $\omega_\nu^{\mathcal{J}}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Proof. We suitably adapt the proof of Lemma 4.23. Especially, we use the same splitting and obtain for the occurring middle summand that

$$\begin{aligned} & \sum_{\nu=1}^n \int_{I_\nu} \|R_h(\partial_t u^{(\ell)}) - R_h(\partial_t \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|^2(t - t_{\nu-1}) dt \\ & \leq C \sum_{\nu=1}^n \int_{I_\nu} \|\partial_t (u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|_V^2(t - t_{\nu-1}) dt \leq C\tau \int_{t_0}^{t_n} \|\partial_t (u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})\|_V^2 dt. \end{aligned}$$

Moreover, we use Lemma 4.27 to bound the term that includes $\partial_t e_{\tau h, \ell}^{\mathcal{J}}$. \square

Theorem 4.29

Let $k - 2\ell = 0$ ($\Leftrightarrow k$ is even). Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} & \sum_{\nu=1}^n \int_{I_\nu} \|\partial_t u^{(\ell)} - \partial_t u_{\tau h}^{(\ell)}\|^2 (t - t_{\nu-1}) dt \\ & \leq C(1 + \tau) \left[h^{2(\kappa+\sigma)} \|u^{(\ell)}\|_{H^1((t_0, t_n), H^{\kappa+1}(\Omega))}^2 + \tau^{2(r-\ell)+1} \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^1(\Omega))}^2 \right. \\ & \quad \left. + \tau^{2(r-\ell+1)} \left(\tau \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^2(\Omega))}^2 + \|f\|_{H^{r+1}((t_0, t_n), L^2(\Omega))}^2 \right) \right], \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Again, we (can) suppose that Assumptions 4.1 and 4.2 are fulfilled. It then only remains to bound the right-hand side of Lemma 4.28 for the choice $\mathcal{J}_\nu = \int_{I_\nu}$. The estimates for the occurring terms are clear, cf. the proofs of Theorems 4.11, 4.20, and 4.24.

However, we want to point out that, in contrast to the proof of Theorem 4.24, for $e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)$ only a bound in the H -norm is needed here. For this it holds (see proof of Theorem 4.11)

$$\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 \leq \begin{cases} 0, & \tilde{P}_h^0 = R_h, \\ Ch^{2(\kappa+\sigma)} \|u^{(\ell)}(t_0^-)\|_{H^{\kappa+1}(\Omega)}^2, & \tilde{P}_h^0 = P_h, \end{cases}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise. This justifies the slightly better spatial order. \square

Remark 4.30

Comparing the estimates of Theorem 4.24 and Theorem 4.29, one may briefly wonder about the additional power of τ . However, this results from the weighting functions $t \mapsto (t - t_{\nu-1})$ used locally on I_ν .

Moreover, note that from Lemma 4.27 and usual estimates for the occurring projection error terms, we have for the fully discrete error with $\mathcal{J}_\nu = \int_{I_\nu}$ and if $k - 2\ell = 0$ ($\Leftrightarrow k$ is even) that

$$\sum_{\nu=1}^n \int_{I_\nu} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 (t - t_{\nu-1}) dt \leq C(f, u) (h^{2(\kappa+\sigma)} + \tau^{2(r-\ell+1)}),$$

which shows an improved convergence behavior with respect to time approximation compared to the respective estimate for the error. Thus, also in the dG-like case $\partial_t R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}$ is superclose to $\partial_t u_{\tau h}^{(\ell)}$. \clubsuit

4.1.5 Pointwise error in the H -norm

For both, cGP-like and dG-like methods, we have control over the time derivative of the discrete error term $e_{\tau h, \ell}^{\mathcal{J}}$ in a (locally weighted) L^2 -norm. This can be used to derive pointwise error estimates.

Lemma 4.31

Suppose that Assumptions 4.1 and 4.2 hold. Moreover, assume that $\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})(t) \in H$ for all $t \in \bar{I}$. Then, for all $n = 1, \dots, N$ it holds

$$\begin{aligned} & \sup_{t \in [t_0, t_n]} \| (u^{(\ell)} - u_{\tau h}^{(\ell)})(t) \|^2 \\ & \leq C \left(\sup_{t \in [t_0, t_n]} \| u^{(\ell)}(t) - R_h u^{(\ell)}(t) \|^2 + \sup_{t \in [t_0, t_n]} \| u^{(\ell)}(t) - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}(t) \|_V^2 + \sup_{t \in [t_0, t_n]} \| e_{\tau h, \ell}^{\mathcal{J}}(t) \|^2 \right) \end{aligned}$$

where

$$\begin{aligned} \sup_{t \in [t_0, t_n]} \| e_{\tau h, \ell}^{\mathcal{J}}(t) \|^2 & \leq C(1 + \tau) \exp((1 + \delta_{1, k-2\ell})(t_{n-1} - t_0)) \\ & \quad \left(\| e_{\tau h, \ell}^{\mathcal{J}}(t_0^-) \|^2 + \sum_{\nu=1}^{n-1} (2 + \tau_{\nu}^{-1}) \| \omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)}) \|_V^2 + \mathcal{J}_{[1, n]} [\| u^{(\ell+1)} - R_h u^{(\ell+1)} \|^2] \right. \\ & \quad + \tau \mathcal{J}_{[1, n]} [\| u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} \|_{\tilde{V}}^2] + \mathcal{J}_{[1, n]} [\| u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} \|_V^2] \\ & \quad \left. + \mathcal{J}_{[1, n]} [\| \Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)}) \|^2] + \tau \delta_{1, k-2\ell} \| e_{\tau h, \ell}^{\mathcal{J}}(t_0^-) \|_V^2 \right) \end{aligned}$$

with $\omega_{\nu}^{\mathcal{J}}(\cdot)$ as defined in (4.7). The exponential factor can be dropped if $\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)}) = 0$ for all $\nu = 1, \dots, n-1$.

Proof. We decompose the error as usual in the three terms

$$u^{(\ell)} - u_{\tau h}^{(\ell)} = (u^{(\ell)} - R_h u^{(\ell)}) + (R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}) + e_{\tau h, \ell}^{\mathcal{J}}, \quad e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}.$$

The stability of R_h is used to estimate the second term by

$$\begin{aligned} \sup_{t \in [t_0, t_n]} \| R_h u^{(\ell)}(t) - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}(t) \|^2 & \leq C_{\text{emb}}^2 \sup_{t \in [t_0, t_n]} \| R_h (u^{(\ell)}(t) - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}(t)) \|_V^2 \\ & \leq C \sup_{t \in [t_0, t_n]} \| u^{(\ell)}(t) - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}(t) \|_V^2. \end{aligned}$$

Hence, using the triangle inequality, the first desired statement follows easily.

We now analyze the third summand $e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}$. Let $s \in [t_0, t_n]$ be fixed, then it holds $s \in [t_{\nu-1}, t_{\nu}]$ for some $1 \leq \nu \leq n$ and

$$e_{\tau h, \ell}^{\mathcal{J}}(s) = e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-) - \int_s^{t_{\nu}} \partial_t e_{\tau h, \ell}^{\mathcal{J}} dt.$$

From this we derive by the triangle inequality and the Cauchy–Schwarz inequality

$$\begin{aligned} \| e_{\tau h, \ell}^{\mathcal{J}}(s) \|^2 & \leq 2 \| e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-) \|^2 + 2 \left\| \int_s^{t_{\nu}} \partial_t e_{\tau h, \ell}^{\mathcal{J}} dt \right\|^2 \leq 2 \| e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-) \|^2 + 2 \left(\int_s^{t_{\nu}} \| \partial_t e_{\tau h, \ell}^{\mathcal{J}} \| dt \right)^2 \\ & \leq 2 \| e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-) \|^2 + 2(t_{\nu} - s) \int_s^{t_{\nu}} \| \partial_t e_{\tau h, \ell}^{\mathcal{J}} \|^2 dt. \end{aligned}$$

Hence, we obtain

$$\sup_{t \in [t_0, t_n]} \|e_{\tau h, \ell}^{\mathcal{J}}(t)\|^2 \leq 2 \max_{\nu=1, \dots, n} \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-)\|^2 + \tau_{\nu} \int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \right).$$

If $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd), the right-hand side then can be bounded by Lemma 4.18 (and Lemma 4.8). Hence, in this case the second desired estimate follows easily.

If $k - 2\ell = 0$ ($\Leftrightarrow k$ is even), we further use a norm equivalence in the finite dimensional polynomial space, cf. [52, (12.18), p. 210], that gives

$$\tau_{\nu} \int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \leq \tilde{C} \int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 (t - t_{\nu-1}) dt$$

for some constant $\tilde{C} > 0$ independent of ν (and τ_{ν}). Thus, it holds

$$\sup_{t \in [t_0, t_n]} \|e_{\tau h, \ell}^{\mathcal{J}}(t)\|^2 \leq C \max_{\nu=1, \dots, n} \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_{\nu}^-)\|^2 + \int_{I_{\nu}} \|\partial_t e_{\tau h, \ell}^{\mathcal{J}}\|^2 (t - t_{\nu-1}) dt \right).$$

Then, in order to bound the right-hand side, Lemma 4.8 and Lemma 4.27 can be applied. To simplify the terms, note that due to $\|w\| \leq C_{\text{emb}} \|w\|_V$ for all $w \in V$ it also holds the estimate $\|w\|_{V'} \leq C_{\text{emb}} \|w\|$ for all $w \in H \subset V'$. \square

Theorem 4.32

Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} & \sup_{t \in [t_0, t_n]} \|(u^{(\ell)} - u_{\tau h}^{(\ell)})(t)\|^2 \\ & \leq C \left[h^{2(\kappa+\sigma)} \left(\|u^{(\ell)}\|_{C([t_0, t_n], H^{\kappa+1}(\Omega))}^2 + \|u^{(\ell)}\|_{H^1((t_0, t_n), H^{\kappa+1}(\Omega))}^2 \right) \right. \\ & \quad + \delta_{1, k-2\ell} \tilde{\sigma} \tau h^{2\kappa} \|u^{(\ell)}(t_0)\|_{H^{\kappa+1}(\Omega)}^2 \\ & \quad + \tau^{2(r-\ell+1)} \left(\|u^{(\ell)}\|_{W^{r-\ell+1, \infty}([t_0, t_n], H^1(\Omega))}^2 + \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^1(\Omega))}^2 \right. \\ & \quad \left. \left. + \tau \|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^2(\Omega))}^2 + \|f\|_{H^{r+1}((t_0, t_n), L^2(\Omega))}^2 \right) \right], \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise. Moreover, $\tilde{\sigma} = 0$ if $\tilde{P}_h^0 = R_h$ and $\tilde{\sigma} = 1$ if $\tilde{P}_h^0 = P_h$.

Proof. Starting with Lemma 4.31 for the case $\mathcal{J}_{\nu} = \int_{I_{\nu}}$, the statement follows from projection error estimates. Most terms have been already bounded earlier. The remaining terms can be bounded using quite similar arguments. \square

Remark 4.33

The convergence behavior with respect to time shown in Theorem 4.32 is of the expected order $r - \ell + 1$. With respect to space, we find order $\kappa + \sigma$ in the dG-like case. For cGP-like methods this spatial order is only obtained if the Ritz projection R_h is chosen for the approximation of the initial value. However, it is not yet clear whether this choice is really necessary to gain spatial order $\kappa + \sigma$ instead of κ or whether we only need it due to our proof technique. \clubsuit

4.1.6 Supercloseness and its consequences

Supercloseness phenomena occur for many different discretizations of various differential problems. Supercloseness here means that the numerical solution is somewhat closer to a certain projection of the exact solution than to the exact solution itself. Often this property can be used to improve the method or the estimates.

Usually, supercloseness is strongly connected to specific properties of the involved projection operator. Therefore, we have a look on an interesting feature of the temporal projection operator $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ at first. The result then will be exploited later.

Lemma 4.34

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $\ell = \lfloor \frac{k}{2} \rfloor$. Denote by X a Banach space over \mathbb{R} . Then, it holds

$$\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v = \tilde{\Pi}_{k-2\ell}^{r-\ell} v = \mathcal{I}_{k-2\ell}^{r-\ell} v \quad \forall v \in P_{r-\ell+1}(I_n, X).$$

Proof. Taking a closer look at the definition of $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$, we see that for $v \in P_{r-\ell+1}(I_n, X)$ the integrator \mathcal{J}_n can be replaced by the integral over I_n due to our assumption (4.5a) that \mathcal{J}_n integrates polynomials of maximal degree $2r - k$ exactly. Together with the observations of Remark 4.2, the first identity is shown.

It remains to prove the second identity. Recalling the definitions of $\tilde{\Pi}_{k-2\ell}^{r-\ell}$ and $\mathcal{I}_{k-2\ell}^{r-\ell}$, we immediately get that

$$\tilde{\Pi}_{k-2\ell}^{r-\ell} v(t_{n-1}^+) = v(t_{n-1}^+) = \mathcal{I}_{k-2\ell}^{r-\ell} v(t_{n-1}^+), \quad \text{if } k - 2\ell = 1,$$

and

$$\tilde{\Pi}_{k-2\ell}^{r-\ell} v(t_n^-) = v(t_n^-) = \mathcal{I}_{k-2\ell}^{r-\ell} v(t_n^-)$$

for any $v \in C(\bar{I}_n, X)$. Moreover, using the exactness of the quadrature rule $Q_{k-2\ell}^{r-\ell}$ up to polynomial degree $2r - k$, we gain for $v \in P_{r-\ell+1}(I_n, X)$ that

$$\int_{I_n} \tilde{\Pi}_{k-2\ell}^{r-\ell} v w \, dt = \int_{I_n} v w \, dt = Q_{k-2\ell}^{r-\ell} [v w] = Q_{k-2\ell}^{r-\ell} [\mathcal{I}_{k-2\ell}^{r-\ell} v w] = \int_{I_n} \mathcal{I}_{k-2\ell}^{r-\ell} v w \, dt \quad \forall w \in P_{r-k+\ell-1}(I_n).$$

Since both $\tilde{\Pi}_{k-2\ell}^{r-\ell} v$ and $\mathcal{I}_{k-2\ell}^{r-\ell} v$ are X -valued polynomials of degree $r - \ell$, which are uniquely determined by these $r - \ell + 1$ conditions, it follows that $\tilde{\Pi}_{k-2\ell}^{r-\ell} v = \mathcal{I}_{k-2\ell}^{r-\ell} v$ holds for all $v \in P_{r-\ell+1}(I_n, X)$. \square

The result of Lemma 4.34 suggests that $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ provides improved approximation properties in the quadrature points of $Q_{k-2\ell}^{r-\ell}$. Now, inspecting the estimates for $e_{\tau h, \ell}^{\mathcal{J}}$ derived above, see Lemma 4.10, we note that the term

$$\mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right]$$

is occurring. So, from our observations it seems that $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$ is an appropriate choice to gain improved estimates. This choice or change of the integrator, however, is possible under certain assumptions on g only, for example, if $g^{(\ell)}$ is a polynomial in time of maximal degree $r - \ell$ on every I_ν , $\nu = 1, \dots, n$.

Lemma 4.35

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, $\ell = \lfloor \frac{k}{2} \rfloor$, and set $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$. Then, in the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) it holds

$$\begin{aligned} \mathcal{J}_{[1,n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] &= \sum_{\nu=1}^n Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \\ &\leq C \tau^{2(r-\ell+2)} \|u^{(\ell)}\|_{H^{r-\ell+2}((t_0, t_n), H^1(\Omega))}^2. \end{aligned}$$

Proof. Let X denote some Banach space over \mathbb{R} . We start defining another operator $\tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} : H^1(I_n, X) \cap C^{k_{\mathcal{J}}+1}(\bar{I}_n, X) \rightarrow P_{r-\ell+1}(I_n, X)$ by

$$\begin{aligned} (v - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} v)(t_{n-1}^+) &= 0, \quad \text{if } k-2\ell = 1, \\ \mathcal{J}_n \left[\partial_t (v - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} v) w \right] + \delta_{0, k-2\ell} (v - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} v)(t_{n-1}^+) w(t_{n-1}^+) &= 0 \quad \forall w \in P_{r-k+\ell}(I_n), \\ \int_{I_n} \partial_t (v - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} v) w \, dt + \delta_{0, k-2\ell} (v - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} v)(t_{n-1}^+) w(t_{n-1}^+) &= 0 \quad \forall w \in \tilde{P}_{r-k+\ell+1}(I_n) \end{aligned}$$

where $\tilde{P}_{r-k+\ell+1}(I_n) := P_{r-k+\ell+1}(I_n) \setminus P_{r-k+\ell}(I_n)$. One easily verifies that $\tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}}$ is a well-defined projection operator onto X -valued polynomials of degree $r-\ell+1$, cf. Definition C.11. Moreover, with the findings of Lemma 4.34 it follows

$$\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v = \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} v = \mathcal{I}_{k-2\ell}^{r-\ell} \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} v \quad \text{for all } v \in C^{\max\{0, k_{\mathcal{J}}\}+1}(\bar{I}_n, X).$$

Thus, we obtain that

$$\begin{aligned} Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] &= Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - \mathcal{I}_{k-2\ell}^{r-\ell} \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} u^{(\ell)}\|_V^2 \right] \\ &= Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} u^{(\ell)}\|_V^2 \right]. \end{aligned}$$

Then, local projection error estimates, see Lemma B.9 or also cf. [21, Theorem 3.1.4, p. 121] or [25, Remark 1.112, p. 62] (where the real-valued case is handled), yield for the term on the right-hand side that

$$\begin{aligned} Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} u^{(\ell)}\|_V^2 \right] &\leq C \tau_\nu \|u^{(\ell)} - \tilde{\Pi}_{k-2\ell, *}^{r-\ell+1, \mathcal{J}} u^{(\ell)}\|_{C(I_\nu, H^1(\Omega))}^2 \\ &\leq C \tau_\nu \left(C \tau_\nu^{r-\ell+2-1/2} \|u^{(\ell)}\|_{H^{r-\ell+2}(I_\nu, H^1(\Omega))} \right)^2 \leq C \tau_\nu^{2(r-\ell+2)} \|u^{(\ell)}\|_{H^{r-\ell+2}(I_\nu, H^1(\Omega))}^2, \end{aligned}$$

which, after summation over $\nu = 1, \dots, n$, gives the desired bound. \square

In previous subsections, we have always chosen $\mathcal{J}_\nu = \int_{I_\nu}$ when concrete convergence orders were shown. In this case, of course, no integrator error occurs. Since for the supercloseness studies we choose $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$, an examination of $\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})$, $\nu = 1, \dots, n$, becomes necessary.

Lemma 4.36

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, $\ell = \lfloor \frac{k}{2} \rfloor$, and set $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$. Moreover, let $j \in \mathbb{Z}$ with $1 \leq j \leq 2r - k + 1$ and $\omega_\nu^{\mathcal{J}}(\cdot)$ be defined by (4.7). Then, in the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) it holds

$$\sum_{\nu=1}^{n-1} (2 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|^2 \leq C (1 + 2\tau) \tau^{2j} \|u^{(\ell+1)}\|_{H^j((t_0, t_{n-1}), H^1(\Omega))}^2.$$

Proof. First of all, by definition we have that

$$\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)}) = \int_{I_\nu} R_h u^{(\ell+1)} dt - Q_{k-2\ell, \nu}^{r-\ell} [R_h u^{(\ell+1)}].$$

Furthermore, let $\check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q}$ denote the local version of an interpolation operator such that $\check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} w \in P_{2r-k}(I_\nu, X)$ interpolates $w \in C(\bar{I}_\nu, X)$ in the $r - \ell + 1$ quadrature points of $Q_{k-2\ell, \nu}^{r-\ell}$ and in $r - k + \ell$ additional points. The exactness of the quadrature rule for polynomials up to degree $2r - k$ yields

$$\begin{aligned} Q_{k-2\ell, \nu}^{r-\ell} [R_h u^{(\ell+1)}] &= Q_{k-2\ell, \nu}^{r-\ell} [\check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} (R_h u^{(\ell+1)})] = \int_{I_\nu} \check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} (R_h u^{(\ell+1)}) dt \\ &= \int_{I_\nu} R_h \check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} u^{(\ell+1)} dt. \end{aligned}$$

So, using the Cauchy–Schwarz inequality and the stability of R_h in the V -norm, it follows

$$\begin{aligned} \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|^2 &= \left\| \int_{I_\nu} R_h u^{(\ell+1)} dt - Q_{k-2\ell, \nu}^{r-\ell} [R_h u^{(\ell+1)}] \right\|^2 \\ &= \left\| \int_{I_\nu} R_h u^{(\ell+1)} - R_h \check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} u^{(\ell+1)} dt \right\|^2 \leq C_{\text{emb}}^2 \tau_\nu \int_{I_\nu} \|R_h (u^{(\ell+1)} - \check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} u^{(\ell+1)})\|_V^2 dt \\ &\leq C \tau_\nu \int_{I_\nu} \|u^{(\ell+1)} - \check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} u^{(\ell+1)}\|_V^2 dt. \end{aligned} \tag{4.15}$$

Therefore, (standard) error estimates for the interpolation operator $\check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q}$ imply

$$\begin{aligned} \sum_{\nu=1}^{n-1} (2 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|^2 \\ \leq C (1 + 2\tau) \int_{t_0}^{t_{n-1}} \|u^{(\ell+1)} - \check{\mathcal{I}}_{k-2\ell}^{r-\ell, Q} u^{(\ell+1)}\|_V^2 dt \leq C (1 + 2\tau) \tau^{2j} \|u^{(\ell+1)}\|_{H^j((t_0, t_{n-1}), H^1(\Omega))}^2, \end{aligned}$$

which completes the proof. \square

Remark 4.37

If u is sufficiently smooth (especially $u^{(\ell+1)} \in H^1((t_0, t_n), H_0^1(\Omega)) \cap H^{2r-k+1}((t_0, t_n), H^1(\Omega)))$, the estimate of Lemma 4.36 ensures a behavior of the quadrature error term of $\mathcal{O}(\tau^{2(2r-k+1)})$. Note that this is in line with the superconvergence order in the time (mesh) points of $2r - k + 1$ seen in Subsection 1.2.3 in the case of non-stiff initial value problems. \clubsuit

It seems that supercloseness only can be proven if the right-hand side g of the discrete method fulfills certain conditions.

Assumption 4.3

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, $\ell = \lfloor \frac{k}{2} \rfloor$. We assume that there is an approximation operator \mathcal{A} satisfying

$$\mathcal{I}_{k-2\ell}^{r-\ell}((\mathcal{A}f)^{(\ell)} - g^{(\ell)}) = 0 \quad \text{and} \quad \|(f - \mathcal{A}f)^{(\ell)}\|_{C(I_n, V')} \leq C \tau_n^{r-\ell+2-1/2} \|f\|_{H^{r+2}(I_n, V')} \quad (4.16)$$

with C independent of τ_n and $1 \leq n \leq N$.

Remark 4.38

We want to give some relevant examples for g where (4.16) can be satisfied.

- (i) For $g = \mathcal{I}_k^r f$: Similar to $\widehat{\mathcal{I}}_k^r$ let the operator $\widehat{\mathcal{I}}_{k,\diamond}^{r+1} : C^{\lfloor \frac{k}{2} \rfloor}([-1, 1]) \rightarrow P_{r+1}([-1, 1])$ use the same interpolation points as $\widehat{\mathcal{I}}_k^r$, cf. (1.15), and one additional interpolation point $\hat{t}^\diamond \in (-1, 1)$. By transformation with T_n from (1.7) we also get interpolation operators $\mathcal{I}_{k,\diamond}^{r+1}$ on \bar{I}_n , $n = 1, \dots, N$.

We now choose $\mathcal{A}f = \mathcal{I}_{k,\diamond}^{r+1} f$. Since, obviously, $g = \mathcal{I}_k^r f = \mathcal{I}_k^r \mathcal{I}_{k,\diamond}^{r+1} f$, we have on I_n that

$$\mathcal{A}f - g = (\text{Id} - \mathcal{I}_k^r) \mathcal{I}_{k,\diamond}^{r+1} f = c \phi_n \quad \text{for some } c \in \mathbb{R}, \phi_n \in P_{r+1}(I_n),$$

where ϕ_n vanishes in all interpolation points of \mathcal{I}_k^r .

Because of $\ell = \lfloor \frac{k}{2} \rfloor$ and $k - \ell - 1 = \lfloor \frac{k-1}{2} \rfloor$ as well as due to the construction of \mathcal{I}_k^r , it holds that $\phi_n = \phi \circ T_n^{-1}$ is the local version of the function $\phi \in P_{r+1}([-1, 1])$ given by

$$\phi(\hat{t}) = (1 - \hat{t})^{\ell+1} (1 + \hat{t})^{k-\ell} P_{r-k}^{(\ell+1, k-\ell)}(\hat{t}). \quad (4.17)$$

Here, $P_{r-k}^{(\ell+1, k-\ell)}$ denotes the $(r-k)$ th Jacobi-polynomial with respect to the weight $(1 - \hat{t})^{\ell+1} (1 + \hat{t})^{k-\ell}$, see Appendix A.2 for details. An easy conclusion from Rodrigues' formula, see (A.2), furthermore gives that

$$\phi^{(\ell)}(\hat{t}) = \tilde{c} (1 - \hat{t}) (1 + \hat{t})^{k-2\ell} P_{r-k+\ell}^{(1, k-2\ell)}(\hat{t})$$

and so $\phi^{(\ell)}$ vanishes in the interpolation points of $\widehat{\mathcal{I}}_{k-2\ell}^{r-\ell}$.

We therefore conclude $\mathcal{I}_{k-2\ell}^{r-\ell}((\mathcal{A}f)^{(\ell)} - g^{(\ell)}) = c \mathcal{I}_{k-2\ell}^{r-\ell} \phi_n^{(\ell)} = 0$. Furthermore, since $\mathcal{A}f = \mathcal{I}_{k,\diamond}^{r+1} f$ is a Hermite interpolation of f of polynomial degree $r+1$, the error estimates are clear.

- (ii) For $g = \mathcal{C}_k^r f$: Here, we choose $\mathcal{A}f = \mathcal{C}_{k+2}^{r+1} f$. Since, obviously, $g = \mathcal{C}_k^r f = \mathcal{I}_k^r \mathcal{C}_{k+2}^{r+1} f$, we again have on I_n that

$$\mathcal{A}f - g = (\text{Id} - \mathcal{I}_k^r) \mathcal{C}_{k+2}^{r+1} f = c \phi_n \quad \text{for some } c \in \mathbb{R}, \phi_n \in P_{r+1}(I_n),$$

where ϕ_n vanishes in all interpolation points of \mathcal{I}_k^r . Therefore, we can conclude as in (i). ♣

Lemma 4.39

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, $\ell = \lfloor \frac{k}{2} \rfloor$, and set $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$. Moreover, assume that there is an approximation operator \mathcal{A} satisfying Assumption 4.3. Then, in the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) it holds

$$\begin{aligned} \mathcal{J}_{[1,n]} \left[\left\| \Pi_{r-k+\ell}^{\mathcal{J}} (f^{(\ell)} - g^{(\ell)}) \right\|_{V'}^2 \right] &= \sum_{\nu=1}^n Q_{k-2\ell, \nu}^{r-\ell} \left[\left\| \Pi_{r-k+\ell}^{\mathcal{J}} (f^{(\ell)} - g^{(\ell)}) \right\|_{V'}^2 \right] \\ &\leq C \tau^{2(r-\ell+2)} \|f\|_{H^{r+2}((t_0, t_n), H^{-1}(\Omega))}^2. \end{aligned}$$

Proof. Because of Remark 4.7, $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$, and $\mathcal{I}_{k-2\ell}^{r-\ell}((\mathcal{A}f)^{(\ell)} - g^{(\ell)}) = 0$, we get

$$\mathcal{J}_\nu \left[\left\| \Pi_{r-k+\ell}^{\mathcal{J}} (f^{(\ell)} - g^{(\ell)}) \right\|_{V'}^2 \right] \leq C Q_{k-2\ell, \nu}^{r-\ell} \left[\|f^{(\ell)} - g^{(\ell)}\|_{V'}^2 \right] = C Q_{k-2\ell, \nu}^{r-\ell} \left[\|f^{(\ell)} - (\mathcal{A}f)^{(\ell)}\|_{V'}^2 \right].$$

With similar arguments as used in the proof of Lemma 4.35 but here applying the assumptions on the error of \mathcal{A} , we further conclude

$$\begin{aligned} Q_{k-2\ell, \nu}^{r-\ell} \left[\|f^{(\ell)} - (\mathcal{A}f)^{(\ell)}\|_{V'}^2 \right] &\leq C \tau_\nu \|f^{(\ell)} - (\mathcal{A}f)^{(\ell)}\|_{C(I_\nu, H^{-1}(\Omega))}^2 \\ &\leq C \tau_\nu \left(C \tau_\nu^{r-\ell+2-1/2} \|f\|_{H^{r+2}(I_\nu, H^{-1}(\Omega))} \right)^2 \leq C \tau_\nu^{2(r-\ell+2)} \|f\|_{H^{r+2}(I_\nu, H^{-1}(\Omega))}^2. \end{aligned}$$

The desired statement follows easily by summation over $\nu = 1, \dots, n$. \square

Summarizing, we get supercloseness results for $e_{\tau h, \ell}^{\mathcal{J}}$ with $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$ under certain assumptions on g . This also implies a lower order superconvergence result for the ℓ th derivative of the error in the time mesh points and an improved convergence order with respect to the quadrature formula $Q_{k-2\ell}^{r-\ell}$. Hereby, recall that the choice $\mathcal{J}_n = Q_{k-2\ell, n}^{r-\ell}$ is possible only if $g^{(\ell)}|_{I_n} \in P_{r-\ell}(I_n, V')$ for all $n = 1, \dots, N$.

Theorem 4.40 (Supercloseness result)

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, $\ell = \lfloor \frac{k}{2} \rfloor$, and set $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$. Moreover, assume that there is an approximation operator \mathcal{A} satisfying Assumption 4.3. Then, in the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) it holds

$$\begin{aligned} &\max_{\nu=1, \dots, n} \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|^2 + \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt + \delta_{0, k-2\ell} \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\ &\leq C(1 + \delta_{1, k-2\ell}(t_n - t_0)) \exp(t_{n-1} - t_0) \\ &\quad \left[h^{2(\kappa+\sigma)} \left((t_n - t_0) \|u^{(\ell+1)}\|_{C((t_0, t_n), H^{\kappa+1}(\Omega))}^2 + \|u^{(\ell)}(t_0)\|_{H^{\kappa+1}(\Omega)}^2 \right) \right. \\ &\quad + \tau^{2(r-\ell+2)} \left(\|u^{(\ell)}\|_{H^{r-\ell+2}((t_0, t_n), H^1(\Omega))}^2 + \|f\|_{H^{r+2}((t_0, t_n), H^{-1}(\Omega))}^2 \right) \\ &\quad \left. + (1 + \tau) \tau^{2 \min\{r-\ell+2, 2r-k+1\}} \|u^{(\ell+1)}\|_{H^{\min\{r-\ell+2, 2r-k+1\}}((t_0, t_{n-1}), H^1(\Omega))}^2 \right], \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Combining Lemma 4.8 and the second estimate of Lemma 4.10, we find

$$\begin{aligned} & \|e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)\|^2 + \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt + \delta_{0, k-2\ell} \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt \\ & \leq C(1 + \delta_{1, k-2\ell}(t_n - t_0)) \exp(t_{n-1} - t_0) \\ & \quad \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 + \sum_{\nu=1}^{n-1} (2 + \tau_\nu^{-1}) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|^2 + \mathcal{J}_{[1, n]} \left[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2 \right] \right. \\ & \quad \left. + \mathcal{J}_{[1, n]} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] + \mathcal{J}_{[1, n]} \left[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'}^2 \right] \right) \end{aligned}$$

with $\omega_\nu^{\mathcal{J}}(\cdot)$ as defined in (4.7). Merging this with the Lemmas 4.35, 4.36, and 4.39 as well as estimating the remaining spatial error terms by (3.16a) and (3.16b), we are done. \square

Corollary 4.41 (Consequences of the supercloseness result)

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, $\ell = \lfloor \frac{k}{2} \rfloor$, and set $\mathcal{Q}_\nu^{r-\ell} = Q_{k-2\ell, \nu}^{r-\ell}$. Moreover, assume that there is an approximation operator \mathcal{A} satisfying Assumption 4.3. Then, in the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) it holds

$$\begin{aligned} & \max_{\nu=1, \dots, n} \|(u^{(\ell)} - u_{\tau h}^{(\ell)})(t_\nu^-)\|^2 + \sum_{\nu=1}^n Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - u_{\tau h}^{(\ell)}\|^2 \right] \\ & \leq C(1 + \delta_{1, k-2\ell}(t_n - t_0)) \exp(t_{n-1} - t_0) \\ & \quad \left[h^{2(\kappa+\sigma)} \left((t_n - t_0) \|u^{(\ell+1)}\|_{C((t_0, t_n), H^{\kappa+1}(\Omega))}^2 + (1 + (t_n - t_0)) \|u^{(\ell)}\|_{C((t_0, t_n), H^{\kappa+1}(\Omega))}^2 \right) \right. \\ & \quad + \tau^{2(r-\ell+2)} \left(\|u^{(\ell)}\|_{H^{r-\ell+2}((t_0, t_n), H^1(\Omega))}^2 + \|f\|_{H^{r+2}((t_0, t_n), H^{-1}(\Omega))}^2 \right) \\ & \quad \left. + (1 + \tau) \tau^{2 \min\{r-\ell+2, 2r-k+1\}} \|u^{(\ell+1)}\|_{H^{\min\{r-\ell+2, 2r-k+1\}}((t_0, t_n), H^1(\Omega))}^2 \right], \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Using the error decomposition

$$u^{(\ell)} - u_{\tau h}^{(\ell)} = (u^{(\ell)} - R_h u^{(\ell)}) + (R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}) + e_{\tau h, \ell}^{\mathcal{J}}, \quad e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)},$$

we find together with (4.7) that

$$\|(u^{(\ell)} - u_{\tau h}^{(\ell)})(t_\nu^-)\| \leq \|(u^{(\ell)} - R_h u^{(\ell)})(t_\nu^-)\| + \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\| + \|e_{\tau h, \ell}^{\mathcal{J}}(t_\nu^-)\|.$$

These terms can be estimated by (3.16), Lemma 4.36, and Theorem 4.40.

Moreover, we obtain

$$\begin{aligned} & Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - u_{\tau h}^{(\ell)}\|^2 \right] \\ & \leq 3 \left(Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - R_h u^{(\ell)}\|^2 \right] + Q_{k-2\ell, \nu}^{r-\ell} \left[\|R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|^2 \right] + Q_{k-2\ell, \nu}^{r-\ell} \left[\|e_{\tau h, \ell}^{\mathcal{J}}\|^2 \right] \right) \\ & \leq C \left(\tau_\nu \|u^{(\ell)} - R_h u^{(\ell)}\|_{C(I_\nu, L^2(\Omega))}^2 + Q_{k-2\ell, \nu}^{r-\ell} \left[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2 \right] + \int_{I_\nu} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt \right). \end{aligned}$$

Here, we used $V \hookrightarrow H$ and the stability of R_h with respect to the norm in V to bound the second term. Moreover, we exploited

$$Q_{k-2\ell,\nu}^{r-\ell} \left[\|e_{\tau h,\ell}^{\mathcal{J}}\|^2 \right] \leq C \tau_\nu \sup_{t \in I_\nu} \|e_{\tau h,\ell}^{\mathcal{J}}\|^2 \leq C \int_{I_\nu} \|e_{\tau h,\ell}^{\mathcal{J}}\|^2 dt,$$

where the last step follows from a norm equivalence since the function $t \mapsto \|e_{\tau h,\ell}^{\mathcal{J}}(t)\|^2$ is in $P_{2(r-\ell)}(I_\nu)$. Summing over $\nu = 1, \dots, n$ and applying (3.16), Lemma 4.35, and Theorem 4.40, we easily complete the proof. \square

Remark 4.42

Theorem 4.40 shows that under certain assumptions the temporal convergence order $r - \ell + 2$ can be obtained for the fully discrete error $e_{\tau h,\ell}^{\mathcal{J}}$ with $\mathcal{J}_\nu = Q_{k-2\ell,\nu}^{r-\ell}$. This is one order higher than in the respective results for the error, cf. Theorem 4.11 and Corollary 4.13.

In Corollary 4.41 we then see an improved temporal convergence behavior for the ℓ th derivative of the error in the time (mesh) points compared to the pointwise error estimate of Theorem 4.32. Moreover, we also obtain an improved estimate if we consider not the exactly integrated, squared H -norm of $u^{(\ell)} - u_{\tau h}^{(\ell)}$, as in Theorem 4.11, but the numerically integrated using quadrature formula $Q_{k-2\ell}^{r-\ell}$. \clubsuit

4.2 Error estimates in the time (mesh) points

We now have estimates for the ℓ th derivative of the error. However, for $k \geq 2$ ($\Leftrightarrow \ell \geq 1$) bounds for the error $u - u_{\tau h}$ itself still are missing. In this section, we want to derive such bounds at least in the time (mesh) points.

4.2.1 Exploiting the collocation conditions

Recall that $\ell = \lfloor \frac{k}{2} \rfloor$, which is the highest derivative order that appears in the collocation conditions on the right interval end. Set $v_{h,i} := (R_h u - u_{\tau h})^{(i)}$ for $0 \leq i \leq \ell - 1$. The V -ellipticity of $a(\cdot, \cdot)$ and the definition of R_h yield

$$\begin{aligned} \alpha \|v_{h,i}(t_n^-)\|_V^2 &\leq a(R_h u^{(i)}(t_n^-) - u_{\tau h}^{(i)}(t_n^-), v_{h,i}(t_n^-)) = a(u^{(i)}(t_n^-) - u_{\tau h}^{(i)}(t_n^-), v_{h,i}(t_n^-)) \\ &= -(u^{(i+1)}(t_n^-) - u_{\tau h}^{(i+1)}(t_n^-), v_{h,i}(t_n^-)) + \langle f^{(i)}(t_n^-) - g^{(i)}(t_n^-), v_{h,i}(t_n^-) \rangle_{V',V}, \end{aligned}$$

where we also used that $u^{(i)}$ solves the i th (temporal) derivative of the differential equation and that $u_{\tau h}$ satisfies (3.17b). Hence, by Cauchy–Schwarz’ inequality, the definition of the V' -norm, and because of $V \hookrightarrow H$, we obtain

$$\begin{aligned} \|(R_h u - u_{\tau h})^{(i)}(t_n^-)\|_V &\leq \frac{C_{\text{emb}}}{\alpha} \|u^{(i+1)}(t_n^-) - u_{\tau h}^{(i+1)}(t_n^-)\| + \frac{1}{\alpha} \|f^{(i)}(t_n^-) - g^{(i)}(t_n^-)\|_{V'}, \\ &\quad \text{for all } 0 \leq i \leq \ell - 1. \end{aligned} \quad (4.18)$$

By the triangle inequality we recursively conclude for all $0 \leq i \leq \ell - 1$ that

$$\begin{aligned}
 \|(u - u_{\tau h})^{(i)}(t_n^-)\| &\leq \|(u - R_h u)^{(i)}(t_n^-)\| + \|(R_h u - u_{\tau h})^{(i)}(t_n^-)\| \\
 &\leq \left(\frac{C_{\text{emb}}^2}{\alpha}\right)^{\ell-i} \|(u - u_{\tau h})^{(\ell)}(t_n^-)\| \\
 &\quad + \sum_{j=i}^{\ell-1} \left(\frac{C_{\text{emb}}^2}{\alpha}\right)^{j-i} \left(\|(u - R_h u)^{(j)}(t_n^-)\| + \frac{C_{\text{emb}}}{\alpha} \|(f - g)^{(j)}(t_n^-)\|_{V'} \right).
 \end{aligned} \tag{4.19}$$

Accordingly, to get estimates for the error in the time (mesh) points, it is sufficient to have a suitable bound for the norm of the ℓ th derivative of the error there. Such bounds can be derived quite similar to those of Lemma 4.10. Therefore, as consequence we get the following result.

Lemma 4.43

Let $0 \leq i \leq \ell$. Then, for all $n = 1, \dots, N$ it holds

$$\begin{aligned}
 &\|(u - u_{\tau h})^{(i)}(t_n^-)\|^2 \\
 &\leq C \sum_{j=i}^{\ell} \|(u - R_h u)^{(j)}(t_n^-)\|^2 + C \sum_{j=i}^{\ell-1} \|(f - g)^{(j)}(t_n^-)\|_{V'}^2 \\
 &\quad + C \exp(t_{n-1} - t_0) \\
 &\quad \left(\|e_{\tau h, \ell}^{\mathcal{J}}(t_0^-)\|^2 + \sum_{\nu=1}^n (2 + \tau_{\nu}^{-1}) \|\omega_{\nu}^{\mathcal{J}}(R_h u^{(\ell)})\|^2 + \mathcal{J}_{[1, n]}[\|u^{(\ell+1)} - R_h u^{(\ell+1)}\|^2] \right. \\
 &\quad \left. + \mathcal{J}_{[1, n]}[\|u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}\|_V^2] + \mathcal{J}_{[1, n]}[\|\Pi_{r-k+\ell}^{\mathcal{J}}(f^{(\ell)} - g^{(\ell)})\|_{V'}^2] \right)
 \end{aligned}$$

with $\omega_{\nu}^{\mathcal{J}}(\cdot)$ as defined in (4.7).

Proof. Because of (4.19), it remains to derive a suitable bound for the norm of the ℓ th derivative of the error in the time (mesh) points. To this end, we split the error as

$$u^{(\ell)} - u_{\tau h}^{(\ell)} = (u^{(\ell)} - R_h u^{(\ell)}) + (R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)}) + e_{\tau h, \ell}^{\mathcal{J}}, \quad e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}.$$

From (4.7) we get for the second summand

$$(R_h u^{(\ell)} - R_h \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} u^{(\ell)})(t_n^-) = (R_h u^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} R_h u^{(\ell)})(t_n^-) = \omega_n^{\mathcal{J}}(R_h u^{(\ell)}).$$

Hence, we conclude that

$$\begin{aligned}
 \|(u - u_{\tau h})^{(i)}(t_n^-)\| &\leq \left(\frac{C_{\text{emb}}^2}{\alpha}\right)^{\ell-i} \left(\|(u - R_h u)^{(\ell)}(t_n^-)\| + \|\omega_n^{\mathcal{J}}(R_h u^{(\ell)})\| + \|e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)\| \right) \\
 &\quad + \sum_{j=i}^{\ell-1} \left(\frac{C_{\text{emb}}^2}{\alpha}\right)^{j-i} \left(\|(u - R_h u)^{(j)}(t_n^-)\| + \frac{C_{\text{emb}}}{\alpha} \|(f - g)^{(j)}(t_n^-)\|_{V'} \right).
 \end{aligned}$$

The desired statement now follows from Lemma 4.8. \square

Proposition 4.44

Let $1 \leq n \leq N$ and $0 \leq i \leq \ell$. Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{f, \Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} \|(u - u_{\tau h})^{(i)}(t_n^-)\|^2 &\leq Ch^{2(\kappa+\sigma)} \left(\|u^{(\ell)}\|_{H^1((t_0, t_n), H^{\kappa+1}(\Omega))}^2 + \sum_{j=i}^{\ell-1} \|u^{(j)}(t_n^-)\|_{H^{\kappa+1}(\Omega)}^2 \right) \\ &\quad + C\tau^{2(r-\ell+1)} \left(\|u^{(\ell)}\|_{H^{r-\ell+1}((t_0, t_n), H^1(\Omega))}^2 + \|f\|_{H^{r+1}((t_0, t_n), H^{-1}(\Omega))}^2 \right), \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Applying Lemma 4.43, used with $\mathcal{J}_\nu = \int_{I_\nu}$, we need to bound certain projection errors only. Most of them have already been estimated in the proof of Theorem 4.11. The remaining terms can be bounded using (3.16). \square

Remark 4.45

For $0 \leq i \leq \ell$ we would expect convergence of order $r - i + 1$ with respect to time from the i th derivative of the error $u - u_{\tau h}$. But Proposition 4.44 only gives the order $r - \ell + 1$, which is suboptimal for $0 \leq i < \ell$. \clubsuit

4.2.2 What about superconvergence!?

For $0 \leq i < \ell$ the estimates of Proposition 4.44 do not show the convergence orders that we would expect from the i th derivative of the error $u - u_{\tau h}$. Hence, we are not satisfied by these estimates.

An obvious approach therefore would be to use superconvergence results in the time (mesh) points to derive more appropriate convergence orders. Exploiting the supercloseness result of Subsection 4.1.6, we can derive a low order superconvergence result for the ℓ th derivative of the error in the time (mesh) points at least for $g \in \{\mathcal{I}_k^r f, \mathcal{C}_k^r f\}$. We, thus, gain the following result.

Proposition 4.46

Let $1 \leq n \leq N$ and $0 \leq i \leq \ell$. Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{\mathcal{I}_k^r f, \mathcal{C}_k^r f\}$. Then, we have the following error estimate

$$\begin{aligned} &\|(u - u_{\tau h})^{(i)}(t_n^-)\|^2 \\ &\leq C(1 + \delta_{1,k-2\ell}(t_n - t_0)) \exp(t_{n-1} - t_0) \\ &\quad \left[h^{2(\kappa+\sigma)} \left((t_n - t_0) \|u^{(\ell+1)}\|_{C((t_0, t_n), H^{\kappa+1}(\Omega))}^2 + \|u^{(\ell)}(t_0)\|_{H^{\kappa+1}(\Omega)}^2 + \sum_{j=i}^{\ell} \|u^{(j)}(t_n)\|_{H^{\kappa+1}(\Omega)}^2 \right) \right. \\ &\quad + \tau^{2(r-\ell+2)} \left(\|u^{(\ell)}\|_{H^{r-\ell+2}((t_0, t_n), H^1(\Omega))}^2 + \|f\|_{H^{r+2}((t_0, t_n), H^{-1}(\Omega))}^2 \right) \\ &\quad \left. + (1 + \tau) \tau^{2 \min\{r-\ell+2, 2r-k+1\}} \|u^{(\ell+1)}\|_{H^{\min\{r-\ell+2, 2r-k+1\}}((t_0, t_n), H^1(\Omega))}^2 \right], \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Recalling the arguments in the proof of Lemma 4.43, we find for $0 \leq i \leq \ell$

$$\begin{aligned} \|(u - u_{\tau h})^{(i)}(t_n^-)\| &\leq C \left(\|(u - R_h u)^{(\ell)}(t_n^-)\| + \|\omega_n^{\mathcal{J}}(R_h u^{(\ell)})\| + \|e_{\tau h, \ell}^{\mathcal{J}}(t_n^-)\| \right) \\ &\quad + C \sum_{j=i}^{\ell-1} \left(\|(u - R_h u)^{(j)}(t_n^-)\| + \|(f - g)^{(j)}(t_n^-)\|_{V'} \right), \end{aligned}$$

where we choose $\mathcal{J}_\nu = Q_{k-2\ell, \nu}^{r-\ell}$, which is possibly due to $g \in \{\mathcal{I}_k^r f, \mathcal{C}_k^r f\}$. Moreover, note that there is an approximation operator \mathcal{A} satisfying (4.16) because of Remark 4.38.

The desired statement then follows from (3.16), Lemma 4.36, and Theorem 4.40. Here, also note that $(f - g)^{(j)}(t_n^-) = 0$ for $0 \leq j \leq \ell$ since $g \in \{\mathcal{I}_k^r f, \mathcal{C}_k^r f\}$. \square

In view of Subsection 1.2.3, we could hope for superconvergence results of the high order $2r - k + 1$ in the time (mesh) points also in the case of parabolic problems. However, known results from the literature, see e.g. [11, Theorem 4.2], [52, p. 211], or also [6, Section 6], suggest that, in order to obtain such higher order superconvergence for dG or cGP methods, certain compatibility conditions are needed. So, inconvenient assumptions on the data would be required and some artificial boundary conditions would have to be imposed that often are quite unrealistic. We therefore look for an alternative approach.

4.2.3 Satisfactory order convergence avoiding superconvergence

As we have seen in the previous subsection, superconvergence estimates are only suitable to a limited extent to derive convergence of satisfactory order. Therefore, we need to find a technique of proof avoiding superconvergence.

In view of Subsection 3.3.2, the fully discrete problem can be interpreted as approximation to the semi-discrete problem (3.11). So, we may transfer the ideas of the (stiff) error analysis of Section 2.3, especially the results of Theorem 2.30. However, to derive appropriate estimates, we then need uniform bounds on the derivatives of the solution u_h to the semi-discrete problem.

Semi-discretization in space revisited

Recalling the stability estimate of Corollary 3.12, it remains to derive uniform bounds for $\|u_h^{(i)}(t_0)\|$ or $\|\bar{U}_h^{(i)}(t_0)\|$, respectively. This, however, is only possible if the initial value $u_{h,0}$ for the semi-discretization in space is suitably chosen.

To this end, let f and, thus, \tilde{F} be at least $(r + 1)$ -times continuously differentiable with respect to time on \bar{I} . Then, by (3.15) we have for $i = 0, \dots, r + 1$ the following iterative connection

$$(u_h^{(i+1)}(t_0^+), v_h) = \langle f^{(i)}(t_0^+), v_h \rangle_{V', V} - a(u_h^{(i)}(t_0^+), v_h) \quad \forall v_h \in V_h$$

or in basis representation

$$\bar{U}_h^{(i+1)}(t_0^+) = \bar{M} M^{-1} \tilde{F}^{(i)}(t_0^+) - \bar{A} \bar{U}_h^{(i)}(t_0^+),$$

where $\bar{U}_h^{(i)}(t_0^+) = \bar{M}U_h^{(i)}(t_0^+) = M^{1/2}U_h^{(i)}(t_0^+)$ and $u_h^{(i)}(t_0^+) = \sum_{j=1}^{\dim(V_h)} (U_h^{(i)}(t_0^+))_j \varphi_j$. Adapting the idea used in (4.1) to define the initial value of the discrete problem, we choose

$$\begin{aligned} u_{h,0}^{[r+2]} &:= P_h \partial_t^{r+2} u_0, \\ u_{h,0}^{[i]} \in V_h, i = r+1, \dots, 0 : \quad a(u_{h,0}^{[i]}, v_h) &= \langle f^{(i)}(t_0^+), v_h \rangle_{V', V} - (u_{h,0}^{[i+1]}, v_h) \quad \forall v_h \in V_h, \end{aligned} \quad (4.20)$$

where $\partial_t^{r+2} u_0$ is generated from u_0 via (3.7). We then set $u_{h,0} := u_{h,0}^{[0]}$. Obviously, by construction it holds that $u_h^{(i)}(t_0^+) = u_{h,0}^{[i]}$ for $i = 0, \dots, r+2$.

Note that a similar approach for the choice of the initial values can be found in [52, pp. 74–75]. There these initial values were needed to guarantee uniform estimates in negative norms for the derivative of the error of the semi-discrete problem down to the initial time $t = t_0$.

Lemma 4.47

Let $f \in C^{r+2}(\bar{I}, V')$ and suppose that $u_{h,0} := u_{h,0}^{[0]}$ with $u_{h,0}^{[i]}$ according to (4.20). Then, for $i = 0, \dots, r+2$ it holds

$$\|u_h^{(i)}(t_0^+)\| = \|\bar{U}_h^{(i)}(t_0^+)\| \leq C \left(\|\partial_t^{r+2} u_0\| + \sum_{j=i}^{r+1} \|f^{(j)}(t_0^+)\|_{V'_h} \right).$$

Proof. The argument is quite similar to that used at the beginning of Subsection 4.2.1. Indeed, using the V -ellipticity of $a(\cdot, \cdot)$, the definition of the norm in V'_h , the Cauchy–Schwarz inequality, and $V \hookrightarrow H$, it follows for $i = 0, \dots, r+1$

$$\begin{aligned} \alpha \|u_{h,0}^{[i]}\|_V^2 &\leq a(u_{h,0}^{[i]}, u_{h,0}^{[i]}) = \langle f^{(i)}(t_0^+), u_{h,0}^{[i]} \rangle_{V', V} - (u_{h,0}^{[i+1]}, u_{h,0}^{[i]}) \\ &\leq \left(\|f^{(i)}(t_0^+)\|_{V'_h} + C_{\text{emb}} \|u_{h,0}^{[i+1]}\| \right) \|u_{h,0}^{[i]}\|_V. \end{aligned}$$

Therefore, we have

$$\|u_{h,0}^{[i]}\| \leq C_{\text{emb}} \|u_{h,0}^{[i]}\|_V \leq \frac{C_{\text{emb}}}{\alpha} \left(\|f^{(i)}(t_0^+)\|_{V'_h} + C_{\text{emb}} \|u_{h,0}^{[i+1]}\| \right).$$

So, by recursion we conclude

$$\|u_{h,0}^{[i]}\| \leq \left(\frac{C_{\text{emb}}^2}{\alpha} \right)^{r+2-i} \|u_{h,0}^{[r+2]}\| + \frac{C_{\text{emb}}}{\alpha} \sum_{j=i}^{r+1} \left(\frac{C_{\text{emb}}^2}{\alpha} \right)^{j-i} \|f^{(j)}(t_0^+)\|_{V'_h}$$

for $i = 0, \dots, r+2$. Because of $u_{h,0}^{[r+2]} = P_h \partial_t^{r+2} u_0$, the desired statement follows easily using the stability of P_h in $\|\cdot\|$ and the fact that $u_h^{(i)}(t_0^+) = u_{h,0}^{[i]}$. \square

We see that the special choice of $u_{h,0}$ guarantees that the norm of $\|u_h^{(i)}(t_0)\| = \|\bar{U}_h^{(i)}(t_0)\|$ with $i = 0, \dots, r+2$ can be estimated with respect to the given data and independent of h . Therefore, together with Corollary 3.12, we obtain the following stability result.

Corollary 4.48

Let $f \in C^{r+2}(\bar{I}, V')$ and let the initial value of the semi-discrete problem be chosen as $u_{h,0} := u_{h,0}^{[0]}$ with $u_{h,0}^{[i]}$ according to (4.20). Then, for $i = 0, \dots, r+2$ it holds

$$\sup_{t \in \bar{I}} \|u_h^{(i)}(t)\| = \sup_{t \in \bar{I}} \|\bar{U}_h^{(i)}(t)\| \leq C \left(\|\partial_t^{r+2} u_0\| + \|f^{(i)}\|_{L^2(\bar{I}, V_h')} + \sum_{j=i}^{r+1} \|f^{(j)}(t_0^+)\|_{V_h'} \right).$$

Remark 4.49

For $0 \leq i \leq r+1$ the estimates of Lemma 4.47 and Corollary 4.48 stay true if $\|\partial_t^{r+2} u_0\|$ is replaced by $\|\partial_t^{r+2} u_0\|_{V_h'}$. Moreover, we only need $f \in C^{r+1}(\bar{I}, V')$ then. ♣

Having the initial values properly defined, we can concretize the error estimates for the semi-discrete approximation. The convergence rates obtained for the model problem are given in the next proposition.

Proposition 4.50

Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16). Moreover, suppose that f be $(r+1)$ -times continuously differentiable with respect to time on \bar{I} and let the initial value of the semi-discrete problem be chosen as $u_{h,0} := u_{h,0}^{[0]}$ with $u_{h,0}^{[i]}$ according to (4.20). Then, for $i = 0, \dots, r+1$ and $t > t_0$ we have the following error estimate

$$\|(u - u_h)^{(i)}(t)\| \leq Ch^{\kappa+\sigma} \left(\|u^{(i)}\|_{H^1((t_0,t), H^{\kappa+1}(\Omega))} + \sum_{j=i}^r \|u^{(j+1)}(t_0)\|_{H^{\kappa+1}(\Omega)} \right),$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Recalling the estimates of Theorem 3.13, we already have

$$\begin{aligned} & \|(u - u_h)^{(i)}(t)\| \\ & \leq \|(u^{(i)} - R_h u^{(i)})(t)\| + \|(R_h u^{(i)} - u_h^{(i)})(t_0^+)\| + C \left(\int_{t_0}^t \|u^{(i+1)} - R_h u^{(i+1)}\|_{V'}^2 ds \right)^{1/2}. \end{aligned}$$

Approximation results for R_h are known, cf. (3.16). So, bounds on $\|(R_h u^{(i)} - u_h^{(i)})(t_0^+)\|$ are needed only. For this we can adapt the argumentation of the proof of Lemma 4.47 and obtain for $v_{h,0}^{[i]} := (R_h u^{(i)} - u_h^{(i)})(t_0^+) \in V_h$ that

$$\begin{aligned} \alpha \|v_{h,0}^{[i]}\|_V^2 & \leq a((R_h u^{(i)} - u_h^{(i)})(t_0^+), v_{h,0}^{[i]}) = a(u^{(i)}(t_0^+), v_{h,0}^{[i]}) - a(u_h^{(i)}(t_0^+), v_{h,0}^{[i]}) \\ & = \langle f^{(i)}(t_0^+) - \partial_t u^{(i)}(t_0^+), v_{h,0}^{[i]} \rangle_{V', V} - \langle f^{(i)}(t_0^+) - \partial_t u_h^{(i)}(t_0^+), v_{h,0}^{[i]} \rangle_{V', V} \\ & = -\langle (u^{(i+1)} - u_h^{(i+1)})(t_0^+), v_{h,0}^{[i]} \rangle_{V', V} \leq \|(u^{(i+1)} - u_h^{(i+1)})(t_0^+)\|_{V_h'} \|v_{h,0}^{[i]}\|_V \\ & \leq \left(\|(u^{(i+1)} - R_h u^{(i+1)})(t_0^+)\|_{V_h'} + C_{\text{emb}} \|(R_h u^{(i+1)} - u_h^{(i+1)})(t_0^+)\| \right) \|v_{h,0}^{[i]}\|_V. \end{aligned}$$

Thus, it follows

$$\begin{aligned} \|(R_h u^{(i)} - u_h^{(i)})(t_0^+)\| &\leq C_{\text{emb}} \|(R_h u^{(i)} - u_h^{(i)})(t_0^+)\|_V \leq \frac{C_{\text{emb}}}{\alpha} \|(u^{(i+1)} - u_h^{(i+1)})(t_0^+)\|_{V'_h} \\ &\leq \frac{C_{\text{emb}}}{\alpha} \left(\|(u^{(i+1)} - R_h u^{(i+1)})(t_0^+)\|_{V'_h} + C_{\text{emb}} \|(R_h u^{(i+1)} - u_h^{(i+1)})(t_0^+)\| \right) \end{aligned}$$

and consequently

$$\begin{aligned} &\|(R_h u^{(i)} - u_h^{(i)})(t_0^+)\| \\ &\leq \left(\frac{C_{\text{emb}}^2}{\alpha} \right)^{r+1-i} \|(R_h u^{(r+1)} - u_h^{(r+1)})(t_0^+)\| + \frac{C_{\text{emb}}}{\alpha} \sum_{j=i}^r \left(\frac{C_{\text{emb}}^2}{\alpha} \right)^{j-i} \|(u^{(j+1)} - R_h u^{(j+1)})(t_0^+)\|_{V'_h} \\ &\leq \frac{C_{\text{emb}}}{\alpha} \left(\left(\frac{C_{\text{emb}}^2}{\alpha} \right)^{r+1-i} \|(\text{Id} - P_h) \partial_t^{r+2} u_0\|_{V'_h} + \sum_{j=i}^r \left(\frac{C_{\text{emb}}^2}{\alpha} \right)^{j-i} \|(u^{(j+1)} - R_h u^{(j+1)})(t_0^+)\|_{V'_h} \right) \end{aligned}$$

for $i = 0, \dots, r+1$. Because of $\|(\text{Id} - P_h) \partial_t^{r+2} u_0\|_{V'_h} = 0$, we overall conclude from (3.16) that

$$\|(u - u_h)^{(i)}(t)\| \leq Ch^{\kappa+\sigma} \left(\|u^{(i)}\|_{H^1((t_0, t), H^{\kappa+1}(\Omega))} + \sum_{j=i}^r \|u^{(j+1)}(t_0)\|_{H^{\kappa+1}(\Omega)} \right),$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise. So, we are done. \square

Transferring the (stiff) error analysis

We now are ready to give estimates in the time (mesh) points of a satisfactory order.

Theorem 4.51

Let $1 \leq n \leq N$ and $0 \leq i \leq \ell = \lfloor \frac{k}{2} \rfloor$. Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{\Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Moreover, suppose that f is $(r+2)$ -times continuously differentiable with respect to time on \bar{I} . Then, we have the following error estimate

$$\begin{aligned} &\|(u - u_{\tau h})^{(i)}(t_n^-)\| + \|(R_h u - u_{\tau h})^{(i)}(t_n^-)\| \\ &\leq Ch^{\kappa+\sigma} \left(\|u^{(i)}\|_{H^1((t_0, t_n), H^{\kappa+1}(\Omega))} + \sum_{j=i}^r \|u^{(j+1)}(t_0)\|_{H^{\kappa+1}(\Omega)} \right) \\ &\quad + C(t_n - t_0) \tau^{r+1-i} \left(\|\partial_t^{r+2} u_0\| + \|f^{(r+2)}\|_{L^2((t_0, t_n), V'_h)} + \sup_{t \in (t_0, t_n)} \|f^{(r+1)}(t)\| \right), \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. Unlike in the proof of Lemma 4.43, we here use a splitting of the error of the form

$$(u - u_{\tau h})^{(i)}(t_n^-) = (u - u_h)^{(i)}(t_n^-) + (u_h - u_{\tau h})^{(i)}(t_n^-) \quad (4.21)$$

where u_h is the solution of the semi-discrete problem (3.8) to the initial value $u_{h,0} = u_{h,0}^{[0]}$ with $u_{h,0}^{[i]}$ defined according to (4.20).

The first term on the right-hand side of (4.21) can be estimated by Proposition 4.50. For the second term we have due to Remark 3.7 that

$$\|(u_h - u_{\tau h})^{(i)}(t_n^-)\| = \|(\bar{U}_h - \bar{U}_{\tau h})^{(i)}(t_n^-)\|.$$

So, we are almost in the same setting as in Section 2.3. We let \bar{U}_h take the role of \bar{u} in Chapter 2 where \tilde{F} replaces f . Similarly, $\bar{U}_{\tau h}$ takes the part of \bar{U} where \tilde{G} replaces g , but the initial value is chosen slightly different. Of course, in contrast to Chapter 2, we do not have $\bar{U}_{\tau h}(t_0) = \bar{U}_h(t_0)$ in general.

Now, revising the arguments of Section 2.3, we see that the concrete choice of the initial value is only needed and used at the end of Theorem 2.30. Therefore, we gain (also noting that $\tilde{C} = 0$ since $\mu = -\alpha C_{\text{emb}}^{-2} < 0$) that

$$\begin{aligned} \|(u_h - u_{\tau h})^{(i)}(t_n^-)\| &= \|(\bar{U}_h - \bar{U}_{\tau h})^{(i)}(t_n^-)\| \\ &\leq \|(\bar{U}_h - \bar{U}_{\tau h})^{(i)}(t_0^-)\| + \sum_{\nu=1}^n C \left(\frac{\tau_\nu}{2}\right)^{r+2-i} \left(\sup_{t \in I_\nu} \|\bar{M} M^{-1} \tilde{F}^{(r+1)}(t)\| + \sup_{t \in I_\nu} \|\bar{U}_h^{(r+2)}(t)\| \right) \\ &\leq \|(u_h - u_{\tau h})^{(i)}(t_0^-)\| + \sum_{\nu=1}^n C \left(\frac{\tau_\nu}{2}\right)^{r+2-i} \left(\sup_{t \in I_\nu} \|f^{(r+1)}(t)\| + \sup_{t \in I_\nu} \|u_h^{(r+2)}(t)\| \right) \end{aligned}$$

for $0 \leq i \leq \lfloor \frac{k}{2} \rfloor$ and $1 \leq n \leq N$. Because of Corollary 4.48, the latter term is uniformly bounded. But we need to study the initial value term. For this, we use the splitting

$$\|(u_h - u_{\tau h})^{(i)}(t_0^-)\| \leq \|(u_h^{(i)} - R_h u^{(i)})(t_0^-)\| + \|(R_h u^{(i)} - u_{\tau h}^{(i)})(t_0^-)\|.$$

Estimates for the first term on the right-hand side have already been derived in the proof of Proposition 4.50. Since the initial values of the fully discrete problem and the semi-discrete problem are defined very similar, cf. (4.1) and (4.20), the second term can be estimated quite analog. Here, also note that due to $g \in \{\Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup \text{if } k \geq 2 \{ \mathcal{I}_{k-2,*}^r f \}$ we have $g^{(i)}(t_0^+) = f^{(i)}(t_0^+)$, $i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1$. Thus, in (4.1) we could use f instead of g . Therefore, we obtain

$$\begin{aligned} \|(u_h - u_{\tau h})^{(i)}(t_0^-)\| &\leq C \|(\text{Id} - P_h) \partial_t^{r+2} u_0\|_{V_h'} + C \|(R_h - \tilde{P}_h^0) \partial_t^{\lfloor \frac{k}{2} \rfloor} u_0\| + C \sum_{j=i}^r \|(u^{(j+1)} - R_h u^{(j+1)})(t_0)\|_{V_h'}. \end{aligned}$$

Noting that $\|(\text{Id} - P_h) \partial_t^{r+2} u_0\|_{V_h'} = 0$ as well as $\|(R_h - \tilde{P}_h^0) \partial_t^{\lfloor \frac{k}{2} \rfloor} u_0\| = 0$ if $\tilde{P}_h^0 = R_h$ and $\|(R_h - \tilde{P}_h^0) \partial_t^{\lfloor \frac{k}{2} \rfloor} u_0\| \leq \|(R_h - \text{Id}) \partial_t^{\lfloor \frac{k}{2} \rfloor} u_0\|$ if $\tilde{P}_h^0 = P_h$, we thus get from (3.16) that

$$\|(u_h - u_{\tau h})^{(i)}(t_0^-)\| \leq C h^{\kappa+\sigma} \sum_{j=\min\{i+1, \lfloor \frac{k}{2} \rfloor\}}^{r+1} \|u^{(j)}(t_0)\|_{H^{\kappa+1}(\Omega)}$$

for $0 \leq i \leq \lfloor \frac{k}{2} \rfloor$.

Combining the above observations, we easily conclude the desired statement. \square

Because of Theorem 4.51, which for $0 \leq i \leq \ell$ shows the expected temporal convergence order $r - i + 1$ for the i th derivative of the error $u - u_{\tau h}$ at least in the time (mesh) points, we do not need superconvergence estimates for error estimates of a satisfactory order anymore.

4.3 Final error estimate

We now have estimates for the ℓ th derivative of the error as well as for the error in the time (mesh) points. For discrete functions (in $P_r(I_n, V_h)$) these information suffice to bound the $L^2(H)$ -norm. Therefore, we split the error in a discrete error part and a remaining projection error part where a suitable projection operator has to be used.

The further error analysis is based on a norm equivalence in the finite dimensional space $P_r(I_n, V_h)$ where V_h is equipped with the norm $\|\cdot\|$. The following statement is proven in Appendix D, see Lemma D.4, in a more general setting.

Lemma 4.52

Let $0 \leq l \leq r$ and let $\|\cdot\|_W$ denote some norm on V_h . Then, the mappings

$$v \mapsto \left(\int_{I_n} \|v(t)\|_W^2 dt \right)^{1/2} \quad \text{and} \quad v \mapsto \left(\left(\frac{\tau_n}{2} \right)^{2l} \int_{I_n} \|v^{(l)}(t)\|_W^2 dt + \sum_{i=0}^{l-1} \left(\frac{\tau_n}{2} \right)^{2i+1} \|v^{(i)}(t_n^-)\|_W^2 \right)^{1/2}$$

define equivalent norms on $P_r(I_n, V_h)$ where the equivalence constants are independent of τ_n and of V_h .

As before, let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $\ell = \lfloor \frac{k}{2} \rfloor$. We generalize the projection operator $\tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}$ introduced in Subsection 4.1.1. So, let X denote some Banach space over \mathbb{R} . For $v \in H^{\ell+1}(I_n, X) \cap C^{k_{\mathcal{J}}+\ell+1}(\bar{I}_n, X)$ let $\bar{\Pi}_k^{r, \mathcal{J}} v \in P_r(I_n, X)$ be determined by

$$\begin{aligned} (\bar{\Pi}_k^{r, \mathcal{J}} v)^{(j)}(t_n^-) &= v^{(j)}(t_n^-), & \text{for } j = 0, \dots, \ell - 1, \\ (\bar{\Pi}_k^{r, \mathcal{J}} v)^{(\ell)}(t) &= \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}}(v^{(\ell)})(t), & \text{for all } t \in I_n, \end{aligned}$$

cf. Definition C.12. Note that the second condition is just an identity of two polynomials of (maximal) degree $r - \ell$.

In order to derive error estimates for $\bar{\Pi}_k^{r, \mathcal{J}}$, we first of all note that for $\ell \geq 1$ we have

$$\begin{aligned} &\|(v - \bar{\Pi}_k^{r, \mathcal{J}} v)(t)\|_X \\ &= \left\| \underbrace{(v - \bar{\Pi}_k^{r, \mathcal{J}} v)(t_n^-)}_{=0} - \int_t^{t_n} (v - \bar{\Pi}_k^{r, \mathcal{J}} v)'(s) ds \right\|_X \leq \int_{I_n} \|(v - \bar{\Pi}_k^{r, \mathcal{J}} v)'(s)\|_X ds \quad \forall t \in I_n, \end{aligned}$$

where we used the fundamental theorem of calculus, the definition of $\bar{\Pi}_k^{r, \mathcal{J}}$, and properties of the Bochner integral. By iteration we then obtain for $t \in I_n$ that

$$\|(v - \bar{\Pi}_k^{r, \mathcal{J}} v)(t)\|_X \leq \tau_n^{\ell-1} \int_{I_n} \|(v - \bar{\Pi}_k^{r, \mathcal{J}} v)^{(\ell)}(s)\|_X ds = \tau_n^{\ell-1} \int_{I_n} \|(v^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell, \mathcal{J}} v^{(\ell)})(s)\|_X ds.$$

Hence, it follows

$$\int_{I_n} \|(v - \bar{\Pi}_k^{r,\mathcal{J}} v)(t)\|_X^2 dt \leq \tau_n^{2\ell} \int_{I_n} \|(v^{(\ell)} - \tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}} v^{(\ell)})(s)\|_X^2 ds, \quad (4.22)$$

where also the Cauchy–Schwarz inequality was used. Bounds on the approximation error of $\tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}}$, which is on the right-hand side, are already known from (4.9).

Now, we are well prepared to start the proof of the error estimate for general variational time discretization methods.

Theorem 4.53

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $\ell = \lfloor \frac{k}{2} \rfloor$. Moreover, suppose that f is $(r+2)$ -times continuously differentiable with respect to time on \bar{I} . Consider the setting of model problem (3.4) with standard spatial discretization satisfying (3.16) and let $g \in \{\Pi_k^r f, \mathcal{I}_k^r f, \mathcal{C}_k^r f\} \cup_{\text{if } k \geq 2} \{\mathcal{I}_{k-2,*}^r f\}$. Then, we have the following error estimate

$$\begin{aligned} & \int_{t_0}^{t_n} \|u - u_{\tau h}\|^2 dt \\ & \leq C(1 + (t_n - t_0))h^{2(\kappa+\sigma)} \left(\|u\|_{H^{\ell+1}((t_0,t_n),H^{\kappa+1}(\Omega))}^2 + \sum_{j=0}^r \|u^{(j+1)}(t_0)\|_{H^{\kappa+1}(\Omega)}^2 \right) \\ & \quad + C(1 + (t_n - t_0)^3)\tau^{2(r+1)} \\ & \quad \left(\|u\|_{H^{r+1}((t_0,t_n),H^1(\Omega))}^2 + \|\partial_t^{r+2} u_0\|^2 + \|f\|_{H^{r+2}((t_0,t_n),H^{-1}(\Omega))}^2 + \sup_{t \in (t_0,t_n)} \|f^{(r+1)}(t)\|^2 \right), \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

Proof. For the proof we choose $\mathcal{J}_\nu = \int_{I_\nu}$ and decompose the error $u - u_{\tau h}$ as follows

$$u - u_{\tau h} = (u - R_h u) + (R_h u - R_h \bar{\Pi}_k^{r,\mathcal{J}} u) + (R_h \bar{\Pi}_k^{r,\mathcal{J}} u - u_{\tau h}).$$

Because of the stability of R_h in $\|\cdot\|_V$, the second summand can be bounded as

$$\int_{t_0}^{t_n} \|R_h u - R_h \bar{\Pi}_k^{r,\mathcal{J}} u\|^2 dt \leq \int_{t_0}^{t_n} C_{\text{emb}}^2 \|R_h(u - \bar{\Pi}_k^{r,\mathcal{J}} u)\|_V^2 dt \leq C \int_{t_0}^{t_n} \|u - \bar{\Pi}_k^{r,\mathcal{J}} u\|_V^2 dt.$$

Thus, the projection error parts can be estimated using the known error bounds for the projection operators R_h , see (3.16), and $\bar{\Pi}_k^{r,\mathcal{J}}$, see (4.9) and (4.22), by

$$\begin{aligned} & \int_{t_0}^{t_n} \|u - R_h u\|^2 dt \leq C h^{2(\kappa+\sigma)} \|u\|_{L^2((t_0,t_n),H^{\kappa+1}(\Omega))}^2, \\ & \int_{t_0}^{t_n} \|R_h u - R_h \bar{\Pi}_k^{r,\mathcal{J}} u\|^2 dt \leq C \tau^{2(r+1)} \|u\|_{H^{r+1}((t_0,t_n),H^1(\Omega))}^2, \end{aligned}$$

where $\sigma = 1$ if the associated stationary problem is H^2 -regular and $\sigma = 0$ otherwise.

It remains to study the fully discrete error part $(R_h \bar{\Pi}_k^{r,\mathcal{J}} u - u_{\tau h}) \in P_r(I_n, V_h)$. The norm equivalence of Lemma 4.52 with $l = \ell$ gives

$$\begin{aligned} & \int_{I_\nu} \|R_h \bar{\Pi}_k^{r,\mathcal{J}} u - u_{\tau h}\|^2 dt \\ & \leq C \left(\left(\frac{\tau_\nu}{2}\right)^{2\ell} \int_{I_\nu} \|(R_h \bar{\Pi}_k^{r,\mathcal{J}} u - u_{\tau h})^{(\ell)}\|^2 dt + \sum_{i=0}^{\ell-1} \left(\frac{\tau_\nu}{2}\right)^{2i+1} \|(R_h \bar{\Pi}_k^{r,\mathcal{J}} u - u_{\tau h})^{(i)}(t_\nu^-)\|^2 \right) \\ & = C \left(\left(\frac{\tau_\nu}{2}\right)^{2\ell} \int_{I_\nu} \|R_h \tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)}\|^2 dt + \sum_{i=0}^{\ell-1} \left(\frac{\tau_\nu}{2}\right)^{2i+1} \|(R_h u - u_{\tau h})^{(i)}(t_\nu^-)\|^2 \right), \end{aligned}$$

where for the second step the definition of $\bar{\Pi}_k^{r,\mathcal{J}}$ was used. Recalling the notation of the two previous sections, especially

$$e_{\tau h, \ell}^{\mathcal{J}} = R_h \tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}} u^{(\ell)} - u_{\tau h}^{(\ell)},$$

a summation over $\nu = 1, \dots, n$ yields

$$\begin{aligned} \int_{t_0}^{t_n} \|R_h \bar{\Pi}_k^{r,\mathcal{J}} u - u_{\tau h}\|^2 dt & \leq C \tau^{2\ell} \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt + C \sum_{\nu=1}^n \sum_{i=0}^{\ell-1} \left(\frac{\tau_\nu}{2}\right)^{2i+1} \|(R_h u - u_{\tau h})^{(i)}(t_\nu^-)\|^2 \\ & \leq C \tau^{2\ell} \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt + C (t_n - t_0) \sum_{i=0}^{\ell-1} \tau^{2i} \max_{\nu=1, \dots, n} \|(R_h u - u_{\tau h})^{(i)}(t_\nu^-)\|^2. \end{aligned}$$

All apparent terms can be estimated by Theorems 4.11 and 4.51. In detail, we have

$$\begin{aligned} \int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|^2 dt & \leq C(1 + (t_n - t_0)) \left[h^{2(\kappa+\sigma)} \|u\|_{H^{\ell+1}((t_0, t_n), H^{\kappa+1}(\Omega))}^2 \right. \\ & \quad \left. + \tau^{2(r-\ell+1)} \left(\|u\|_{H^{r+1}((t_0, t_n), H^1(\Omega))}^2 + \|f\|_{H^{r+1}((t_0, t_n), H^{-1}(\Omega))}^2 \right) \right] \end{aligned}$$

and for $0 \leq i \leq \ell - 1$

$$\begin{aligned} & \max_{\nu=1, \dots, n} \|(R_h u - u_{\tau h})^{(i)}(t_\nu^-)\| \\ & \leq C h^{\kappa+\sigma} \left(\|u\|_{H^\ell((t_0, t_n), H^{\kappa+1}(\Omega))} + \sum_{j=0}^r \|u^{(j+1)}(t_0)\|_{H^{\kappa+1}(\Omega)} \right) \\ & \quad + C(t_n - t_0) \tau^{r+1-i} \left(\|\partial_t^{r+2} u_0\| + \|f^{(r+2)}\|_{L^2((t_0, t_n), V_h')} + \sup_{t \in (t_0, t_n)} \|f^{(r+1)}(t)\| \right) \end{aligned}$$

with σ as above. Hence, it follows

$$\begin{aligned} & \int_{t_0}^{t_n} \|R_h \bar{\Pi}_k^{r,\mathcal{J}} u - u_{\tau h}\|^2 dt \\ & \leq C(1 + (t_n - t_0)) h^{2(\kappa+\sigma)} \left(\|u\|_{H^{\ell+1}((t_0, t_n), H^{\kappa+1}(\Omega))}^2 + \sum_{j=0}^r \|u^{(j+1)}(t_0)\|_{H^{\kappa+1}(\Omega)}^2 \right) \\ & \quad + C(1 + (t_n - t_0)^3) \tau^{2(r+1)} \\ & \quad \left(\|u\|_{H^{r+1}((t_0, t_n), H^1(\Omega))}^2 + \|\partial_t^{r+2} u_0\|^2 + \|f\|_{H^{r+2}((t_0, t_n), H^{-1}(\Omega))}^2 + \sup_{t \in (t_0, t_n)} \|f^{(r+1)}(t)\|^2 \right). \end{aligned}$$

Summarizing, we get the desired statement and the proof is completed. \square

Remark 4.54

The estimate of Theorem 4.53 is of optimal order with respect to space and time. In the case that $k \in \{0, 1\}$ a similar estimate was already proven in Theorem 4.11. Otherwise, for $k \geq 2$, in the proof of Theorem 4.53 the results of Theorem 4.51 and, thus, the (stiff) error analysis were reused to gain estimates for $\|(R_h u - u_{\tau h})^{(i)}(t_\nu^-)\|^2$ of sufficiently high order. However, note that for $k \in \{2, 3\}$ and $g \in \{\mathcal{I}_k^r f, \mathcal{C}_k^r f\}$ we could alternatively use Proposition 4.46, which was derived from the supercloseness result of Subsection 4.1.6, to get final error estimates of optimal order.

Still another way to prove optimal order $L^2(H)$ -estimates for $\mathbf{VTD}_k^r(g)$ with $k \in \{2, 3\}$ and specially chosen g is presented in [9, 12] in the context of the parabolic wave equations and in [27] for linear first order partial differential equations. The approach strongly exploits the connection between \mathbf{VTD}_k^r and postprocessed \mathbf{VTD}_{k-2}^{r-1} methods. Moreover, for the argument it is quite crucial that the difference between the dimensions of trial space and test space in the variational condition and, thus, k is not too large. Therefore, this approach seems to be limited to small k . \clubsuit

Remark 4.55 (Comments on the choice of the discrete initial condition)

By (4.1) the discrete initial condition is determined in a very special way. First the initial value of the ℓ th derivative is projected by a spatial approximation operator and then the discrete initial values for lower derivatives are determined via the differential equation. This special choice is exploited at several points in the analysis.

On the one hand, it guarantees that the collocation conditions also hold at t_0^- . This is used in the proof of Lemma 4.1 and in the argumentation of Section 4.2, especially in Theorem 4.51. On the other hand, directly projecting the initial value of the ℓ th derivative makes the estimation of the initial error in Section 4.1 quite straightforward. Furthermore, estimates for the initial error of higher derivatives in the H -norm give respective estimates for lower derivatives in the V -norm, which can be easily seen by adapting the argumentation of Subsection 4.2.1. Suitable error estimates for the initial error of the ℓ th derivative therefore also yield appropriate estimates for all lower derivatives.

In contrast, the latter argumentation does not work in the other direction since control on the stronger V -norm of the initial error of the i th derivative is necessary to bound the H -norm of the initial error of the $(i + 1)$ th derivative. Actually, this is also apparent in numerical experiments, cf. Tables 4.1 and 4.2. Especially for large k , considerably larger errors and reduced convergence orders are observed if, instead of by (4.1), the initial values $\partial_t^i u_{\tau h}(t_0^-) \in V_h$, $i = 0, \dots, \lfloor \frac{k}{2} \rfloor$, are determined by

$$\begin{aligned} \partial_t^0 u_{\tau h}(t_0^-) &= P_h u_0, \\ \partial_t^i u_{\tau h}(t_0^-) &\in V_h \text{ with } i = 1, \dots, \lfloor \frac{k}{2} \rfloor : \\ (\partial_t^i u_{\tau h}(t_0^-), v_h) &= \langle g^{(i-1)}(t_0^+), v_h \rangle_{V', V} - a(\partial_t^{i-1} u_{\tau h}(t_0^-), v_h) \quad \forall v_h \in V_h. \end{aligned} \tag{4.23}$$

With this in mind, further research to allow a more flexible or easily implemented choice of initial conditions would be appropriate. \clubsuit

Remark 4.56 (Comments on postprocessing)

Note that, due to the findings of Subsection 3.3.2, the postprocessing according to Theorem 1.32 can be applied also in the considered parabolic setting. Especially, for the solution $u_{\tau h}$ of $\mathbf{VTD}_k^r(\mathcal{I}_k^r f)$ or $\mathbf{VTD}_k^r(\mathcal{C}_k^r f)$, $0 \leq k < r$, postprocessing yields the solution $\tilde{u}_{\tau h}$ of $\mathbf{VTD}_{k+2}^{r+1}(\mathcal{I}_{k,*}^{r+1} f)$ or $\mathbf{VTD}_{k+2}^{r+1}(\mathcal{C}_{k+2}^{r+1} f)$, respectively. In order to estimate the $L^2(H)$ -norm of the error $u - \tilde{u}_{\tau h}$, then Theorem 4.53 can be used. However, a careful inspection of the influences of the discrete initial condition may be required. ♣

Remark 4.57 (Comments on estimates in the $L^2(V)$ -norm)

Adapting the arguments in the proof of Theorem 4.53, also estimates in the V -norm can be proven. Of course, we then build on the results of Subsection 4.1.3 to bound $\int_{t_0}^{t_n} \|e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt$. Moreover, note that by (4.18) in the time (mesh) points error control in the H -norm implies certain error control in the V -norm. Therefore, for $0 \leq i \leq \ell - 1$ the estimates of Theorem 4.51 for $\|(u - u_{\tau h})^{(i+1)}(t_n^-)\|$ enable upper bounds for $\|(R_h u - u_{\tau h})^{(i)}(t_n^-)\|_V$. However, since results for the $(i + 1)$ th derivative are used to bound the i th derivative, we loose one convergence order in τ and, thus, obtain a suboptimal estimate only.

With a more involved proof, this loss of order in the $L^2(V)$ -estimate can be avoided for $\mathbf{VTD}_k^r(g)$ with $0 \leq k < r$ and $g \in \{\mathcal{I}_k^r f, \mathcal{C}_k^r f\}$ such that we then find

$$\left(\int_{t_0}^{t_n} \|u - u_{\tau h}\|_V^2 dt \right)^{1/2} \leq C(f, u)(h^\kappa + \tau^{r+1}).$$

Furthermore, we can drop Assumptions 4.1 and 4.2, which are otherwise needed if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd). For brevity, we shortly sketch the main ideas only.

Adapting and combining some of the arguments used in the proofs of Lemma 4.19 and Theorem 4.53, we get (also using (4.7) if $k - 2\ell = 1$ ($\Leftrightarrow k$ is odd))

$$\begin{aligned} & \int_{t_0}^{t_n} \|R_h \bar{\Pi}_k^{r, \mathcal{J}} u - u_{\tau h}\|_V^2 dt \\ & \leq C\tau^{2\ell} \int_{t_0}^{t_n} \|\Pi_{r-k+\ell} e_{\tau h, \ell}^{\mathcal{J}}\|_V^2 dt + C(t_n - t_0) \sum_{i=0}^{\ell - \delta_{0, k-2\ell}} \tau^{2i} \max_{\nu=1, \dots, n} \|(R_h u - u_{\tau h})^{(i)}(t_\nu^-)\|_V^2 \\ & \quad + C\delta_{1, k-2\ell} \tau^{2\ell} \sum_{\nu=1}^n \left(\frac{\tau_\nu}{2}\right) \|\omega_\nu^{\mathcal{J}}(R_h u^{(\ell)})\|_V^2. \end{aligned}$$

We already showed how to suitably estimate the first and the third term on the right-hand side, so we focus on the second term. Applying the postprocessing to the solution $u_{\tau h}$ of $\mathbf{VTD}_k^r(\mathcal{I}_k^r f)$ or $\mathbf{VTD}_k^r(\mathcal{C}_k^r f)$, $0 \leq k < r$, we obtain the solution $\tilde{u}_{\tau h}$ of $\mathbf{VTD}_{k+2}^{r+1}(\mathcal{I}_{k,*}^{r+1} f)$ or $\mathbf{VTD}_{k+2}^{r+1}(\mathcal{C}_{k+2}^{r+1} f)$, respectively. Then, from Theorem 4.51 (and with a suitable choice of the discrete initial condition) we find

$$\|(u - \tilde{u}_{\tau h})^{(i)}(t_\nu^-)\| \leq C(f, u)(h^{\kappa+\sigma} + \tau^{(r+1)+1-i}) \quad \text{for } 0 \leq i \leq \ell + 1,$$

which, by (4.18), gives $\|(R_h u - \tilde{u}_{\tau h})^{(i)}(t_\nu^-)\|_V \leq C(f, u)(h^{\kappa+\sigma} + \tau^{r+1-i})$ for $0 \leq i \leq \ell$. Hence, since by construction of the postprocessing $\tilde{u}_{\tau h}^{(i)}(t_\nu^-) = u_{\tau h}^{(i)}(t_\nu^-)$ holds true for $0 \leq i \leq \ell$, we also gain a suitable bound for the remaining terms. ♣

Remark 4.58 (Comments on estimates for the time derivative(s) of the error)

Suitably adapting the arguments of Theorem 4.53 and Remark 4.57, also estimates for the i th time derivative, $0 \leq i \leq \ell$, of the error in the $L^2(H)$ - and the $L^2(V)$ -norm are possible. The respective convergence order in τ then is reduced by i . \clubsuit

Remark 4.59 (Comments on pointwise estimates in the H -norm)

If $u|_{I_\nu} \in C^\ell(I_\nu, H)$, we easily find that

$$\sup_{t \in I_\nu} \|(u - u_{\tau h})(t)\| \leq \sum_{i=0}^{\ell-1} \tau_\nu^i \|(u - u_{\tau h})^{(i)}(t_\nu^-)\| + \tau_\nu^\ell \sup_{t \in I_\nu} \|(u - u_{\tau h})^{(\ell)}(t)\|,$$

where amongst others the fundamental theorem of calculus and properties of the Bochner integral are used to show this. Therefore, Theorem 4.32 and Theorem 4.51 imply appropriate bounds for the pointwise error of $u - u_{\tau h}$ in the H -norm. \clubsuit

Remark 4.60 (Superconvergence in time (mesh) points for cascadic interpolation)

Recalling the observations of Subsection 1.4.3, especially Remark 1.38, we have that the solutions $u_{\tau h}$ of $\mathbf{VTD}_k^r(\mathcal{C}_k^r f)$ and $\tilde{u}_{\tau h}$ of $\mathbf{VTD}_{2r-k}^{2r-k}(\mathcal{I}_{2r-k}^{2r-k} f)$ coincide in the time (mesh) points t_ν^- . Therefore, with a suitably chosen discrete initial condition, from Theorem 4.51 it follows

$$\|(u - u_{\tau h})(t_\nu^-)\| = \|(u - \tilde{u}_{\tau h})(t_\nu^-)\| \leq C(f, u)(h^{\kappa+\sigma} + \tau^{2r-k+1}).$$

Hence, by using cascadic interpolation of the right-hand side f , we can recover the high superconvergence order in the time (mesh) points of $2r - k + 1$ as known for non-stiff initial value problems, cf. Subsection 1.2.3. \clubsuit

4.4 Numerical results

In this section, we want to illustrate our theoretical findings and error estimates by some numerical results. For simplicity, we only consider test problems that are one-dimensional with respect to space, even more concrete, we always consider $\Omega = (0, 1)$.

To this end, for different test situations, the error in the (semi-)norms

$$\begin{aligned} \|v\|_{L^2(L^2)} &= \left(\int_I \int_\Omega |v(t, x)|^2 dx dt \right)^{1/2}, & |v|_{L^2(H^1)} &= \left(\int_I \int_\Omega |\partial_x v(t, x)|^2 dx dt \right)^{1/2}, \\ |v|_{H^1(L^2)} &= \left(\int_I \int_\Omega |\partial_t v(t, x)|^2 dx dt \right)^{1/2}, & |v|_{H^1(H^1)} &= \left(\int_I \int_\Omega |\partial_t \partial_x v(t, x)|^2 dx dt \right)^{1/2}, \\ \|v\|_{\ell^\infty(L^2)} &= \max_{1 \leq n \leq N} \left(\int_\Omega |v(t_n^-, x)|^2 dx \right)^{1/2} \end{aligned}$$

is investigated numerically. The numerical experiments were performed with the software Julia [18], where the floating point data type `BigFloat` with 512 bits was used for all calculations.

We start with the following test problem known from [11, Section 5]. Note that the right-hand side given in the reference had to be corrected.

Example (cf. [11, Section 5])

We consider the one-dimensional heat equation

$$\begin{aligned} u_t(t, x) - u_{xx}(t, x) &= f(t, x) \quad \text{for } (t, x) \in (0, 3) \times (0, 1), \\ u(t, 0) &= u(t, 1) = 0 \quad \text{for } t \in (0, 3), \\ u(0, x) &= 0 \quad \text{for } x \in (0, 1) \end{aligned} \quad (4.24a)$$

with

$$f(t, x) = 3x \cos\left(\frac{3\pi}{2}x\right) \cos(3t) + \left(3\pi \sin\left(\frac{3\pi}{2}x\right) + \left(\frac{3\pi}{2}\right)^2 x \cos\left(\frac{3\pi}{2}x\right)\right) \sin(3t), \quad (4.24b)$$

which results in

$$u(t, x) = x \cos\left(\frac{3\pi}{2}x\right) \sin(3t)$$

as exact solution.

The errors in different (semi-)norms of the $Q_k^6\text{-VTD}_k^6$ method, $k \in \{3, 4\}$, in time and continuous finite elements of piecewise polynomial degree $\kappa \in \{5, 6, 7\}$ in space are considered for problem (4.24). Hereby, the discrete initial values are determined according to (4.1) with $\tilde{P}_h^0 = P_h$. The same number of mesh intervals in t and in x direction is used such that we have $\tau = 3h$. Therefore, we expect that the minimum of temporal and spatial convergence order can be seen, which, according to Theorem 4.53, in the $L^2(L^2)$ -norm is $\min\{r+1, \kappa+1\}$. Moreover, according to Remark 4.57 and Remark 4.58, we expect that the temporal order is reduced by one when the error of the derivative in time is considered and analogously that the spatial order reduces by one if the derivative in space occurs.

The numerical results of Figure 4.1 nicely support all these expectations. Firstly, the orders of convergence turn out to be independent of k . Secondly, for $\kappa = 6$, we see $L^2(L^2)$ -order $\min\{6+1, 6+1\} = 7$ while the $L^2(H^1)$ -, $H^1(L^2)$ -, and $H^1(H^1)$ -order only is 6, which exactly meets our prediction since these (semi-)norms contain at least a derivative in time or space. Thirdly, for $\kappa = 5$, the spatial order is less than the temporal order and, as predicted, we also see maximal order 6 which reduces to 5 if the first derivative with respect to space is contained in the respective (semi-)norm. Fourthly, for $\kappa = 7$, the behavior is just the other way around. In this setting the spatial order is greater than the temporal order, thus, the temporal order is the restricting one. So, we have and also see order 7 if no temporal derivative is involved and order 6 if the respective (semi-)norm includes the first time derivative.

Next, we want to numerically investigate the consequences of different choices for the discrete initial condition. To this end, we consider selected $Q_k^r\text{-VTD}_k^r$ methods in time in combination with continuous finite elements of piecewise polynomial degree $\kappa = r$ for problem (4.24) where once the discrete initial values are determined in “downward direction” according to (4.1) with $\tilde{P}_h^0 = P_h$ and once in “upward direction” by (4.23). We again use meshes with the same number N of uniform subintervals in space and time where $N = 2^i$, $i = 3, \dots, 8$. The errors in the $L^2(L^2)$ -norm as well as the associated experimental orders of convergence for $r \in \{12, 13\}$ and $k \in \{10, 11\}$ are given in Tables 4.1 and 4.2.

If the discrete initial values are defined “downward” via (4.1), we obtain $L^2(L^2)$ -order $r+1$ for all considered methods, which exactly meets our theoretical prediction. In comparison,

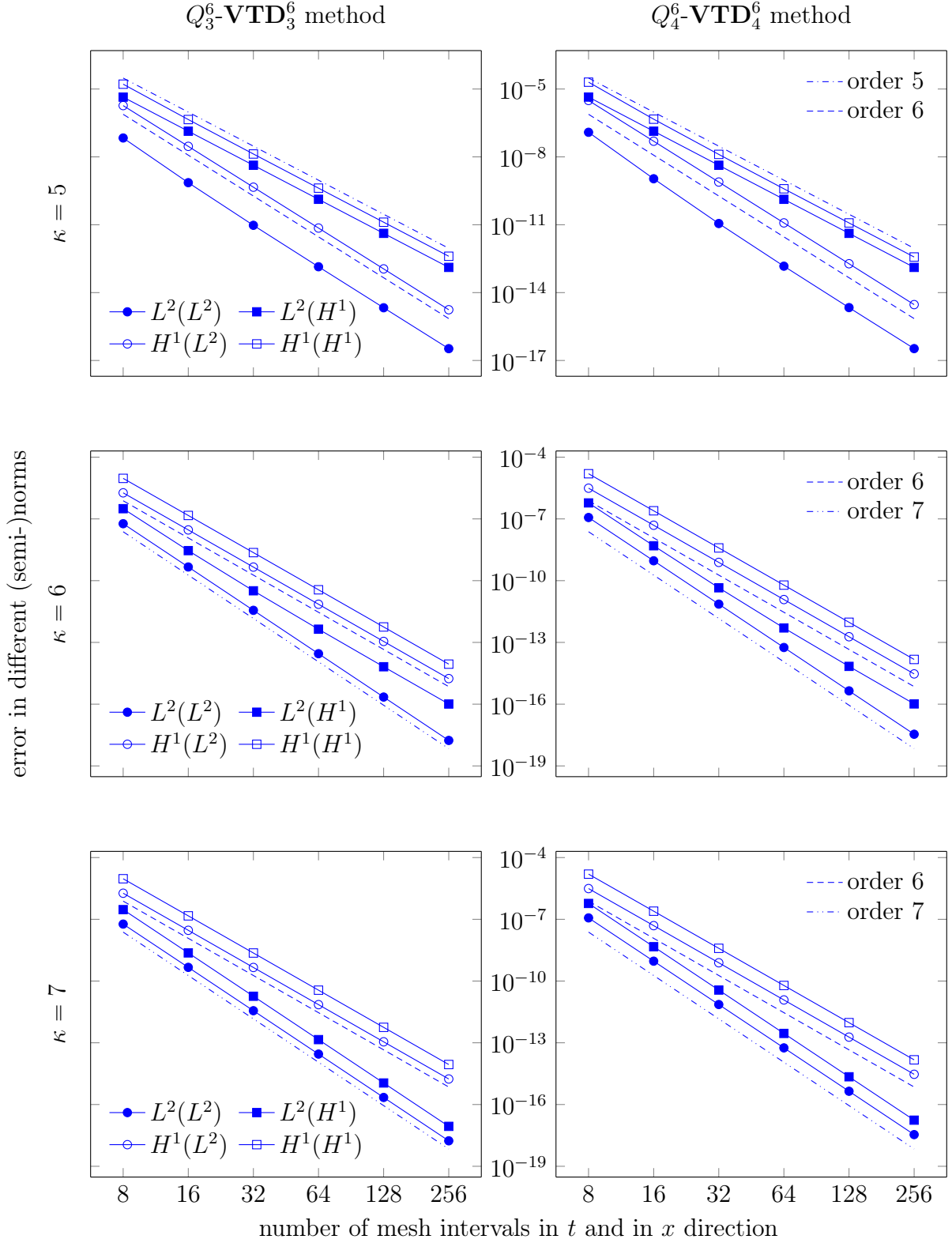


Figure 4.1: Errors in different (semi-)norms of the $Q_k^6\text{-VTD}_k^6$ method, $k \in \{3, 4\}$, in time and continuous P_κ -finite elements, $\kappa \in \{5, 6, 7\}$, in space for problem (4.24)

Table 4.1: Errors in $L^2(L^2)$ -norm and experimental orders of convergence for Q_k^{12} -**VTD** $_k^{12}$, $k \in \{10, 11\}$, in time with discrete initial values determined “down-/upward” via (4.1)/(4.23) and continuous P_{12} -finite elements in space for problem (4.24)

N	discrete initial values via (4.1)				discrete initial values via (4.23)			
	Q_{10}^{12} - VTD $_{10}^{12}$		Q_{11}^{12} - VTD $_{11}^{12}$		Q_{10}^{12} - VTD $_{10}^{12}$		Q_{11}^{12} - VTD $_{11}^{12}$	
	error	eoc	error	eoc	error	eoc	error	eoc
8	1.384e-15	12.987	4.053e-15	12.986	5.798e-11	12.307	7.661e-06	10.768
16	1.705e-19	12.997	4.995e-19	12.996	1.144e-14	12.429	4.393e-09	10.920
32	2.086e-23	12.999	6.113e-23	12.999	2.075e-18	12.460	2.268e-12	10.969
64	2.548e-27	13.000	7.468e-27	13.000	3.684e-22	12.435	1.131e-15	10.986
128	3.111e-31	13.000	9.119e-31	13.000	6.652e-26	12.313	5.579e-19	10.986
256	3.797e-35		1.113e-34		1.307e-29		2.750e-22	

Table 4.2: Errors in $L^2(L^2)$ -norm and experimental orders of convergence for Q_k^{13} -**VTD** $_k^{13}$, $k \in \{10, 11\}$, in time with discrete initial values determined “down-/upward” via (4.1)/(4.23) and continuous P_{13} -finite elements in space for problem (4.24)

N	discrete initial values via (4.1)				discrete initial values via (4.23)			
	Q_{10}^{13} - VTD $_{10}^{13}$		Q_{11}^{13} - VTD $_{11}^{13}$		Q_{10}^{13} - VTD $_{10}^{13}$		Q_{11}^{13} - VTD $_{11}^{13}$	
	error	eoc	error	eoc	error	eoc	error	eoc
8	2.128e-17	14.000	5.166e-17	14.000	1.488e-12	12.526	1.729e-07	11.077
16	1.299e-21	14.000	3.153e-21	14.000	2.522e-16	12.510	8.007e-11	11.029
32	7.931e-26	14.000	1.924e-25	14.000	4.324e-20	12.503	3.832e-14	11.011
64	4.840e-30	14.000	1.174e-29	14.000	7.449e-24	12.501	1.857e-17	11.004
128	2.954e-34	14.000	7.168e-34	14.000	1.285e-27	12.500	9.041e-21	11.002
256	1.803e-38		4.375e-38		2.218e-31		4.409e-24	

the computed errors are considerably larger and the associated experimental convergence orders are clearly reduced if the discrete initial values are defined “upward” via (4.23). This suggests that the rather complicated construction (4.1) for the discrete initial value is really necessary.

A closer look at the computational results of Tables 4.1 and 4.2 moreover shows that, depending on whether r and k are even or odd, there are significant differences in the gap between the $L^2(L^2)$ -convergence order $r + 1$ expected for “downward” initial values and the experimental order of convergence obtained for “upward” initial values. Therefore, for closer examinations, the deficit compared to $r + 1$ of the $L^2(L^2)$ -convergence orders obtained for discrete initial values determined via (4.23) are given in Table 4.3 for Q_k^r -**VTD** $_k^r$ methods with $5 \leq k \leq r \leq 13$. Here, the deficits are calculated using the experimental orders of convergence computed from the $L^2(L^2)$ -errors for $N \in \{128, 256\}$.

First of all, from our computational results we see no deficit in the $L^2(L^2)$ -convergence order for $k \leq 6$. However, for $k \geq 7$ the situation is quite different. For odd r , we see a

Table 4.3: Deficit compared to $r + 1$ of the experimental $L^2(L^2)$ -orders of convergence for $Q_k^r\text{-VTD}_k^r$, $5 \leq k \leq r \leq 13$, in time with discrete initial values determined “upward” via (4.23) and continuous P_r -finite elements in space for problem (4.24)

r	$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$	$k = 11$	$k = 12$	$k = 13$
5	0.000								
6	0.000	0.000							
7	0.000	0.000	0.999						
8	0.000	0.000	0.735	0.228					
9	0.000	0.000	1.000	0.500	1.998				
10	0.000	0.000	0.808	0.302	1.138	0.640			
11	0.000	0.000	1.001	0.501	1.999	1.499	2.997		
12	0.000	0.000	0.847	0.329	1.184	0.687	2.014	1.514	
13	0.000	0.000	1.000	0.501	2.000	1.500	2.998	2.498	3.996

deficit of 1 for $k = 7$ and of 0.5 for $k = 8$. Further, incrementing $k \geq 7$ by two, increments the observed deficit by one. For even r , we also see an enlargement of the gap between $r + 1$ and the obtained experimental $L^2(L^2)$ -order when $k \geq 7$ is incremented by two. But the increase of the deficit is lower as for odd r and we do not observe clear full or half convergence orders. While the differences between even and odd k may be explained by the different stability properties or differences in the construction of dG-like and cGP-like methods, the observed differences between even and odd r here at the example are rather surprising and we have no direct explanation for them.

In order to examine certain specific features of the variational time discretizations more easily, we in addition study a problem with a solution that is polynomial in space and, thus, allows to almost exclude the spatial error.

Example

We consider the instationary convection-diffusion-reaction problem

$$\begin{aligned}
 u_t(t, x) - u_{xx}(t, x) + (1 + x^2)u_x(t, x) + (1 + 2x)u(t, x) &= f(t, x) \quad \text{for } (t, x) \in (0, 2) \times (0, 1), \\
 u(t, 0) = u(t, 1) = 0 &\quad \text{for } t \in (0, 2), \\
 u(0, x) = x^2(1 - x) &\quad \text{for } x \in (0, 1)
 \end{aligned} \tag{4.25a}$$

with f chosen such that

$$u(t, x) = x(1 - x)(x \cos(t) - \sin(2t)) \tag{4.25b}$$

is the exact solution.

For approximation in space we use continuous, piecewise cubic finite elements where the spatial interval $(0, 1)$ is decomposed into 10 uniform subintervals. Note that this proper choice of the trial space for the spatial discretization ensures that the error in space is

negligible. So, the numerical results for problem (4.25) reflect the error behavior of the variational time discretization method.

At first, we have a look at the superconvergence behavior in the time mesh points. However, according to Subsection 4.2.2, in this parabolic setting for $Q_k^r\text{-VTD}_k^r$ methods we only can expect a low order superconvergence and not the high superconvergence order of $2r - k + 1$ as known from the analysis of non-stiff ode systems. This can also be seen in the results presented in Table 4.4 where errors in the $L^2(L^2)$ - and the $\ell^\infty(L^2)$ -norm as well as their associated experimental orders of convergence are given for the $Q_0^6\text{-VTD}_0^6$ method on time meshes with N uniform subintervals where $N = 2^i$, $i = 5, \dots, 13$. While the $L^2(L^2)$ -order is $r + 1 = 7$ as predicted by theory, the high superconvergence order, which is $2r - k + 1 = 13$, is clearly not obtained, even for quite small time steps. This suggests that in general additional compatibility conditions, as those in [52, p. 211], really are needed for higher order superconvergence.

The situation is quite different if we apply cascadic interpolation to the function f on the right-hand side. Corresponding computational results for $Q_0^6\text{-VTD}_0^6(\mathcal{C}_0^6 f)$ are also given in Table 4.4. We observe that the error in the $L^2(L^2)$ -norm is almost the same as for the standard method without cascade, but the $\ell^\infty(L^2)$ -norm is considerably smaller. Moreover, now the desired high superconvergence order of 13 is achieved even for quite large time steps, which is in accordance with the theoretical result of Remark 4.60. This suggests that the application of cascadic interpolation can be quite advantageous.

Table 4.4: Errors and experimental orders of convergence for $Q_0^6\text{-VTD}_0^6$ without and with interpolation cascade of the right-hand side for problem (4.25)

N	without cascade				with cascade			
	$\ u - u_{\tau h}\ _{L^2(L^2)}$		$\ u - u_{\tau h}\ _{\ell^\infty(L^2)}$		$\ u - u_{\tau h}\ _{L^2(L^2)}$		$\ u - u_{\tau h}\ _{\ell^\infty(L^2)}$	
	error	eoc	error	eoc	error	eoc	error	eoc
32	2.016e-15	7.000	4.595e-20	9.263	2.016e-15	7.000	3.598e-28	12.963
64	1.575e-17	7.000	7.477e-23	9.236	1.575e-17	7.000	4.506e-32	12.986
128	1.231e-19	7.000	1.240e-25	9.261	1.231e-19	7.000	5.554e-36	12.994
256	9.616e-22	7.000	2.022e-28	9.517	9.616e-22	7.000	6.809e-40	12.997
512	7.513e-24	7.000	2.761e-31	9.796	7.513e-24	7.000	8.328e-44	12.999
1024	5.869e-26	7.000	3.105e-34	9.963	5.869e-26	7.000	1.018e-47	12.999
2048	4.586e-28	7.000	3.112e-37	10.828	4.586e-28	7.000	1.243e-51	13.000
4096	3.582e-30	7.000	1.712e-40	11.866	3.582e-30	7.000	1.517e-55	13.000
8192	2.799e-32		4.586e-44		2.799e-32		1.852e-59	

In Table 4.5 computational results for the $Q_3^6\text{-VTD}_3^6$ method without and with interpolation cascade are presented. The behavior is quite similar as for $Q_0^6\text{-VTD}_0^6$. While the errors in the $L^2(L^2)$ -norm are almost equal for the standard method and the method with cascadic interpolation and show the predicted convergence order $r + 1 = 7$, the errors in the time mesh points reveal considerable differences between both methods. So, with cascade the high superconvergence order $2r - k + 1 = 10$ is obtained already for coarse grids, whereas without cascade this order is clearly underachieved. Though, the differences are

not as prominent as for $Q_0^6\text{-VTD}_0^6$, which is certainly due to the smaller difference between $L^2(L^2)$ -order 7 and high superconvergence order 10.

Table 4.5: Errors and experimental orders of convergence for $Q_3^6\text{-VTD}_3^6$ without and with interpolation cascade of the right-hand side for problem (4.25)

N	without cascade				with cascade			
	$\ u - u_{\tau h}\ _{L^2(L^2)}$		$\ u - u_{\tau h}\ _{\ell^\infty(L^2)}$		$\ u - u_{\tau h}\ _{L^2(L^2)}$		$\ u - u_{\tau h}\ _{\ell^\infty(L^2)}$	
	error	eoc	error	eoc	error	eoc	error	eoc
32	5.071e-15	7.000	1.315e-18	9.331	5.072e-15	7.000	2.817e-21	9.983
64	3.962e-17	7.000	2.042e-21	9.290	3.962e-17	7.000	2.784e-24	9.996
128	3.095e-19	7.000	3.263e-24	9.267	3.095e-19	7.000	2.726e-27	9.999
256	2.418e-21	7.000	5.295e-27	9.242	2.418e-21	7.000	2.663e-30	10.000
512	1.889e-23	7.000	8.746e-30	9.255	1.889e-23	7.000	2.601e-33	10.000
1024	1.476e-25	7.000	1.431e-32	9.414	1.476e-25	7.000	2.541e-36	10.000
2048	1.153e-27	7.000	2.099e-35	9.572	1.153e-27	7.000	2.481e-39	10.000
4096	9.008e-30	7.000	2.758e-38	9.719	9.008e-30	7.000	2.423e-42	10.000
8192	7.038e-32		3.273e-41		7.038e-32		2.366e-45	

In order to enable also an easy comparison of the variational time discretization methods for different choices of the method parameter k , we present in Table 4.6 the computational results for different versions of $Q_k^6\text{-VTD}_k^6$ with $k = 0, \dots, 6$ for problem (4.25). In addition to the standard method and the method with cascadic interpolation, we now also consider the postprocessing of the methods without and with cascade. Note that the errors, given in various (semi-)norms, are those obtained for a time mesh consisting of $N = 256$ uniform subintervals. Moreover, the listed associated experimental orders of convergence were calculated from the errors for $N \in \{256, 512\}$.

The numerical results of Table 4.6 once again reflect many features of the variational time discretization methods that we have observed and discussed earlier. We will therefore highlight only a few aspects.

For $r = k = 6$ using the interpolation cascade has no effect on the computational results. This is because the methods $Q_6^6\text{-VTD}_6^6$ and $Q_6^6\text{-VTD}_6^6(\mathcal{C}_6^6 f)$ are equivalent. Moreover, postprocessing has no effect on the $\ell^\infty(L^2)$ -norm of the error if $0 \leq k \leq r = 6$ and on the $\ell^\infty(L^2)$ -norm of the first time derivative of the error if $2 \leq k \leq r = 6$. This is because postprocessing, by construction, preserves function and derivative values up to derivative order $\lfloor \frac{k}{2} \rfloor$ in the time mesh points.

The errors in the $L^2(L^2)$ -norm and the $H^1(L^2)$ -semi-norm are hardly influenced by the usage of cascadic interpolation. Without postprocessing we see, as expected, $L^2(L^2)$ -order $r + 1 = 7$ and $H^1(L^2)$ -order $r = 6$. Moreover, postprocessing increases the $L^2(L^2)$ -order if $0 \leq k \leq r - 1 = 5$ and the $H^1(L^2)$ -order if $0 \leq k \leq r = 6$ by one. This is in accordance with our theoretical and numerical results from Sections 1.3 and 1.4, also see Remark 4.56.

When using the interpolation cascade for the right-hand side, we observe the high superconvergence order $2r - k + 1 = 13 - k$ for the error in the $\ell^\infty(L^2)$ -norm. For $2 \leq k \leq r = 6$ before postprocessing and $0 \leq k \leq r = 6$ after postprocessing, respectively, this supercon-

vergence order is also obtained for the first time derivative of the error in the time mesh points. Without cascadic interpolation the convergence orders for the errors in the time mesh points are partly considerably smaller. However, a low order superconvergence behavior can be observed for all $0 \leq k \leq r - 1 = 5$, which is in accordance with our theoretical findings.

Summarizing, the numerical results nicely show the properties of the considered variational time discretization methods. The convergence behavior expected from our theoretical error estimates is met and well illustrated.

Table 4.6: Errors and experimental orders of convergence for different versions of $Q_k^6\text{-VTD}_k^6$, $k = 0, \dots, 6$, for problem (4.25)

k	$\ u - u_{\tau h}\ _{L^2(L^2)}$		$ u - u_{\tau h} _{H^1(L^2)}$		$\ u - u_{\tau h}\ _{\ell^\infty(L^2)}$		$\ \partial_t(u - u_{\tau h})\ _{\ell^\infty(L^2)}$	
	error	eoc	error	eoc	error	eoc	error	eoc
(i) standard method								
0	9.616e-22	7.000	4.548e-18	6.000	2.022e-28	9.517	4.301e-18	6.000
1	1.007e-21	7.000	2.342e-18	6.000	6.102e-28	9.404	7.987e-18	6.000
2	1.645e-21	7.000	3.460e-18	6.000	9.839e-28	9.225	1.236e-24	8.339
3	2.418e-21	7.000	3.662e-18	6.000	5.295e-27	9.242	3.989e-24	8.257
4	4.836e-21	7.000	6.190e-18	6.000	1.665e-25	8.983	1.136e-23	8.248
5	1.026e-20	7.000	9.650e-18	6.000	2.828e-23	8.000	3.496e-22	7.989
6	4.139e-20	7.000	2.228e-17	6.000	4.939e-21	6.995	5.754e-20	6.995
(ii) with cascadic interpolation								
0	9.616e-22	7.000	4.548e-18	6.000	6.809e-40	12.997	4.301e-18	6.000
1	1.007e-21	7.000	2.342e-18	6.000	1.137e-36	12.000	7.987e-18	6.000
2	1.645e-21	7.000	3.460e-18	6.000	1.884e-33	10.997	2.195e-32	10.997
3	2.418e-21	7.000	3.662e-18	6.000	2.663e-30	10.000	3.103e-29	10.000
4	4.836e-21	7.000	6.190e-18	6.000	3.731e-27	8.996	4.346e-26	8.996
5	1.026e-20	7.000	9.650e-18	6.000	4.317e-24	8.000	5.029e-23	8.000
6	4.139e-20	7.000	2.228e-17	6.000	4.939e-21	6.995	5.754e-20	6.995
(iii) with postprocessing								
0	6.855e-25	7.999	1.694e-21	6.998	2.022e-28	9.517	6.056e-25	8.777
1	9.741e-25	8.000	1.781e-21	7.000	6.102e-28	9.404	1.657e-24	8.746
2	1.801e-24	7.999	2.903e-21	6.999	9.839e-28	9.225	1.236e-24	8.339
3	3.307e-24	8.000	4.275e-21	7.000	5.295e-27	9.242	3.989e-24	8.257
4	8.527e-24	7.999	8.537e-21	6.999	1.665e-25	8.983	1.136e-23	8.248
5	2.201e-23	8.000	1.813e-20	7.000	2.828e-23	8.000	3.496e-22	7.989
6	4.543e-21	6.999	4.833e-20	6.999	4.939e-21	6.995	5.754e-20	6.995
(iv) with cascadic interpolation and postprocessing								
0	6.868e-25	8.000	1.700e-21	7.000	6.809e-40	12.997	7.932e-39	12.997
1	9.741e-25	8.000	1.781e-21	7.000	1.137e-36	12.000	1.324e-35	12.000
2	1.804e-24	8.000	2.908e-21	7.000	1.884e-33	10.997	2.195e-32	10.997
3	3.307e-24	8.000	4.275e-21	7.000	2.663e-30	10.000	3.103e-29	10.000
4	8.539e-24	8.000	8.551e-21	7.000	3.731e-27	8.996	4.346e-26	8.996
5	3.336e-23	8.000	1.814e-20	7.000	4.317e-24	8.000	5.029e-23	8.000
6	4.543e-21	6.999	4.833e-20	6.999	4.939e-21	6.995	5.754e-20	6.995

Summary and Outlook

We have considered a family of variational time discretization schemes \mathbf{VTD}_k^r with parameters $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, that generalizes the well-known discontinuous Galerkin (dG) method and continuous Galerkin–Petrov (cGP) method. Generalizing the methods and studying the entire family was interesting for several reasons.

On the one hand, the new schemes have useful properties. So, for example, a higher regularity of the discrete solution can be provided. Indeed, in dependence of k we obtain discrete solutions that are $\lfloor \frac{k-1}{2} \rfloor$ -times continuously differentiable with respect to time. Further, holding the local polynomial ansatz degree r constant, the number of unknowns decreases with increasing k . In the extreme case $r = k$ the number of unknowns is (almost) halved. Moreover, under appropriate conditions superconvergence behavior in the time mesh points can be observed also for derivatives up to order $\lfloor \frac{k}{2} \rfloor$.

On the other hand, the unified analysis as well as the observed connections to other discretization schemes, as collocation methods with multiple nodes or Runge–Kutta-like methods, and the observed connections between different variational time discretization methods via postprocessing provide interesting insights and lead to alternative proof techniques. So, for example, in the case of cascadic interpolation we now have a nice and short justification of superconvergence in the time mesh points.

Finally, we want to summarize briefly some of the important results and raise further questions that came up during the extensive investigations.

In Part I, the \mathbf{VTD}_k^r methods, $0 \leq k \leq r$, were studied for initial value problems. For non-stiff ode systems, in Section 1.2, a unified error analysis was established that can be applied in a rather abstract setting which also allows numerical integration and approximation of the “right-hand side”. Beyond pointwise error estimates also superconvergence in time mesh points was shown. However, especially for $k \geq 4$ Assumption 1.4 almost precluded to study nonlinear problems. Therefore, we should ask whether or not similar results can also be proven under weaker assumptions that allow nonlinear problems also for large k . Moreover, it would be quite interesting to investigate the variational time discretization methods in the context of integral equations.

In Section 1.3 a postprocessing technique was provided that under suitable assumptions can be used to improve the discrete solution. In this context we also found out that some of the variational time discretization methods are connected to collocation methods with multiple nodes. Here, we could ask whether for all considered variational time discretization methods the postprocessing can be used to drive an adaptive time step control as it is known from [3] for dG and cGP methods.

Further, in Section 1.4 we considered affine linear problems with time-independent coefficients. We introduced the idea of cascadic interpolation of the right-hand side function in order to enable multiple postprocessing steps. As easy consequence we got a nice proof of superconvergence in the case that the interpolation cascade is used. However, it is open

whether or not the idea of cascadic interpolation can be generalized to more general problems with coefficients that depend on time. Moreover, it would be nice to find a mathematical explanation for the computational results of Table 1.4 which showed that postprocessing based on jumps and postprocessing based on residuals behave very different if multiple postprocessing steps are applied.

In Chapter 2 we derived error estimates also for stiff ode systems. To this end, we fitted the variational time discretization methods into a Runge–Kutta-like framework, see Sections 2.1 and 2.2. Then, in Section 2.3 we transferred the techniques that are usually used to prove the B -convergence of Runge–Kutta methods in order to derive stiff error estimates for the \mathbf{VTD}_k^r methods. However, also here we have restricted the investigations to affine linear problems with time-independent coefficients. Thus, it is still an open question whether for the whole family of methods similar estimates can be shown also for more general problem classes, e.g., for affine linear problems with time-dependent coefficients or certain semilinear problems.

Part II was devoted to the study of variational time discretization methods in the context of parabolic problems with time-independent spatial differential operators and homogeneous boundary conditions. At first, in Chapter 3, we collected some well-known results on the regularity of solutions and the semi-discretization in space that were needed later. Moreover, we presented a full discretization in space and time that was obtained by applying a variational time discretization scheme to the semi-discretization in space.

Then, in Chapter 4 the findings from Part I were combined and transferred to prove error estimates for $\mathbf{VTD}_k^r(g)$ also in the parabolic setting. Using that the $\lfloor \frac{k}{2} \rfloor$ th derivative of the solution of $\mathbf{VTD}_k^r(g)$ actually solves a dG or cGP scheme, respectively, we started in Section 4.1 by showing error estimates for this derivative. Although most estimates were already known from the analysis of dG and cGP methods, this section was very interesting since, firstly, also the application of quadrature rules for approximate integration in time was allowed and, secondly, dG and cGP were studied at once, which nicely showed the similarities and differences in the analysis of the two methods.

Next, in Section 4.2, we had a look on the error in the time (mesh) points. Since in general a superconvergence behavior cannot be observed or at least does not provide sufficiently high orders of convergence, we had to reuse the (stiff) error estimates of Section 2.3 to prove satisfactory error estimates. This, however, also shows that a very detailed investigation of superconvergence could be quite worthwhile since proper adjustments of the methods may be possible if the crucial reason for the lack of high superconvergence is found. In this regard note that we already showed that the high superconvergence order is obtained when cascadic interpolation is used.

Finally, combining all these results, we concluded in Section 4.3 optimal error estimates for full discretizations in space and time that use variational time discretization schemes for approximation in time. Now, an obvious next step would be to allow also inhomogeneous boundary conditions. Here, using the interpolation cascade may also help to treat the issue of order-reduction known in this context. Other approaches that tackle this problem and may be generalized are presented for the dG method in [54, Chapter 3] and for Runge–Kutta methods in [7, 8]. Moreover, in further research, the variational time discretizations may be analyzed also for other problems as, for example, the wave equation or the transient Stokes problem. For wave equations a first step in this direction has been already gone in [9, 12].

Appendix

A Miscellaneous Results

In this section, we want to collect and partly prove miscellaneous results that are needed in this thesis.

A.1 Discrete Gronwall inequality

Discrete versions of the Gronwall lemma are well-known from the literature, see e.g. [52, Lemma 10.5, p. 175], [26, Exercise 67.1, p. 159, Exercise 68.3, pp. 174–175], or [23]. We here prove one further, less common variant.

Lemma A.1 (Discrete Gronwall lemma)

Let $(a_n)_{n \in \mathbb{N}_0}$, $(A_n)_{n \in \mathbb{N}}$, $(B_n)_{n \in \mathbb{N}}$, and $(w_n)_{n \in \mathbb{N}}$ be sequences of real numbers satisfying

$$w_n > 0 \quad \text{and} \quad a_n + A_n \leq B_n + w_n a_{n-1}, \quad n \geq 1.$$

Then, for all $n \geq 1$ it holds

$$a_n + \sum_{\nu=1}^n \left(\prod_{k=\nu+1}^n w_k \right) A_\nu \leq a_0 \left(\prod_{k=1}^n w_k \right) + \sum_{\nu=1}^n \left(\prod_{k=\nu+1}^n w_k \right) B_\nu.$$

If additionally $a_0, A_n, B_n \geq 0$ and $w_n \geq 1$ for all $n \geq 1$, then it follows

$$a_n + \sum_{\nu=1}^n A_\nu \leq \exp \left(\sum_{k=1}^n (w_k - 1) \right) \left(a_0 + \sum_{\nu=1}^n B_\nu \right)$$

for all $n \geq 1$.

Proof. We define some auxiliary variables by $\tilde{a}_n := a_n \left(\prod_{k=1}^n w_k^{-1} \right)$, $n \geq 0$. Then, from the presumed inequality we gain

$$\tilde{a}_\nu - \tilde{a}_{\nu-1} = \left(\prod_{k=1}^\nu w_k^{-1} \right) (a_\nu - w_\nu a_{\nu-1}) \leq \left(\prod_{k=1}^\nu w_k^{-1} \right) (B_\nu - A_\nu).$$

A summation over $\nu = 1, \dots, n$ yields

$$\tilde{a}_n \leq \tilde{a}_0 + \sum_{\nu=1}^n \left(\prod_{k=1}^\nu w_k^{-1} \right) (B_\nu - A_\nu).$$

Recalling the definition of \tilde{a}_n and rearranging the terms, we obtain

$$a_n + \sum_{\nu=1}^n \left(\prod_{k=\nu+1}^n w_k \right) A_\nu \leq a_0 \left(\prod_{k=1}^n w_k \right) + \sum_{\nu=1}^n \left(\prod_{k=\nu+1}^n w_k \right) B_\nu,$$

which is the first estimate.

Because of the additional assumption $w_n \geq 1$ for $n \geq 1$, we have

$$1 \leq \prod_{k=\nu+1}^n w_k \leq \prod_{k=1}^n w_k \leq \prod_{k=1}^n \exp(w_k - 1) = \exp\left(\sum_{k=1}^n (w_k - 1)\right).$$

For $a_0, A_n, B_n \geq 0$ this enables to bound the left-hand side of the first estimate from below and the right-hand side from above in the desired way. \square

A.2 Something about Jacobi-polynomials

The Jacobi-polynomials, denoted by $P_n^{(\alpha, \beta)}(t)$ for $n \in \mathbb{N}_0$, $\alpha, \beta > -1$, form an orthogonal system with respect to the weighting function $w(t) = (1-t)^\alpha(1+t)^\beta$ in the interval $(-1, 1)$, see [1, 22.2.1, p. 774]. They are normalized by setting

$$P_n^{(\alpha, \beta)}(1) = \binom{n + \alpha}{n} \quad (\text{A.1})$$

and satisfy

$$\int_{-1}^1 P_m^{(\alpha, \beta)}(t) P_n^{(\alpha, \beta)}(t) (1-t)^\alpha (1+t)^\beta dt = \frac{2^{\alpha+\beta+1}}{2n + \alpha + \beta + 1} \frac{\Gamma(n + \alpha + 1) \Gamma(n + \beta + 1)}{n! \Gamma(n + \alpha + \beta + 1)} \delta_{m,n}.$$

Hereby, $\Gamma(t)$ is the gamma function and $\delta_{m,n}$ the Kronecker symbol. Note that for $n \in \mathbb{N}_0$ the identity $\Gamma(n + 1) = n!$ holds.

Furthermore, the Jacobi-polynomials satisfy the Rodrigues' formula

$$P_n^{(\alpha, \beta)}(t) = \frac{(-1)^n}{2^n n!} (1-t)^{-\alpha} (1+t)^{-\beta} \frac{d^n}{dt^n} \left[(1-t)^{n+\alpha} (1+t)^{n+\beta} \right],$$

see [1, 22.11.1, p. 785]. From this identity we easily conclude for $0 \leq k \leq n$ that

$$\frac{d^k}{dt^k} \left[(1-t)^{\alpha+k} (1+t)^{\beta+k} P_{n-k}^{(\alpha+k, \beta+k)}(t) \right] = \frac{(-1)^k 2^k n!}{(n-k)!} (1-t)^\alpha (1+t)^\beta P_n^{(\alpha, \beta)}(t). \quad (\text{A.2})$$

B Abstract Projection Operators for Banach Space-Valued Functions

Piecewise polynomial projection operators of real-valued or vector-valued functions are well studied as this is part of the standard finite element interpolation theory, see e.g. [21, Section 3.1] or [25, Chapter 1]. However, to the best of our knowledge for Banach space-valued functions, apart from results on special projection operators, there are no general studies in the literature. Therefore, in this section, a rather abstract definition and rigorous error analysis is presented at least for the univariate, Banach space-valued case.

Here, standard notation for the occurring function spaces is used. For details, especially on the definitions of the norms in Sobolev and Bochner–Sobolev spaces, see page 82. In addition, we list some literature on the basics of Banach space-valued functions and Bochner integration at the end of this section, see Appendix B.3, for easy reference.

B.1 Abstract definition and commutation properties

We start considering assumptions that enable the definition of an abstract polynomial projection operator.

Lemma B.1

Let X denote a Banach space over \mathbb{R} , let $r \in \mathbb{Z}$, $r \geq 0$, and $a, b \in \mathbb{R}$, $a < b$. Assume that there is a Banach space $V((a, b), X)$ with $P_r((a, b), X) \subseteq V((a, b), X) \subseteq L^1((a, b), X)$ and that there are $r + 1$ bounded linear operators $\mathcal{N}_i^X : V((a, b), X) \rightarrow X$, $i = 0, \dots, r$, such that the mapping

$$P_r((a, b), X) \ni v \mapsto (\mathcal{N}_j^X(v))_{j=0, \dots, r} \in X^{r+1} \quad \text{is an isomorphism.} \quad (\text{B.1})$$

Moreover, suppose that there exist functions $\phi_i \in P_r((a, b))$, $i = 0, \dots, r$, such that

$$\mathcal{N}_j^X(w\phi_i) = \delta_{i,j}w \quad \forall i, j = 0, \dots, r, \forall w \in X. \quad (\text{B.2})$$

Then, the projection operator

$$\Pi^X : V((a, b), X) \rightarrow P_r((a, b), X), v \mapsto \sum_{i=0}^r \mathcal{N}_i^X(v)\phi_i,$$

is well-defined and preserves X -valued polynomials of maximal degree r .

If furthermore $k, m \in \mathbb{Z}$, $k, m \geq 0$, and $p, q \in [1, \infty]$ are chosen such that the embedding $W^{k+1,p}((a, b), X) \hookrightarrow V((a, b), X)$ holds true, then Π^X is a bounded linear operator from $W^{k+1,p}((a, b), X)$ to $W^{m,q}((a, b), X)$.

Proof. Obviously, since $\mathcal{N}_i^X : V((a, b), X) \rightarrow X$, $i = 0, \dots, r$, are bounded linear operators, for every function in $v \in V((a, b), X)$ we have that $\Pi^X v = \sum_{i=0}^r \mathcal{N}_i^X(v) \phi_i \in P_r((a, b), X)$ is well-defined. Because of $P_r((a, b), X) \subseteq V((a, b), X)$, we can also apply Π^X to $\Pi^X v$.

We further need to show that Π^X preserves X -valued polynomials of maximal degree r and, thus, also $\Pi^X(\Pi^X v) = \Pi^X v$ for all $v \in V((a, b), X)$. So, let $v \in P_r((a, b), X)$ be arbitrarily chosen. Since v is uniquely described by $(\mathcal{N}_j^X(v))_{j=0, \dots, r}$, it suffices to verify that $\mathcal{N}_j^X(\Pi^X v) = \mathcal{N}_j^X(v)$ for all $j = 0, \dots, r$. But this follows easily from (B.2)

$$\mathcal{N}_j^X(\Pi^X v) = \sum_{i=0}^r \mathcal{N}_j^X(\mathcal{N}_i^X(v) \phi_i) = \sum_{i=0}^r \delta_{i,j} \mathcal{N}_i^X(v) = \mathcal{N}_j^X(v).$$

Therefore, Π^X is a projection operator onto $P_r((a, b), X)$.

Now, it only remains to prove the boundedness of Π^X . Let $v \in W^{k+1,p}((a, b), X)$. Then, due to $W^{k+1,p}((a, b), X) \hookrightarrow V((a, b), X)$, it also holds $v \in V((a, b), X)$ and \mathcal{N}_i^X , $i = 0, \dots, r$, are bounded linear operators from $W^{k+1,p}((a, b), X)$ to X . Moreover, obviously we have $\Pi^X v \in P_r((a, b), X) \subset W^{m,q}((a, b), X)$. Therefore,

$$\begin{aligned} \|\Pi^X v\|_{W^{m,q}((a,b),X)} &\leq \sum_{i=0}^r \|\mathcal{N}_i^X(v) \phi_i\|_{W^{m,q}((a,b),X)} = \sum_{i=0}^r \|\mathcal{N}_i^X(v)\|_X \|\phi_i\|_{W^{m,q}((a,b),\mathbb{R})} \\ &\leq \sum_{i=0}^r C_{\mathcal{N}_i^X} \|v\|_{W^{k+1,p}((a,b),X)} \|\phi_i\|_{W^{m,q}((a,b),\mathbb{R})} = C \|v\|_{W^{k+1,p}((a,b),X)} \end{aligned}$$

with $C = \sum_{i=0}^r C_{\mathcal{N}_i^X} \|\phi_i\|_{W^{m,q}((a,b),\mathbb{R})}$. □

Remark B.2

Note that the existence of suitable $\phi_i \in P_r((a, b))$, $i = 0, \dots, r$, fulfilling (B.2) already follows from the assumptions (B.1) on \mathcal{N}_i^X , $i = 0, \dots, r$, if there are associated linear operators $\mathcal{N}_i^{\mathbb{R}} : P_r((a, b)) \rightarrow \mathbb{R}$, $i = 0, \dots, r$, that satisfy

$$\mathcal{N}_i^{\mathbb{R}}(\langle g, v \rangle_{X',X}) = \langle g, \mathcal{N}_i^X(v) \rangle_{X',X} \quad \forall g \in X', \forall v \in P_r((a, b), X). \quad (\text{B.3})$$

Indeed, let $\tilde{w} \in X$ with $\|\tilde{w}\|_X = 1$. Then, since $P_r((a, b), X) \ni v \mapsto (\mathcal{N}_i^X(v))_{i=0, \dots, r} \in X^{r+1}$ is an isomorphism, there exist functions $\phi_i^{\tilde{w}} \in P_r((a, b), X)$, $i = 0, \dots, r$, such that

$$\mathcal{N}_j^X(\phi_i^{\tilde{w}}) = \delta_{i,j} \tilde{w} \quad \forall i, j = 0, \dots, r.$$

Now, by Hahn–Banach’s theorem there is a $g_{\tilde{w}} \in X'$ satisfying $\langle g_{\tilde{w}}, \tilde{w} \rangle_{X',X} = \|\tilde{w}\|_X = 1$. Using this, we define $\phi_i \in P_r((a, b))$, $i = 0, \dots, r$, by

$$\phi_i = \langle g_{\tilde{w}}, \phi_i^{\tilde{w}} \rangle_{X',X}.$$

It remains to prove that ϕ_i , $i = 0, \dots, r$, fulfill (B.2). So, let $w \in X$ and $g \in X'$ be arbitrarily chosen. Then, using the properties of $\mathcal{N}_j^{\mathbb{R}}$ and of the duality pairing, we obtain

$$\begin{aligned} \langle g, \mathcal{N}_j^X(w \phi_i) \rangle_{X',X} &= \mathcal{N}_j^{\mathbb{R}}(\langle g, w \phi_i \rangle_{X',X}) = \mathcal{N}_j^{\mathbb{R}}(\langle g, w \rangle_{X',X} \phi_i) = \langle g, w \rangle_{X',X} \mathcal{N}_j^{\mathbb{R}}(\phi_i) \\ &= \langle g, w \rangle_{X',X} \mathcal{N}_j^{\mathbb{R}}(\langle g_{\tilde{w}}, \phi_i^{\tilde{w}} \rangle_{X',X}) = \langle g, w \rangle_{X',X} \langle g_{\tilde{w}}, \mathcal{N}_j^X(\phi_i^{\tilde{w}}) \rangle_{X',X} \\ &= \langle g, w \rangle_{X',X} \langle g_{\tilde{w}}, \delta_{i,j} \tilde{w} \rangle_{X',X} = \langle g, w \rangle_{X',X} \delta_{i,j} \langle g_{\tilde{w}}, \tilde{w} \rangle_{X',X} \\ &= \langle g, \delta_{i,j} w \rangle_{X',X} \end{aligned}$$

for $i, j = 0, \dots, r$. Since this holds for arbitrary $g \in X'$, it follows $\mathcal{N}_j^X(w \phi_i) = \delta_{i,j} w$. ♣

Remark B.3

Often the linear operators that are used to define projection operators by the approach of Lemma B.1 have the basic form

$$\mathcal{N}^X : W^{l,1}((a,b), X) \rightarrow X, v \mapsto \int_a^b v^{(l)}(t) t^j dt, \quad \text{with } l, j \in \mathbb{Z}, l \geq 0, j \geq 0, \quad (\text{B.4a})$$

where the integral is interpreted in Bochner sense,

$$\mathcal{N}^X : C^l([a,b], X) \rightarrow X, v \mapsto v^{(l)}(t^*), \quad \text{with } l \in \mathbb{Z}, l \geq 0, \text{ and } t^* \in [a,b], \quad (\text{B.4b})$$

or are linear combinations of those operators. Of course, these operators are bounded.

The properties of the Bochner integral, see [26, Example 64.15, p. 114] or [50, (10.11), p. 182], and of (derivatives of) Banach space-valued functions, see [26, Corollary 64.32 and Lemma 64.34, p. 118] and [57, beginning of the proof of Proposition 3.6, p. 77], also guarantee that (B.3) is satisfied by the linear operators in (B.4) and their linear combinations. ♣

The next lemma shows that under certain conditions the well-definedness of the Banach space-valued projection operator already follows from that of its real-valued analogon.

Lemma B.4

Let X be a Banach space over \mathbb{R} , let $r \in \mathbb{Z}$, $r \geq 0$, and $a, b \in \mathbb{R}$, $a < b$. Suppose that $\mathcal{N}_i^X : P_r((a,b), X) \rightarrow X$, $i = 0, \dots, r$, are linear operators and that $\mathcal{N}_i^{\mathbb{R}} : P_r((a,b)) \rightarrow \mathbb{R}$, $i = 0, \dots, r$, are associated linear operators that fulfill (B.3). Furthermore, assume that the mapping $P_r((a,b)) \ni v \mapsto (\mathcal{N}_j^{\mathbb{R}}(v))_{j=0,\dots,r} \in \mathbb{R}^{r+1}$ is an isomorphism. Then, the functions $\phi_i \in P_r((a,b))$, $i = 0, \dots, r$, that are well-defined by $\mathcal{N}_j^{\mathbb{R}}(\phi_i) = \delta_{i,j}$, $i, j = 0, \dots, r$, also satisfy (B.2). Moreover, it holds (B.1).

Proof. Let $w \in X$ and $g \in X'$ be arbitrarily chosen. Then,

$$\begin{aligned} \langle g, \mathcal{N}_j^X(w\phi_i) \rangle_{X',X} &= \mathcal{N}_j^{\mathbb{R}}(\langle g, w\phi_i \rangle_{X',X}) = \mathcal{N}_j^{\mathbb{R}}(\langle g, w \rangle_{X',X} \phi_i) = \langle g, w \rangle_{X',X} \mathcal{N}_j^{\mathbb{R}}(\phi_i) \\ &= \langle g, w \rangle_{X',X} \delta_{i,j} = \langle g, \delta_{i,j} w \rangle_{X',X} \end{aligned}$$

and, thus, $\mathcal{N}_j^X(w\phi_i) = \delta_{i,j} w$, which is (B.2).

To show (B.1), we first of all note that the mapping $P_r((a,b), X) \ni v \mapsto (\mathcal{N}_j^X(v))_{j=0,\dots,r}$ obviously preserves the vector space structure, so it remains to show bijectivity. The surjectivity is quite clear since for $(v_j)_{j=0,\dots,r} \in X^{r+1}$ we have that $\sum_{i=0}^r v_i \phi_i \in P_r((a,b), X)$ satisfies $\mathcal{N}_j^X(v) = \sum_{i=0}^r \mathcal{N}_j^X(v_i \phi_i) = v_j$ for all $j = 0, \dots, r$ due to (B.2). In order to prove injectivity, let $v, w \in P_r((a,b), X)$ satisfy $\mathcal{N}_j^X(v) = \mathcal{N}_j^X(w)$ for all $j = 0, \dots, r$. Then, for arbitrary $g \in X'$ we also have $\mathcal{N}_j^{\mathbb{R}}(\langle g, v \rangle_{X',X}) = \mathcal{N}_j^{\mathbb{R}}(\langle g, w \rangle_{X',X})$ for all $j = 0, \dots, r$. Since $\langle g, v \rangle_{X',X}$ and $\langle g, w \rangle_{X',X}$ are in $P_r((a,b))$ and, thus, are uniquely determined by $(\mathcal{N}_j^{\mathbb{R}}(\cdot))_{j=0,\dots,r}$, it follows $\langle g, v \rangle_{X',X} = \langle g, w \rangle_{X',X}$. This, of course, holds pointwise in (a,b) and for arbitrary $g \in X'$. Hence, $v = w$. Therefore, also (B.1) is verified. \square

An important commutation property of the projection operator is presented in the following corollary. We mainly use it in Chapter 4 to guarantee that the projections in time commute with bounded linear operators in space.

Corollary B.5

Let X, Y be Banach spaces over \mathbb{R} , let $r \in \mathbb{Z}$, $r \geq 0$, and $a, b \in \mathbb{R}$, $a < b$. Suppose that for $Z \in \{X, Y, \mathbb{R}\}$ there are Banach spaces $V((a, b), Z)$ with

$$P_r((a, b), Z) \subseteq V((a, b), Z) \subseteq L^1((a, b), Z)$$

and bounded linear operators $\mathcal{N}_i^Z : V((a, b), Z) \rightarrow Z$, $i = 0, \dots, r$, satisfying

$$\mathcal{N}_i^{\mathbb{R}}(\langle g, v \rangle_{Z', Z}) = \langle g, \mathcal{N}_i^Z(v) \rangle_{Z', Z} \quad \forall g \in Z', \forall v \in V((a, b), Z),$$

where we tacitly assume that the term on the left-hand side is well-defined, i.e., we assume that $\langle g, v \rangle_{Z', Z} \in V((a, b), \mathbb{R})$ for all $g \in Z'$, $v \in V((a, b), Z)$. Moreover, presume that the mapping $P_r((a, b)) \ni v \mapsto (\mathcal{N}_j^{\mathbb{R}}(v))_{j=0, \dots, r} \in \mathbb{R}^{r+1}$ is an isomorphism and let $\phi_i \in P_r((a, b))$, $i = 0, \dots, r$, satisfy $\mathcal{N}_j^{\mathbb{R}}(\phi_i) = \delta_{i,j}$, $i, j = 0, \dots, r$.

Then, for $Z \in \{X, Y, \mathbb{R}\}$ the projection operators

$$\Pi^Z : V((a, b), Z) \rightarrow P_r((a, b), Z), \quad v \mapsto \sum_{i=0}^r \mathcal{N}_i^Z(v) \phi_i,$$

are well-defined and preserve Z -valued polynomials of maximal degree r .

Let, in addition, $K : X \rightarrow Y$ be a bounded linear operator. For functions $v : [a, b] \rightarrow X$ set $(Kv)(t) := K(v(t))$ for $t \in [a, b]$. If, furthermore, $K(\mathcal{N}_i^X(v)) = \mathcal{N}_i^Y(Kv)$, $i = 0, \dots, r$, for all $v \in V((a, b), X)$, where we tacitly assume that the term on the right-hand side is well-defined, i.e., we assume that $Kv \in V((a, b), Y)$ for all $v \in V((a, b), X)$, then it holds $K(\Pi^X v) = \Pi^Y(Kv)$ for all $v \in V((a, b), X)$.

Proof. Combining Lemma B.4 and Lemma B.1, the stated assumptions imply the well-definedness of the projection operators Π^Z , $Z \in \{X, Y, \mathbb{R}\}$.

The commutation property $K(\Pi^X v) = \Pi^Y(Kv)$ for all $v \in V((a, b), X)$ follows from

$$(K(\Pi^X v))(t) = K((\Pi^X v)(t)) = \sum_{i=0}^r K(\mathcal{N}_i^X(v)) \phi_i(t) = \sum_{i=0}^r \mathcal{N}_i^Y(Kv) \phi_i(t) = (\Pi^Y(Kv))(t)$$

for all $t \in (a, b)$, where especially the linearity of K was used. \square

Remark B.6

Let X, Y be Banach spaces over \mathbb{R} and let $K : X \rightarrow Y$ be a bounded linear operator. For functions $v : [a, b] \rightarrow X$ set $(Kv)(t) := K(v(t))$ for $t \in [a, b]$. Then, from [26, Corollary 64.14, p. 114] and [26, Lemma 64.34, also note Corollary 64.32, p. 118] we have that $K(\mathcal{N}^X(v)) = \mathcal{N}^Y(Kv)$ holds for linear operators of the form (B.4), where especially \mathcal{N}^Y is well-defined for Kv if \mathcal{N}^X is well-defined for v .

Consequently, if Π^Z , $Z \in \{X, Y\}$, are projection operators defined by the approach of Lemma B.1 or Corollary B.5, respectively, where all $\mathcal{N}_i^Z : V((a, b), Z) \rightarrow Z$, $i = 0, \dots, r$, are linear combinations of linear operators of the form (B.4), then $K(\Pi^X v) = \Pi^Y(Kv)$ for all $v \in V((a, b), X)$. \clubsuit

B.2 Projection error estimates

Of course, we are also interested in error estimates for projections of Banach space-valued functions. But, to prove these, we first have a look on some auxiliary results. Here, for convenience, we use on $W^{k,p}((a,b), X)$ the semi-norm $|v|_{W^{k,p}((a,b), X)} := \|\partial_t^k v\|_{L^p((a,b), X)}$.

Lemma B.7 (Poincaré/Friedrichs' inequality)

Let $p \in [1, \infty]$ and $a, b \in \mathbb{R}$, $a < b$. Moreover, let X denote some Banach space. Suppose that $v \in W^{1,p}((a,b), X)$ and $v(t^*) = 0$ for some $t^* \in [a, b]$. Then, it holds

$$\|v\|_{L^p((a,b), X)} \leq (b-a) |v|_{W^{1,p}((a,b), X)}.$$

Furthermore, for $p \in [1, \infty)$ we have

$$\|v\|_{L^\infty((a,b), X)} \leq (b-a)^{(p-1)/p} |v|_{W^{1,p}((a,b), X)}.$$

Note that $v(t^*) = 0$ is a well-defined condition since embedding results yield that functions in $W^{1,p}((a,b), X)$ are continuous on $[a, b]$, see [26, Lemma 64.37(i), p. 120] or [50, Proposition 10.8, p. 190].

Proof. First of all, using that $v(t^*) = 0$ and applying the fundamental theorem of calculus, which also hold for functions in $W^{1,p}((a,b), X)$, see [50, (10.16), p. 187], also note [50, Proposition 10.8, p. 190], we gain for $t \in [a, b]$

$$\|v(t)\|_X = \|v(t) - v(t^*)\|_X = \left\| \int_{t^*}^t \partial_t v(s) \, ds \right\|_X \leq \int_a^b \|\partial_t v(s)\|_X \, ds,$$

where for the last step also properties of the Bochner integral, see [50, Theorem 10.4, p. 182], were exploited.

For $p = \infty$ we now obtain

$$\|v(t)\|_X \leq \int_a^b \|\partial_t v(s)\|_X \, ds \leq (b-a) \|\partial_t v\|_{L^\infty((a,b), X)}.$$

Otherwise, for $p \in (1, \infty)$ applying the Hölder inequality with p and $q = p' = \frac{p}{p-1}$, it follows

$$\begin{aligned} \|v(t)\|_X &\leq \int_a^b \|\partial_t v(s)\|_X \, ds \leq \left(\int_a^b 1^q \, ds \right)^{1/q} \left(\int_a^b \|\partial_t v(s)\|_X^p \, ds \right)^{1/p} \\ &\leq (b-a)^{(p-1)/p} \|\partial_t v\|_{L^p((a,b), X)}. \end{aligned}$$

Summarizing, we have already shown

$$\|v\|_{L^\infty((a,b), X)} \leq \begin{cases} (b-a) \|\partial_t v\|_{L^\infty((a,b), X)}, & p = \infty, \\ (b-a)^{(p-1)/p} \|\partial_t v\|_{L^p((a,b), X)}, & p \in [1, \infty). \end{cases}$$

So, for $p = \infty$ we are done. Further, for $p \in [1, \infty)$ we conclude from this estimate

$$\begin{aligned} \|v\|_{L^p((a,b), X)}^p &= \int_a^b \|v(t)\|_X^p \, dt \leq \|v\|_{L^\infty((a,b), X)}^p \int_a^b 1 \, dt \\ &\leq (b-a)^{p-1} \|\partial_t v\|_{L^p((a,b), X)}^p (b-a) = (b-a)^p \|\partial_t v\|_{L^p((a,b), X)}^p, \end{aligned}$$

which completes the proof. \square

Next, a Banach space-valued version of Deny–Lions’ lemma is shown for the case of univariate functions. The result of this lemma is then, as in standard finite element interpolation theory, one of the main arguments in the proof of projection error estimates.

Lemma B.8 (Banach space-valued Deny–Lions lemma: univariate case)

Let $p \in [1, \infty]$, $k \in \mathbb{Z}$, $k \geq 0$, and $a, b \in \mathbb{R}$, $a < b$. Moreover, let X denote some Banach space. Suppose that $v \in W^{k+1,p}((a, b), X)$. Then, it holds

$$\inf_{q \in P_k((a, b), X)} \|v - q\|_{W^{k+1,p}((a, b), X)} \leq (\exp(1) + 1) \max \{1, (b - a)^{k+1}\} |v|_{W^{k+1,p}((a, b), X)}.$$

Proof. From embedding theorems, which also hold for Banach space-valued functions, cf. [26, Lemma 64.37(i), p. 120], we have that

$$W^{k+1,p}((a, b), X) \hookrightarrow C^k([a, b], X).$$

Thus, $v \in C^k([a, b], X)$ and so we can choose $q_* \in P_k((a, b), X)$ as k th order Taylor polynomial of v at some point $t^* \in (a, b)$, see also [57, (9) in Chapter 3, p. 77], i.e.,

$$q_*(t) = \sum_{i=0}^k \left(\frac{1}{i!} (t - t^*)^i \right) v^{(i)}(t^*) \quad \forall t \in (a, b).$$

By construction it furthermore holds for $0 \leq j \leq k$ that $q_*^{(j)} \in P_{k-j}((a, b), X)$ is the $(k-j)$ th order Taylor polynomial of $v^{(j)}$ at t^* . Then, for $0 \leq j \leq k-1$ Taylor’s theorem [57, Proposition 3.6, p. 77] yields for all $t \in (a, b)$

$$\begin{aligned} & v^{(j)}(t) - q_*^{(j)}(t) \\ &= \left(v^{(j)}(t) - \sum_{i=0}^{k-j-1} \left(\frac{1}{i!} (t - t^*)^i \right) v^{(i+j)}(t^*) \right) - \frac{1}{(k-j)!} (t - t^*)^{k-j} v^{(k)}(t^*) \\ &= \int_0^1 \frac{(1-s)^{k-j-1}}{(k-j-1)!} (t - t^*)^{k-j} v^{(k)}(t^* + s(t - t^*)) \, ds - \frac{1}{(k-j)!} (t - t^*)^{k-j} v^{(k)}(t^*) \\ &= \int_0^1 \frac{(1-s)^{k-j-1}}{(k-j-1)!} (t - t^*)^{k-j} (v^{(k)}(t^* + s(t - t^*)) - v^{(k)}(t^*)) \, ds \\ &= \int_{t^*}^t \frac{(t-s)^{k-j-1}}{(k-j-1)!} (v^{(k)}(s) - v^{(k)}(t^*)) \, ds. \end{aligned}$$

Thus, using Hölder’s inequality, it follows for $0 \leq j \leq k-1$

$$\begin{aligned} \|v^{(j)}(t) - q_*^{(j)}(t)\|_X &\leq \left\| \frac{(t-\cdot)^{k-j-1}}{(k-j-1)!} \right\|_{L^1((a, b))} \|v^{(k)}(\cdot) - v^{(k)}(t^*)\|_{L^\infty((a, b), X)} \\ &\leq \frac{(b-a)^{k-j}}{(k-j)!} \|v^{(k)}(\cdot) - v^{(k)}(t^*)\|_{L^\infty((a, b), X)}. \end{aligned}$$

Therefore, we gain for all $0 \leq j \leq k$ that

$$\begin{aligned} \|v^{(j)} - q_*^{(j)}\|_{L^\infty((a, b), X)} &\leq \frac{(b-a)^{k-j}}{(k-j)!} \|v^{(k)}(\cdot) - v^{(k)}(t^*)\|_{L^\infty((a, b), X)} \\ &\leq \frac{(b-a)^{k-j}}{(k-j)!} \begin{cases} (b-a) |v|_{W^{k+1,\infty}((a, b), X)}, & p = \infty, \\ (b-a)^{(p-1)/p} |v|_{W^{k+1,p}((a, b), X)}, & p \in [1, \infty), \end{cases} \end{aligned}$$

where Lemma B.7 is applied for the last inequality.

Altogether, also noting that $q_*^{(k+1)} \equiv 0$, we now obtain for $p = \infty$ that

$$\begin{aligned} \inf_{q \in P_k((a,b),X)} \|v - q\|_{W^{k+1,\infty}((a,b),X)} &\leq \|v - q_*\|_{W^{k+1,\infty}((a,b),X)} = \max_{0 \leq j \leq k+1} \|v^{(j)} - q_*^{(j)}\|_{L^\infty((a,b),X)} \\ &\leq \max \{1, (b-a)^{k+1}\} |v|_{W^{k+1,\infty}((a,b),X)}. \end{aligned}$$

Additionally using Hölder's inequality, we similarly conclude for $p \in [1, \infty)$ that

$$\begin{aligned} \inf_{q \in P_k((a,b),X)} \|v - q\|_{W^{k+1,p}((a,b),X)} &\leq \|v - q_*\|_{W^{k+1,p}((a,b),X)} \leq \left(\sum_{j=0}^k \|v^{(j)} - q_*^{(j)}\|_{L^p((a,b),X)}^p + |v|_{W^{k+1,p}((a,b),X)}^p \right)^{1/p} \\ &\leq \left(\sum_{j=0}^k (b-a) \|v^{(j)} - q_*^{(j)}\|_{L^\infty((a,b),X)}^p + |v|_{W^{k+1,p}((a,b),X)}^p \right)^{1/p} \\ &\leq \left(\sum_{j=0}^k \left(\frac{(b-a)^{k-j+1}}{(k-j)!} \right)^p |v|_{W^{k+1,p}((a,b),X)}^p + |v|_{W^{k+1,p}((a,b),X)}^p \right)^{1/p} \\ &\leq \underbrace{\left(\sum_{j=0}^k \left(\frac{1}{(k-j)!} \right)^p + 1 \right)^{1/p}}_{\leq (\exp(1)+1)^{1/p}} \max \{1, (b-a)^{k+1}\} |v|_{W^{k+1,p}((a,b),X)}. \end{aligned}$$

This completes the proof. \square

We now are well prepared to prove the following (local) projection error estimates, see [21, Theorem 3.1.4, p. 121] for an analogous result in the case of real-valued functions.

Lemma B.9

Let X denote a Banach space and let $\hat{\Pi}$ denote some approximation operator for X -valued functions on $[\alpha, \beta]$ with $\alpha, \beta \in \mathbb{R}$, $\alpha < \beta$. Moreover, let $k, m \in \mathbb{Z}$, $k, m \geq 0$, and $p, q \in [1, \infty]$ be chosen such that

- $W^{k+1,p}((\alpha, \beta), X) \hookrightarrow W^{m,q}((\alpha, \beta), X)$,
- $\hat{\Pi}\hat{v} = \hat{v}$ for all $\hat{v} \in P_k((\alpha, \beta), X)$, and
- $\hat{\Pi}$ is a bounded linear operator from $W^{k+1,p}((\alpha, \beta), X)$ to $W^{m,q}((\alpha, \beta), X)$.

Then, for the transformed version Π of $\hat{\Pi}$ on (a, b) with $a, b \in \mathbb{R}$, $a < b$, that is defined by $\Pi v = (\hat{\Pi}(v \circ T_{(a,b)})) \circ T_{(a,b)}^{-1}$ with $T_{(a,b)} : (\alpha, \beta) \ni \hat{t} \mapsto a + \frac{b-a}{\beta-\alpha}(\hat{t} - \alpha) \in (a, b)$, we have

$$|v - \Pi v|_{W^{m,q}((a,b),X)} \leq C \left(\frac{b-a}{\beta-\alpha} \right)^{k-m+1+1/q-1/p} |v|_{W^{k+1,p}((a,b),X)} \quad \forall v \in W^{k+1,p}((a,b),X),$$

where C is independent of the interval (a, b) .

Proof. First of all, because of $W^{k+1,p}((\alpha, \beta), X) \hookrightarrow W^{m,q}((\alpha, \beta), X)$ and due to the respective assumption on $\hat{\Pi}$, we have that $(\text{Id} - \hat{\Pi})$ also is a bounded linear operator from $W^{k+1,p}((\alpha, \beta), X)$ to $W^{m,q}((\alpha, \beta), X)$.

Since $\hat{\Pi}$ preserves polynomials of degree less than or equal to k , we obtain that

$$\hat{v} - \hat{\Pi}\hat{v} = (\text{Id} - \hat{\Pi})(\hat{v} - \hat{q}) \quad \forall \hat{v} \in W^{k+1,p}((\alpha, \beta), X), \forall \hat{q} \in P_k((\alpha, \beta), X).$$

Using the boundedness of $(\text{Id} - \hat{\Pi})$, we thus conclude

$$|\hat{v} - \hat{\Pi}\hat{v}|_{W^{m,q}((\alpha, \beta), X)} \leq C \inf_{\hat{q} \in P_k((\alpha, \beta), X)} \|\hat{v} - \hat{q}\|_{W^{k+1,p}((\alpha, \beta), X)} \leq C |\hat{v}|_{W^{k+1,p}((\alpha, \beta), X)},$$

where Lemma B.8 was applied in the last step.

The desired statement then follows by transformation. Of course, it holds

$$(v - \Pi v) \circ T_{(a,b)} = (v \circ T_{(a,b)}) - \hat{\Pi}(v \circ T_{(a,b)}) = \hat{v} - \hat{\Pi}\hat{v}$$

with $\hat{v} = v \circ T_{(a,b)}$. Therefore, we gain

$$\begin{aligned} |v - \Pi v|_{W^{m,q}((a,b), X)} &= \left(\frac{b-a}{\beta-\alpha} \right)^{1/q-m} |\hat{v} - \hat{\Pi}\hat{v}|_{W^{m,q}((\alpha, \beta), X)} \\ &\leq C \left(\frac{b-a}{\beta-\alpha} \right)^{1/q-m} |\hat{v}|_{W^{k+1,p}((\alpha, \beta), X)} = C \left(\frac{b-a}{\beta-\alpha} \right)^{(1/q-m)+(k+1-1/p)} |v|_{W^{k+1,p}((a,b), X)}. \end{aligned}$$

This completes the proof. \square

B.3 Literature references on basics of Banach space-valued functions

So far, various results on Banach space-valued functions, especially from Bochner–Sobolev spaces, were used. In order to provide direct references to further details and the context of these results, we briefly list some literature on the basics of Banach space-valued functions.

For details on continuous and (strong) differentiable functions of one real variable with values in Banach spaces, we refer to [57, Sections 3.1 and 3.2]. A brief overview of the Bochner integral theory is given in [26, Section 64.1] and [50, Section 10.1] for univariate functions. For more general considerations of the Bochner integral and Bochner spaces see [43, Section 1.2]. Further, for the definition of weak derivatives and associated function spaces for univariate Banach space-valued functions, we refer to [26, Section 64.2]. A discussion of Banach space-valued Sobolev spaces with arguments in a multidimensional domain can be found in [43, Section 2.5]. Last but not least, embedding results for Banach space-valued functions in the univariate case are presented, for example, in [26, Lemma 64.37 and Theorem 64.39, p. 120]. In addition, in [10, Section 5] it was shown that Sobolev–Gagliardo–Nirenberg inequalities and Morrey’s embedding theorem carry over from the real-valued to the Banach space-valued case.

C Operators for Interpolation and Projection in Time

In this section, we collect the definitions of the temporal interpolation and projection operators that are used especially in Part II of this thesis. Moreover, we discuss their well-definedness and some of their properties. Here, we restrict ourselves to the local operators on an arbitrary mesh interval $I_n = (t_{n-1}, t_n]$. Throughout this section, let X denote a Banach space over \mathbb{R} .

C.1 Interpolation operators

We start with important interpolation operators and the associated operator of cascadic interpolation.

Definition C.1 (Standard \mathbf{VTD}_k^r interpolation)

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Then, $\mathcal{I}_k^r : C^{\lfloor \frac{k}{2} \rfloor}(\bar{I}_n, X) \rightarrow P_r(I_n, X)$ is defined by

$$\begin{aligned} (\mathcal{I}_k^r v)^{(i)}(t_{n-1}^+) &= v^{(i)}(t_{n-1}^+), & \text{for } i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor, \\ (\mathcal{I}_k^r v)^{(i)}(t_n^-) &= v^{(i)}(t_n^-), & \text{for } i = 0, \dots, \lfloor \frac{k}{2} \rfloor, \\ \mathcal{I}_k^r v(t_{n,i}) &= v(t_{n,i}), & \text{for } i = 1, \dots, r-k, \end{aligned}$$

with $t_{n,i} := \frac{t_n + t_{n-1}}{2} + \frac{\tau_n}{2} \hat{t}_i$, where \hat{t}_i , $i = 1, \dots, r-k$, denote the zeros of the $(r-k)$ th Jacobi-polynomial $P_{r-k}^{(\lfloor \frac{k}{2} \rfloor + 1, \lfloor \frac{k-1}{2} \rfloor + 1)}$ with respect to the weight $(1 + \hat{t})^{\lfloor \frac{k-1}{2} \rfloor + 1} (1 - \hat{t})^{\lfloor \frac{k}{2} \rfloor + 1}$.

The interpolation operator is of Hermite-type and, in any case, the number of linear independent interpolation conditions is

$$r - k + \lfloor \frac{k}{2} \rfloor + 1 + \lfloor \frac{k-1}{2} \rfloor + 1 = r - k + k - 1 + 2 = r + 1.$$

Hence, the interpolation operator \mathcal{I}_k^r is well-defined. ♣

Definition C.2 (Extended \mathbf{VTD}_k^r interpolation)

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Then, $\mathcal{I}_{k,*}^{r+1} : C^{\lfloor \frac{k+1}{2} \rfloor}(\bar{I}_n, X) \rightarrow P_{r+1}(I_n, X)$ is defined by

$$\begin{aligned} (\mathcal{I}_{k,*}^{r+1} v)^{(i)}(t_{n-1}^+) &= v^{(i)}(t_{n-1}^+), & \text{for } i = 0, \dots, \lfloor \frac{k+1}{2} \rfloor, \\ (\mathcal{I}_{k,*}^{r+1} v)^{(i)}(t_n^-) &= v^{(i)}(t_n^-), & \text{for } i = 0, \dots, \lfloor \frac{k}{2} \rfloor, \\ \mathcal{I}_{k,*}^{r+1} v(t_{n,i}) &= v(t_{n,i}), & \text{for } i = 1, \dots, r-k, \end{aligned}$$

with $t_{n,i} := \frac{t_n + t_{n-1}}{2} + \frac{\tau_n}{2} \hat{t}_i$, where \hat{t}_i , $i = 1, \dots, r-k$, denote the zeros of the $(r-k)$ th Jacobi-polynomial $P_{r-k}^{(\lfloor \frac{k}{2} \rfloor + 1, \lfloor \frac{k-1}{2} \rfloor + 1)}$ with respect to the weight $(1 + \hat{t})^{\lfloor \frac{k-1}{2} \rfloor + 1} (1 - \hat{t})^{\lfloor \frac{k}{2} \rfloor + 1}$.

In short, $\mathcal{I}_{k,*}^{r+1}v \in P_{r+1}(I_n, X)$ satisfies for $v \in C^{\lfloor \frac{k+1}{2} \rfloor}(\bar{I}_n, X)$ all $r+1$ interpolation conditions of \mathcal{I}_k^r and additionally interpolates the $\lfloor \frac{k+1}{2} \rfloor$ th derivative at t_{n-1}^+ .

Note that we also used the two other extensions $\mathcal{I}_{k,\otimes}^{r+1} : C^{\lfloor \frac{k}{2} \rfloor + 1}(\bar{I}_n, X) \rightarrow P_{r+1}(I_n, X)$ and $\mathcal{I}_{k,\diamond}^{r+1} : C^{\lfloor \frac{k}{2} \rfloor}(\bar{I}_n, X) \rightarrow P_{r+1}(I_n, X)$ of \mathcal{I}_k^r that additionally interpolate the $(\lfloor \frac{k}{2} \rfloor + 1)$ th derivative at t_n^- or the function value in one further inner point, respectively. ♣

Definition C.3 (Cascadic interpolation)

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Then, $\mathcal{C}_k^r : C^{r - \lfloor \frac{k+1}{2} \rfloor}(\bar{I}_n, X) \rightarrow P_r(I_n, X)$ is defined by

$$\mathcal{C}_k^r := \mathcal{I}_k^r \circ \mathcal{I}_{k+2}^{r+1} \circ \dots \circ \mathcal{I}_{2r-k}^{2r-k}.$$

Of course, in general \mathcal{C}_k^r itself is not an interpolation operator. But it is a composition of interpolation operators. ♣

C.2 Projection operators

Here, we present the projection operators that are involved in the error analysis. Note that for all considered projection operators, because of Remark B.3 and Lemma B.5, it is sufficient to study the well-definedness for the real-valued version of the operators.

Definition C.4 (L^2 -projection onto polynomials of maximal degree m)

Let $m \in \mathbb{Z}$, $m \geq 0$. Then, $\Pi_m : L^2(I_n, X) \rightarrow P_m(I_n, X)$ is defined by

$$\int_{I_n} (v - \Pi_m v) w \, dt = 0 \quad \forall w \in P_m(I_n),$$

i.e., $\Pi_m v \in P_m(I_n, X)$ denotes the L^2 -projection of $v \in L^2(I_n, X)$ onto polynomials of maximal degree m .

In order to show that Π_m is a well-defined projection operator, it suffices to consider the case $X = \mathbb{R}$. For $v \in P_m(I_n)$ we can choose $w = v - \Pi_m v \in P_m(I_n)$ as test function, which then yields $\int_{I_n} (v - \Pi_m v)^2 \, dt = 0$. Thus, $v = \Pi_m v$ for all $v \in P_m(I_n)$, which implies the well-definedness if $X = \mathbb{R}$. ♣

Definition C.5

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$. Then, $\Pi_k^r : C^{\lfloor \frac{k-1}{2} \rfloor}(\bar{I}_n, X) \rightarrow P_r(I_n, X)$ is defined by

$$\begin{aligned} (v - \Pi_k^r v)^{(i)}(t_{n-1}^+) &= 0, & \text{if } k \geq 1, i = 0, \dots, \lfloor \frac{k-1}{2} \rfloor, \\ (v - \Pi_k^r v)^{(i)}(t_n^-) &= 0, & \text{if } k \geq 2, i = 0, \dots, \lfloor \frac{k}{2} \rfloor - 1, \\ \int_{I_n} (v - \Pi_k^r v) w \, dt &= 0 & \forall w \in P_{r-k}(I_n). \end{aligned}$$

Note that it holds $\Pi_0^r v = \Pi_r v$ for all $v \in L^2(I_n, X)$.

For $v \in P_r(I_n)$ and using the point conditions at t_{n-1}^+ and t_n^- , we get from polynomial long division that $(v - \Pi_k^r v)(t) = (t - t_{n-1})^{\lfloor \frac{k-1}{2} \rfloor + 1} (t_n - t)^{\lfloor \frac{k}{2} \rfloor} \tilde{w}(t)$ where $\tilde{w} \in P_{r-k}(I_n)$. So,

choosing test function $w = \tilde{w}$, we gain that $\int_{I_n} (t - t_{n-1})^{\lfloor \frac{k-1}{2} \rfloor + 1} (t_n - t)^{\lfloor \frac{k}{2} \rfloor} \tilde{w}^2(t) dt = 0$ from which we conclude that $\tilde{w} \equiv 0$ and, thus, $v - \Pi_k^r v \equiv 0$ for $v \in P_r(I_n)$. This implies the well-definedness of Π_k^r if $X = \mathbb{R}$. \clubsuit

Definition C.6 (Standard dG/cGP projection)

Let $l \in \{0, 1\}$ and $m \in \mathbb{Z}$, $m \geq l$. Then, $\tilde{\Pi}_l^m : C(\bar{I}_n, X) \rightarrow P_m(I_n, X)$ is defined by

$$(v - \tilde{\Pi}_l^m v)(t_{n-1}^+) = 0, \quad \text{if } l = 1, \quad (\text{C.1a})$$

$$(v - \tilde{\Pi}_l^m v)(t_n^-) = 0, \quad (\text{C.1b})$$

$$\int_{I_n} (v - \tilde{\Pi}_l^m v) w dt = 0 \quad \forall w \in P_{m-l-1}(I_n). \quad (\text{C.1c})$$

Here, note that $P_{-1}(I_n, X)$ is interpreted as $\{0\}$ so that the variational condition (C.1c) drops out if $m = l$. For $v \in H^1(I_n, X) \subset C(\bar{I}_n, X)$ the projection can be equivalently defined by

$$(v - \tilde{\Pi}_l^m v)(t_n^-) = 0, \quad \text{if } l = 1, \quad (\text{C.2a})$$

$$\int_{I_n} \partial_t (v - \tilde{\Pi}_l^m v) w dt + \delta_{0,l} (v - \tilde{\Pi}_l^m v)(t_{n-1}^+) w(t_{n-1}^+) = 0 \quad \forall w \in P_{m-l}(I_n). \quad (\text{C.2b})$$

Furthermore, we formally set $(v - \tilde{\Pi}_l^m v)(t_0^-) = 0$.

Note that $\tilde{\Pi}_1^m$ ($l = 1$) is the standard projection used in the analysis of the Galerkin-Petrov time stepping, see for example [4, Section 4.1] and [11, Sections 2–4]. On the other hand, the projection $\tilde{\Pi}_0^m$ ($l = 0$) is the standard in the analysis of the discontinuous Galerkin time stepping method, see for example [52, Chapter 12, esp. Theorem 12.1] and [5, Section 3]. For a study of the well-definedness we refer to that of the more general operator in Definition C.10. \clubsuit

Remark C.7

In the case $l = 1$ we also could use in (C.2a)

$$(v - \tilde{\Pi}_l^m v)(t_{n-1}^+) = 0 \quad \text{instead of} \quad (v - \tilde{\Pi}_l^m v)(t_n^-) = 0,$$

which can be easily shown with the fundamental theorem of calculus. \clubsuit

Lemma C.8

Let $l \in \{0, 1\}$ and $m \in \mathbb{Z}$, $m \geq l$. Then, for $v \in H^1(I_n, X)$ the two definitions (C.1) and (C.2) of $\tilde{\Pi}_l^m v \in P_m(I_n, X)$ given in Definition C.6 are equivalent.

Proof. Let $v \in H^1(I_n, X)$.

(C.1) \Rightarrow (C.2): Obviously, (C.1b) implies that (C.2a) holds. It remains to prove (C.2b). From integration by parts we obtain for arbitrary $w \in P_{m-l}(I_n)$ that

$$\begin{aligned} & \int_{I_n} \partial_t (v - \tilde{\Pi}_l^m v) w dt + \delta_{0,l} (v - \tilde{\Pi}_l^m v)(t_{n-1}^+) w(t_{n-1}^+) \\ &= - \int_{I_n} (v - \tilde{\Pi}_l^m v) \partial_t w dt + (v - \tilde{\Pi}_l^m v)(t_n^-) w(t_n^-) - (1 - \delta_{0,l}) (v - \tilde{\Pi}_l^m v)(t_{n-1}^+) w(t_{n-1}^+). \end{aligned}$$

Now, the first two terms on the right-hand side vanish due to (C.1c) and (C.1b). The last term vanishes for $l = 0$ because of $1 - \delta_{0,l} = 0$ and for $l = 1$ because of (C.1a). Thus, (C.2b) holds.

(C.2) \Rightarrow (C.1): First of all, for $l = 1$ we gain from (C.2b) with $w \equiv 1$, the fundamental theorem of calculus, and (C.2a) that

$$0 = - \int_{I_n} \partial_t (v - \tilde{\Pi}_1^m v) dt = -(v - \tilde{\Pi}_1^m v) \Big|_{t_{n-1}^+}^{t_n^-} = (v - \tilde{\Pi}_1^m v)(t_{n-1}^+),$$

which is (C.1a). Furthermore, (C.1b) follows for $l = 1$ directly from (C.2a) and for $l = 0$ from (C.2b) again tested with $w \equiv 1$ after applying the fundamental theorem of calculus, respectively.

We now want to show (C.1c). From (C.2b) we obtain by integration by parts for all $\tilde{w} \in P_{m-l}(I_n)$ that

$$\begin{aligned} 0 &= \int_{I_n} \partial_t (v - \tilde{\Pi}_l^m v) \tilde{w} dt + \delta_{0,l} (v - \tilde{\Pi}_l^m v)(t_{n-1}^+) \tilde{w}(t_{n-1}^+) \\ &= - \int_{I_n} (v - \tilde{\Pi}_l^m v) \partial_t \tilde{w} dt + (v - \tilde{\Pi}_l^m v)(t_n^-) \tilde{w}(t_n^-) - (1 - \delta_{0,l}) (v - \tilde{\Pi}_l^m v)(t_{n-1}^+) \tilde{w}(t_{n-1}^+). \end{aligned}$$

The boundary term at t_n^- vanishes due to the already proven (C.1b), whereas the boundary term at t_{n-1}^+ vanishes because of the already proven (C.1a) if $l = 1$ or because of $1 - \delta_{0,l} = 0$ if $l = 0$, respectively. Hence, it remains to verify that for every $w \in P_{m-l-1}(I_n)$ we can choose a $\tilde{w} \in P_{m-l}(I_n)$ such that $w = \partial_t \tilde{w}$. But this holds if \tilde{w} is an antiderivative of w . \square

The following projections generalize the above operators such that also more general integrators, as e.g. quadrature formulas of sufficiently high degree of exactness, can be involved. Such projection operators are needed especially in Subsection 4.1.6 where specific quadrature rules are chosen in order to show supercloseness and superconvergence results. For simplicity, we restrict ourselves to integrators of the following form.

Assumption C.1

We assume that the integrator \mathcal{J}_n either represents the exact integral over I_n , i.e., $\mathcal{J}_n = \int_{I_n}$, (in which case $k_{\mathcal{J}} = -1$) or the application of a quadrature formula based on function and derivative values of the integrand in \bar{I}_n (in which case $k_{\mathcal{J}} \geq 0$ denotes the largest derivative order that is needed for \mathcal{J}_n).

In particular, Assumption C.1 yields that for all $g \in X'$ it holds

$$\langle g, \mathcal{J}_n[v] \rangle_{X',X} = \mathcal{J}_n \left[\langle g, v \rangle_{X',X} \right] \quad \forall v \in C^{k_{\mathcal{J}}}(\bar{I}_n, X), \quad (\text{C.3})$$

where we used properties of the Bochner integral (if $k_{\mathcal{J}} = -1$) or of (derivatives of) Banach space-valued functions (if $k_{\mathcal{J}} \geq 0$), also cf. Remark B.3.

Definition C.9

Let $m \in \mathbb{Z}$, $m \geq 0$. Furthermore, let $\mathcal{J}_n : C^{k_{\mathcal{J}}}(\bar{I}_n, X) \rightarrow X$ with $k_{\mathcal{J}} \geq -1$ be an integrator on \bar{I}_n that satisfies Assumption C.1 and integrates X -valued polynomials of maximal degree $2m$ exactly. Then, $\Pi_m^{\mathcal{J}} : C^{k_{\mathcal{J}}}(\bar{I}_n, X) \rightarrow P_m(I_n, X)$ is defined by

$$\mathcal{J}_n[(v - \Pi_m^{\mathcal{J}}v)w] = 0 \quad \forall w \in P_m(I_n),$$

i.e., as a generalization of the L^2 -projection to integrators beyond the (Bochner) integral over I_n .

Note that for $v \in P_m(I_n)$ we have, due to the exactness of \mathcal{J}_n , that

$$0 = \mathcal{J}_n[(v - \Pi_m^{\mathcal{J}}v)w] = \int_{I_n} (v - \Pi_m^{\mathcal{J}}v)w \, dt \quad \forall w \in P_m(I_n).$$

So, choosing $w = v - \Pi_m^{\mathcal{J}}v \in P_m(I_n)$, we find $0 = \int_{I_n} (v - \Pi_m^{\mathcal{J}}v)^2 \, dt$ and, thus, $\Pi_m^{\mathcal{J}}v = v$ for $v \in P_m(I_n)$. Hence, $\Pi_m^{\mathcal{J}}$ is a well-defined projection operator if $X = \mathbb{R}$. \clubsuit

Definition C.10 (Generalized dG/cGP projection)

Let $l \in \{0, 1\}$ and $m \in \mathbb{Z}$, $m \geq l$. Furthermore, let $\mathcal{J}_n : C^{k_{\mathcal{J}}}(\bar{I}_n, X) \rightarrow X$ with $k_{\mathcal{J}} \geq -1$ be an integrator on \bar{I}_n that satisfies Assumption C.1 and integrates X -valued polynomials of maximal degree $2m - l - 1$ exactly. Then, $\tilde{\Pi}_l^{m, \mathcal{J}} : H^1(I_n, X) \cap C^{k_{\mathcal{J}}+1}(\bar{I}_n, X) \rightarrow P_m(I_n, X)$ is defined by

$$\begin{aligned} (v - \tilde{\Pi}_l^{m, \mathcal{J}}v)(t_{n-1}^+) &= 0, \quad \text{if } l = 1, \\ \mathcal{J}_n[\partial_t(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)w] + \delta_{0,l}(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)(t_{n-1}^+)w(t_{n-1}^+) &= 0 \quad \forall w \in P_{m-l}(I_n). \end{aligned}$$

This is a generalization of the standard dG/cGP projection for cases where the integrator is not simply the (Bochner) integral over I_n . Note that the domain of definition is chosen as $H^1(I_n, X) \cap C^{k_{\mathcal{J}}+1}(\bar{I}_n, X)$ in order to guarantee that all expressions are well-defined in the case of exact integration ($k_{\mathcal{J}} = -1$) as well as in the case of approximate integration using quadrature formulas that may require function and derivative values in \bar{I}_n ($k_{\mathcal{J}} \geq 0$). From the definition it follows

$$\begin{aligned} (v - \tilde{\Pi}_l^{m, \mathcal{J}}v)(t_n^-) &= \int_{I_n} \partial_t(v - \tilde{\Pi}_l^{m, \mathcal{J}}v) \, dt + (v - \tilde{\Pi}_l^{m, \mathcal{J}}v)(t_{n-1}^+) \\ &= \int_{I_n} \partial_t v \, dt - \mathcal{J}_n[\partial_t v] + \mathcal{J}_n[\partial_t(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)] + \delta_{0,l}(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)(t_{n-1}^+) \\ &= \int_{I_n} \partial_t v \, dt - \mathcal{J}_n[\partial_t v], \end{aligned}$$

where the fundamental theorem of calculus and the linearity of \mathcal{J}_n were used.

In order to verify that $\tilde{\Pi}_l^{m, \mathcal{J}}$ is well-defined, it suffices to show that $v = \tilde{\Pi}_l^{m, \mathcal{J}}v$ for all $v \in P_m(I_n)$. First of all, using the exactness of \mathcal{J}_n , we obtain for $v \in P_m(I_n)$ that

$$\begin{aligned} \int_{I_n} \partial_t(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)w \, dt + \delta_{0,l}(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)(t_{n-1}^+)w(t_{n-1}^+) \\ = \mathcal{J}_n[\partial_t(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)w] + \delta_{0,l}(v - \tilde{\Pi}_l^{m, \mathcal{J}}v)(t_{n-1}^+)w(t_{n-1}^+) = 0 \quad \forall w \in P_{m-l}(I_n). \end{aligned}$$

Therefore, choosing $w = (\partial_t v - \partial_t \tilde{\Pi}_l^{m,\mathcal{J}} v)(t - t_{n-1})^{1-l} \in P_{m-l}(I_n)$, we gain that

$$\int_{I_n} (\partial_t v - \partial_t \tilde{\Pi}_l^{m,\mathcal{J}} v)^2 (t - t_{n-1})^{1-l} dt = 0 \quad \text{and, thus,} \quad \partial_t v = \partial_t \tilde{\Pi}_l^{m,\mathcal{J}} v.$$

So, because of $(v - \tilde{\Pi}_l^{m,\mathcal{J}} v)(t_{n-1}^+) = 0$, which obviously holds for $l = 1$ and follows by choosing $w \equiv 1$ for $l = 0$, we easily conclude that $v = \tilde{\Pi}_l^{m,\mathcal{J}} v$ for $v \in P_m(I_n)$. \clubsuit

Definition C.11 (Extended generalized dG/cGP projection)

Let $l \in \{0, 1\}$ and $m \in \mathbb{Z}$, $m \geq l$. Furthermore, let $\mathcal{J}_n : C^{k_{\mathcal{J}}}(\bar{I}_n, X) \rightarrow X$ with $k_{\mathcal{J}} \geq -1$ be an integrator on \bar{I}_n that satisfies Assumption C.1 and integrates X -valued polynomials of maximal degree $2m - l$ exactly. Then, $\tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} : H^1(I_n, X) \cap C^{k_{\mathcal{J}}+1}(\bar{I}_n, X) \rightarrow P_{m+1}(I_n, X)$ is defined by

$$\begin{aligned} (v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v)(t_{n-1}^+) &= 0, \quad \text{if } l = 1, \\ \mathcal{J}_n \left[\partial_t (v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v) w \right] + \delta_{0,l} (v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v)(t_{n-1}^+) w(t_{n-1}^+) &= 0 \quad \forall w \in P_{m-l}(I_n), \\ \int_{I_n} \partial_t (v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v) w dt + \delta_{0,l} (v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v)(t_{n-1}^+) w(t_{n-1}^+) &= 0 \quad \forall w \in \tilde{P}_{m-l+1}(I_n) \end{aligned}$$

where $\tilde{P}_{m-l+1}(I_n) := P_{m-l+1}(I_n) \setminus P_{m-l}(I_n)$. Note that in comparison to the definition of $\tilde{\Pi}_l^{m,\mathcal{J}}$ the assumption on the integrator is slightly stronger, an additional condition is added, and the operator now maps to $P_{m+1}(I_n, X)$. Therefore, it obviously holds $\tilde{\Pi}_l^{m,\mathcal{J}} v = \tilde{\Pi}_l^{m,\mathcal{J}} \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v$ for all $v \in H^1(I_n, X) \cap C^{k_{\mathcal{J}}+1}(\bar{I}_n, X)$ and we have

$$(v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v)(t_n^-) = \int_{I_n} \partial_t v dt - \mathcal{J}_n[\partial_t v].$$

The operator $\tilde{\Pi}_{l,*}^{m+1,\mathcal{J}}$ is well-defined, which can be shown similar as for $\tilde{\Pi}_l^{m,\mathcal{J}}$. Especially, note that we now even get for all $v \in P_{m+1}(I_n)$ that

$$\int_{I_n} \partial_t (v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v) w dt + \delta_{0,l} (v - \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v)(t_{n-1}^+) w(t_{n-1}^+) = 0 \quad \forall w \in P_{m-l+1}(I_n)$$

due to the additional condition. Choosing $w = (\partial_t v - \partial_t \tilde{\Pi}_{l,*}^{m+1,\mathcal{J}} v)(t - t_{n-1})^{1-l} \in P_{m-l+1}(I_n)$, we easily complete the argument as for $\tilde{\Pi}_l^{m,\mathcal{J}}$. \clubsuit

Definition C.12

Let $r, k \in \mathbb{Z}$, $0 \leq k \leq r$, and $\ell = \lfloor \frac{k}{2} \rfloor$. Furthermore, let $\mathcal{J}_n : C^{k_{\mathcal{J}}}(\bar{I}_n, X) \rightarrow X$ with $k_{\mathcal{J}} \geq -1$ be an integrator on \bar{I}_n that satisfies Assumption C.1 and integrates X -valued polynomials of maximal degree $2r - k - 1$ exactly. Then, $\bar{\Pi}_k^{r,\mathcal{J}} : H^{\ell+1}(I_n, X) \cap C^{k_{\mathcal{J}}+\ell+1}(\bar{I}_n, X) \rightarrow P_r(I_n, X)$ is defined by

$$\begin{aligned} (\bar{\Pi}_k^{r,\mathcal{J}} v)^{(j)}(t_n^-) &= v^{(j)}(t_n^-), & \text{for } j = 0, \dots, \ell - 1, \\ (\bar{\Pi}_k^{r,\mathcal{J}} v)^{(\ell)}(t) &= \tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}}(v^{(\ell)})(t), & \text{for all } t \in I_n. \end{aligned}$$

Note that by definition of $\ell = \lfloor \frac{k}{2} \rfloor \geq 0$ we always have $r - \ell \geq k - 2\ell \in \{0, 1\}$. Hence, $\tilde{\Pi}_{k-2\ell}^{r-\ell,\mathcal{J}}(v^{(\ell)})$ is well-defined according to Definition C.10.

The projection operator $\bar{\Pi}_k^{r,\mathcal{J}}$ is a further generalization of the generalized dG/cGP projection of Definition C.10. Of course, it holds $\bar{\Pi}_k^{r,\mathcal{J}} = \tilde{\Pi}_k^{r,\mathcal{J}}$ for $k \in \{0, 1\}$ and $r \geq k$. ♣

C.3 Some commutation properties

We already studied commutation properties for the abstract projection operators in Corollary B.5. Nevertheless, here we consider an important special case and one of its consequences. Also concrete proofs are given.

Lemma C.13

Let $m \in \mathbb{Z}$, $m \geq 0$, and let X be a Banach space. Suppose that \mathcal{J}_n is an integrator on \bar{I}_n that satisfies Assumption C.1 and integrates polynomials of maximal degree $2m$ exactly. Then, for all $g \in X'$ it holds

$$\langle g, \Pi_m^{\mathcal{J}} v \rangle_{X', X} = \Pi_m^{\mathcal{J}} \langle g, v \rangle_{X', X} \quad \forall v \in C^{k_{\mathcal{J}}}(\bar{I}_n, X).$$

Proof. Let $w \in P_m(I_n)$ be arbitrarily chosen. From the definition of $\Pi_m^{\mathcal{J}}$ it follows

$$\mathcal{J}_n \left[\Pi_m^{\mathcal{J}} (\langle g, v \rangle_{X', X}) w \right] = \mathcal{J}_n \left[\langle g, v \rangle_{X', X} w \right].$$

Further, the linearity of the duality pairing, (C.3), and again the definition of $\Pi_m^{\mathcal{J}}$ give

$$\mathcal{J}_n \left[\langle g, v - \Pi_m^{\mathcal{J}} v \rangle_{X', X} w \right] = \mathcal{J}_n \left[\langle g, (v - \Pi_m^{\mathcal{J}} v) w \rangle_{X', X} \right] = \left\langle g, \mathcal{J}_n \left[(v - \Pi_m^{\mathcal{J}} v) w \right] \right\rangle_{X', X} = 0.$$

Hence, altogether we have shown that

$$\mathcal{J}_n \left[\Pi_m^{\mathcal{J}} (\langle g, v \rangle_{X', X}) w \right] = \mathcal{J}_n \left[\langle g, \Pi_m^{\mathcal{J}} v \rangle_{X', X} w \right] \quad \forall w \in P_m(I_n).$$

Since both $\Pi_m^{\mathcal{J}} (\langle g, v \rangle_{X', X})$ and $\langle g, \Pi_m^{\mathcal{J}} v \rangle_{X', X}$ are in $P_m(I_n)$, the integrands on both sides of this equation are polynomials of maximal degree $2m$. Hence, \mathcal{J}_n can be replaced by \int_{I_n} . We then easily conclude the desired identity. \square

The proof of the following corollary exemplifies how to handle test functions that, unlike in the definition of the projection operator, are not real-valued polynomials but Banach space-valued polynomials.

Corollary C.14

Let $m \in \mathbb{Z}$, $m \geq 0$, and let X, Y be Banach spaces as well as $B(\cdot, \cdot) : X \times Y \rightarrow \mathbb{R}$ a continuous bilinear form. Suppose that \mathcal{J}_n is an integrator on \bar{I}_n that satisfies Assumption C.1 and integrates polynomials of maximal degree $2m$ exactly. Then, for $v \in C^{k_{\mathcal{J}}}(\bar{I}_n, X)$ it holds

$$\mathcal{J}_n \left[B(v, w) \right] = \mathcal{J}_n \left[B(\Pi_m^{\mathcal{J}} v, w) \right] \quad \forall w \in P_m(I_n, Y).$$

Proof. Let $w \in P_m(I_n, Y)$ be arbitrarily chosen. Then, there are $w_i \in Y$ and $p_i \in P_i(I_n)$, $i = 0, \dots, m$, such that $w(t) = \sum_{i=0}^m p_i(t)w_i$. Here, the p_i are typically chosen such that $p_i(t) = t^i$. Using the bilinearity of $B(\cdot, \cdot)$ and the linearity of the integrator, we obtain

$$\mathcal{J}_n[B(v - \Pi_m^\mathcal{J}v, w)] = \sum_{i=0}^m \mathcal{J}_n[B(v - \Pi_m^\mathcal{J}v, p_i w_i)] = \sum_{i=0}^m \mathcal{J}_n\left[\left(B(v, w_i) - B(\Pi_m^\mathcal{J}v, w_i)\right)p_i\right].$$

Applying Lemma C.13 with $g_i \in X'$, $i = 0, \dots, m$, defined by $\langle g_i, z \rangle_{X', X} := B(z, w_i)$ for all $z \in X$, it follows

$$\mathcal{J}_n\left[\left(B(v, w_i) - B(\Pi_m^\mathcal{J}v, w_i)\right)p_i\right] = \mathcal{J}_n\left[\left(B(v, w_i) - \Pi_m^\mathcal{J}B(v, w_i)\right)p_i\right] = 0,$$

where the definition of $\Pi_m^\mathcal{J}$ is used for the last step. Overall, this easily yields the desired statement. \square

C.4 Some stability results

We want to prove a stability result for the projection operator $\Pi_m^\mathcal{J}$, which is used in Section 4.1, see Remark 4.7, to keep the presentation simple. The assumptions on \mathcal{J}_n that occur in the following lemma are those known from (4.5) with $r = m$ and $k = 0$.

Lemma C.15

Let $m \in \mathbb{Z}$, $m \geq 0$, and let V be a Hilbert space. Suppose that \mathcal{J}_n is an integrator on \bar{I}_n that satisfies Assumption C.1 and integrates polynomials of maximal degree $2m$ exactly. Furthermore, let \mathcal{J}_n provide the monotonicity property $\mathcal{J}_n[v] \leq \mathcal{J}_n[w]$ if $v(t) \leq w(t)$ for all $t \in \bar{I}_n$ as well as the Cauchy–Schwarz-type inequality $\mathcal{J}_n[vw] \leq (\mathcal{J}_n[v^2])^{1/2} (\mathcal{J}_n[w^2])^{1/2}$, where we tacitly assume that for v and w all occurring expressions are well-defined. Then, it holds

$$\left(\mathcal{J}_n\left[\|\Pi_m^\mathcal{J}v(\cdot)\|_{V'}^2\right]\right)^{1/2} \leq \left(\mathcal{J}_n\left[\|v(\cdot)\|_{V'}^2\right]\right)^{1/2}$$

for all $v \in C^{k_\mathcal{J}}(\bar{I}_n, V')$.

Proof. Since V is a Hilbert space, also V' is a Hilbert space and its norm $\|\cdot\|_{V'}$ is induced by an inner product, say $(\cdot, \cdot)_{V'}$. Hence, we have

$$\mathcal{J}_n\left[\|\Pi_m^\mathcal{J}v(\cdot)\|_{V'}^2\right] = \mathcal{J}_n\left[(\Pi_m^\mathcal{J}v(\cdot), \Pi_m^\mathcal{J}v(\cdot))_{V'}\right] = \mathcal{J}_n\left[(v(\cdot), \Pi_m^\mathcal{J}v(\cdot))_{V'}\right], \quad (\text{C.4})$$

where in the last step Corollary C.14 was applied with $X = Y = V'$ and $B(\cdot, \cdot) = (\cdot, \cdot)_{V'}$.

Further, for all $t \in \bar{I}_n$ we get by Cauchy–Schwarz' inequality that

$$(v(t), \Pi_m^\mathcal{J}v(t))_{V'} \leq \|v(t)\|_{V'} \|\Pi_m^\mathcal{J}v(t)\|_{V'}.$$

Therefore, using the assumed properties of \mathcal{J}_n , we conclude

$$\mathcal{J}_n\left[(v(\cdot), \Pi_m^\mathcal{J}v(\cdot))_{V'}\right] \leq \mathcal{J}_n\left[\|v(\cdot)\|_{V'} \|\Pi_m^\mathcal{J}v(\cdot)\|_{V'}\right] \leq \left(\mathcal{J}_n\left[\|v(\cdot)\|_{V'}^2\right]\right)^{1/2} \left(\mathcal{J}_n\left[\|\Pi_m^\mathcal{J}v(\cdot)\|_{V'}^2\right]\right)^{1/2}.$$

So, combining this with (C.4) and dividing by $\left(\mathcal{J}_n\left[\|\Pi_m^\mathcal{J}v(\cdot)\|_{V'}^2\right]\right)^{1/2}$, we are done. \square

D Norm Equivalences for Hilbert Space-Valued Polynomials

In the error analysis of Chapter 4 norm equivalences for polynomial spaces are exploited at several places. However, while for real-valued polynomials of fixed maximal degree the equivalence of different norms follows immediately since the space is finite dimensional, for Hilbert space-valued polynomials the situation is not that clear. On the one hand, for infinite dimensional Hilbert spaces W also the space of W -valued polynomials of maximal degree, say $m \in \mathbb{Z}$, $m \geq 0$, is infinite dimensional. On the other hand, norm equivalence constants should ideally not depend on the specific Hilbert space.

However, for the two sorts of norm equivalences that were needed in our analysis, we show now that the norm equivalences for Hilbert space-valued polynomials hold with the same constants as their real-valued analogs.

Let $J \subset \mathbb{R}$ be an interval and X a Banach space. Then,

$$P_m(J, X) := \left\{ v \in C(J, X) : v(t) = \sum_{i=0}^m t^i v_i \text{ with } v_i \in X \right\}$$

defines the space of X -valued polynomials of maximal degree m .

In the following, let W denote a Hilbert space. Then, W possesses an orthonormal basis B , say $B = \{b_\alpha : \alpha \in A\}$, see [55, Theorem 3.10(a), p. 44]. Thus, for $v \in P_m(J, W)$, we have

$$v(t) = \sum_{i=0}^m t^i v_i = \sum_{i=0}^m t^i \left[\sum_{\alpha \in A} (v_i, b_\alpha)_W b_\alpha \right] = \sum_{\alpha \in A} \left[\sum_{i=0}^m t^i (v_i, b_\alpha)_W \right] b_\alpha = \sum_{\alpha \in A} g_\alpha(t) b_\alpha,$$

where $g_\alpha(t) = (v(t), b_\alpha)_W = \sum_{i=0}^m t^i (v_i, b_\alpha)_W \in P_m(J, \mathbb{R})$ for all $\alpha \in A$. Furthermore, by Parseval's identity it follows

$$\|v(t)\|_W^2 = \sum_{\alpha \in A} |(v(t), b_\alpha)_W|^2 = \sum_{\alpha \in A} |g_\alpha(t)|^2. \quad (\text{D.1})$$

Easily, we also get that

$$\|\partial_t^k v(t)\|_W^2 = \sum_{\alpha \in A} |(\partial_t^k v(t), b_\alpha)_W|^2 = \sum_{\alpha \in A} |\partial_t^k (v(t), b_\alpha)_W|^2 = \sum_{\alpha \in A} |g_\alpha^{(k)}(t)|^2 \quad (\text{D.2})$$

for $k \geq 0$.

In the following, we restrict ourselves to the study of norm equivalences for polynomials that are defined on an arbitrary mesh interval $I_n = (t_{n-1}, t_n]$.

D.1 Norm equivalence used for the cGP-like case

We start considering the norm equivalence used in Section 4.1, cf. Lemma 4.3.

Lemma D.1

Let $m \in \mathbb{Z}$, $m \geq 0$. The two mappings

$$\varphi \mapsto \left(\int_{I_n} |\varphi(t)|^2 dt \right)^{1/2}$$

and

$$\varphi \mapsto \left(\int_{I_n} |\Pi_{m-1}\varphi(t)|^2 dt + \left(\frac{\tau_n}{2}\right) |\varphi(t_n)|^2 \right)^{1/2}$$

define equivalent norms on $P_m(I_n, \mathbb{R})$ where the equivalence constants are independent of τ_n . The involved operator Π_{m-1} is the L^2 -projection onto $P_{m-1}(I_n, \mathbb{R})$, cf. Definition C.4. Further, we agree that in the case $m = 0$ we read $\Pi_{-1}\varphi \equiv 0$.

Proof. Using the affine transformation T_n , defined in (1.7), and associating to $\varphi \in P_m(I_n, \mathbb{R})$ the function $\hat{\varphi} \in P_m((-1, 1], \mathbb{R})$ given by $\hat{\varphi}(\hat{t}) := \varphi(T_n(\hat{t}))$, it suffices to prove that

$$\hat{\varphi} \mapsto \left(\int_{-1}^1 |\hat{\varphi}(\hat{t})|^2 d\hat{t} \right)^{1/2} \quad \text{and} \quad \hat{\varphi} \mapsto \left(\int_{-1}^1 |\hat{\Pi}_{m-1}\hat{\varphi}(\hat{t})|^2 d\hat{t} + |\hat{\varphi}(1)|^2 \right)^{1/2}$$

are equivalent norms on $P_m((-1, 1], \mathbb{R})$. Here, note that the $\frac{\tau_n}{2}$ factor in the local I_n version of the second mapping is due to the transformation.

Obviously, the first expression is a norm. Since $P_m((-1, 1], \mathbb{R})$ is finite dimensional all norms on this space are equivalent. So, it remains to prove that also the second expression is a norm. Obviously, it is a semi-norm. We need to show that for all $\hat{\varphi} \in P_m((-1, 1], \mathbb{R})$

$$\left(\int_{-1}^1 |\hat{\Pi}_{m-1}\hat{\varphi}(\hat{t})|^2 d\hat{t} + |\hat{\varphi}(1)|^2 \right)^{1/2} = 0 \quad \text{implies that} \quad \hat{\varphi} \equiv 0.$$

Here, when the expression on the left-hand side equals zero, also every single (non-negative) term needs to vanish.

From $\hat{\Pi}_{m-1}\hat{\varphi} \equiv 0$ it follows that $\hat{\varphi}$ is orthogonal to all polynomials of degree less than or equal to $m - 1$ with respect to the inner product in $L^2((-1, 1])$. Thus, because of $\hat{\varphi} \in P_m((-1, 1], \mathbb{R})$, we have that $\hat{\varphi}$ is a multiple of the m th Legendre polynomial $P_m^{(0,0)}$, i.e., there is a $c \in \mathbb{R}$ such that $\hat{\varphi}(\hat{t}) = cP_m^{(0,0)}(\hat{t})$. Since $P_m^{(0,0)}(1) \neq 0$, see (A.1), we then conclude from $0 = \hat{\varphi}(1) = cP_m^{(0,0)}(1)$ that $c = 0$. Hence, it holds $\hat{\varphi} \equiv 0$ and we are done. \square

Lemma D.2

Let $m \in \mathbb{Z}$, $m \geq 0$, and let W be a Hilbert space. Then, the mappings

$$v \mapsto \left(\int_{I_n} \|v(t)\|_W^2 dt \right)^{1/2}$$

and

$$v \mapsto \left(\int_{I_n} \|\Pi_{m-1}v(t)\|_W^2 dt + \left(\frac{\tau_n}{2}\right) \|v(t_n)\|_W^2 \right)^{1/2}$$

define equivalent norms on $P_m(I_n, W)$ where the equivalence constants are independent of τ_n and of the space W .

Proof. Let $v \in P_m(I_n, W)$ be arbitrarily chosen. Then, the polynomial v can be represented by $v(t) = \sum_{i=0}^m t^i v_i$ with $v_i \in W$. We define $\widetilde{W} := \text{span}\{v_0, v_1, \dots, v_m\} \subset W$. Equipped with the $\|\cdot\|_W$ -norm, \widetilde{W} is a finite dimensional Hilbert space. It, of course, has a orthonormal basis $\{b_1, \dots, b_d\}$, where $d \leq m+1$ is the dimension of \widetilde{W} .

By construction it holds $v \in P_m(I_n, \widetilde{W})$. Therefore, by (D.1) it follows

$$\int_{I_n} \|v(t)\|_W^2 dt = \int_{I_n} \sum_{j=1}^d |(v(t), b_j)_W|^2 dt = \sum_{j=1}^d \int_{I_n} |(v(t), b_j)_W|^2 dt$$

and similarly

$$\begin{aligned} \int_{I_n} \|\Pi_{m-1}v(t)\|_W^2 dt + \left(\frac{\tau_n}{2}\right) \|v(t_n)\|_W^2 &= \sum_{j=1}^d \left(\int_{I_n} |(\Pi_{m-1}v(t), b_j)_W|^2 dt + \left(\frac{\tau_n}{2}\right) |(v(t_n), b_j)_W|^2 \right) \\ &= \sum_{j=1}^d \left(\int_{I_n} |\Pi_{m-1}(v(t), b_j)_W|^2 dt + \left(\frac{\tau_n}{2}\right) |(v(t_n), b_j)_W|^2 \right). \end{aligned}$$

Here, in the last step, the projection operator Π_{m-1} can be pulled out of the inner product due to Lemma C.13 since for every $j = 1, \dots, d$ the expression $(\cdot, b_j)_W$ defines a function in W' , i.e., there is a $g_j \in W'$ such that $\langle g_j, w \rangle_{W', W} = (w, b_j)_W$ for all $w \in W$.

But for every $j = 1, \dots, d$ the function $t \mapsto (v(t), b_j)_W$ is in $P_m(I_n, \mathbb{R})$. Thus, from Lemma D.1 we have

$$\begin{aligned} C_1 \int_{I_n} |(v(t), b_j)_W|^2 dt \\ \leq \int_{I_n} |\Pi_{m-1}(v(t), b_j)_W|^2 dt + \left(\frac{\tau_n}{2}\right) |(v(t_n), b_j)_W|^2 \leq C_2 \int_{I_n} |(v(t), b_j)_W|^2 dt, \end{aligned}$$

where C_1 and C_2 do not depend on τ_n and b_j . Summing up over $j = 1, \dots, d$ and exploiting the identities proven above, we immediately get

$$C_1 \int_{I_n} \|v(t)\|_W^2 dt \leq \int_{I_n} \|\Pi_{m-1}v(t)\|_W^2 dt + \left(\frac{\tau_n}{2}\right) \|v(t_n)\|_W^2 \leq C_2 \int_{I_n} \|v(t)\|_W^2 dt.$$

Since the constants are independent of b_j , they are also independent of \widetilde{W} and v , respectively. So, since $v \in P_m(I_n, W)$ was arbitrarily chosen, we are done. \square

D.2 Norm equivalence used for final error estimate

We now study the norm equivalence needed in Section 4.3, cf. Lemma 4.52.

Lemma D.3

Let $m, l \in \mathbb{Z}$, $0 \leq l \leq m$. The two mappings

$$\varphi \mapsto \left(\int_{I_n} |\varphi(t)|^2 dt \right)^{1/2}$$

and

$$\varphi \mapsto \left(\left(\frac{\tau_n}{2} \right)^{2l} \int_{I_n} |\varphi^{(l)}(t)|^2 dt + \sum_{i=0}^{l-1} \left(\frac{\tau_n}{2} \right)^{2i+1} |\varphi^{(i)}(t_n^-)|^2 \right)^{1/2}$$

define equivalent norms on $P_m(I_n, \mathbb{R})$ where the equivalence constants are independent of τ_n .

Proof. Using the affine transformation T_n , defined in (1.7), and associating to $\varphi \in P_m(I_n, \mathbb{R})$ the function $\hat{\varphi} \in P_m((-1, 1], \mathbb{R})$ given by $\hat{\varphi}(\hat{t}) := \varphi(T_n(\hat{t}))$, it suffices to prove that

$$\hat{\varphi} \mapsto \left(\int_{-1}^1 |\hat{\varphi}(\hat{t})|^2 d\hat{t} \right)^{1/2} \quad \text{and} \quad \hat{\varphi} \mapsto \left(\int_{-1}^1 |\hat{\varphi}^{(l)}(\hat{t})|^2 d\hat{t} + \sum_{i=0}^{l-1} |\hat{\varphi}^{(i)}(1^-)|^2 \right)^{1/2}$$

are equivalent norms on $P_m((-1, 1], \mathbb{R})$. Here, note that the $(\frac{\tau_n}{2})^{2l}$ and $(\frac{\tau_n}{2})^{2i+1}$ factors in the local I_n version of the second mapping are due to the transformation.

Obviously, the first expression is a norm. Since $P_m((-1, 1], \mathbb{R})$ is finite dimensional all norms on this space are equivalent. So it remains to prove that also the second expression is a norm. Obviously, it is a semi-norm. We need to show that for all $\hat{\varphi} \in P_m((-1, 1], \mathbb{R})$

$$\left(\int_{-1}^1 |\hat{\varphi}^{(l)}(\hat{t})|^2 d\hat{t} + \sum_{i=0}^{l-1} |\hat{\varphi}^{(i)}(1^-)|^2 \right)^{1/2} = 0 \quad \text{implies that} \quad \hat{\varphi} \equiv 0.$$

Here, when the expression on the left-hand side equals zero, also every single (non-negative) term needs to vanish.

Owing to $\hat{\varphi}^{(l)} \equiv 0$, we have that $\hat{\varphi}^{(l-1)}$ is constant. Combining this with $\hat{\varphi}^{(l-1)}(1^-) = 0$, it follows $\hat{\varphi}^{(l-1)} \equiv 0$. Then, because of $\hat{\varphi}^{(i)}(1^-) = 0$ for all $i = 0, \dots, l-1$, we recursively conclude $\hat{\varphi} \equiv 0$. \square

Lemma D.4

Let $m, l \in \mathbb{Z}$, $0 \leq l \leq m$, and let W be a Hilbert space. Then, the mappings

$$v \mapsto \left(\int_{I_n} \|v(t)\|_W^2 dt \right)^{1/2}$$

and

$$v \mapsto \left(\left(\frac{\tau_n}{2} \right)^{2l} \int_{I_n} \|v^{(l)}(t)\|_W^2 dt + \sum_{i=0}^{l-1} \left(\frac{\tau_n}{2} \right)^{2i+1} \|v^{(i)}(t_n^-)\|_W^2 \right)^{1/2}$$

define equivalent norms on $P_m(I_n, W)$ where the equivalence constants are independent of τ_n and of the space W .

Proof. Let $v \in P_m(I_n, W)$ be arbitrarily chosen. Then, the polynomial v can be represented by $v(t) = \sum_{i=0}^m t^i v_i$ with $v_i \in W$. We define $\widetilde{W} := \text{span}\{v_0, v_1, \dots, v_m\} \subset W$. Equipped with the $\|\cdot\|_W$ -norm, \widetilde{W} is a finite dimensional Hilbert space. It, of course, has a orthonormal basis $\{b_1, \dots, b_d\}$, where $d \leq m+1$ is the dimension of \widetilde{W} .

Therefore, since by construction $v \in P_m(I_n, \widetilde{W})$, we obtain by Parseval's identity, cf. (D.1), that

$$\int_{I_n} \|v(t)\|_W^2 dt = \int_{I_n} \sum_{j=1}^d |(v(t), b_j)_W|^2 dt = \sum_{j=1}^d \int_{I_n} |(v(t), b_j)_W|^2 dt$$

and similarly, cf. (D.2), that

$$\begin{aligned} & \left(\frac{\tau_n}{2}\right)^{2l} \int_{I_n} \|v^{(l)}(t)\|_W^2 dt + \sum_{i=0}^{l-1} \left(\frac{\tau_n}{2}\right)^{2i+1} \|v^{(i)}(t_n^-)\|_W^2 \\ &= \sum_{j=1}^d \left(\left(\frac{\tau_n}{2}\right)^{2l} \int_{I_n} |\partial_t^l (v, b_j)_W(t)|^2 dt + \sum_{i=0}^{l-1} \left(\frac{\tau_n}{2}\right)^{2i+1} |\partial_t^i (v, b_j)_W(t_n^-)|^2 \right). \end{aligned}$$

But for every $j = 1, \dots, d$ the function $t \mapsto (v(t), b_j)_W$ is in $P_m(I_n, \mathbb{R})$. Thus, from Lemma D.3 we have

$$\begin{aligned} & C_1 \int_{I_n} |(v(t), b_j)_W|^2 dt \\ & \leq \left(\frac{\tau_n}{2}\right)^{2l} \int_{I_n} |\partial_t^l (v, b_j)_W(t)|^2 dt + \sum_{i=0}^{l-1} \left(\frac{\tau_n}{2}\right)^{2i+1} |\partial_t^i (v, b_j)_W(t_n^-)|^2 \leq C_2 \int_{I_n} |(v(t), b_j)_W|^2 dt, \end{aligned}$$

where C_1 and C_2 do not depend on τ_n and b_j . Summing up over $j = 1, \dots, d$ and exploiting the identities proven above, we immediately get

$$C_1 \int_{I_n} \|v(t)\|_W^2 dt \leq \left(\frac{\tau_n}{2}\right)^{2l} \int_{I_n} \|v^{(l)}(t)\|_W^2 dt + \sum_{i=0}^{l-1} \left(\frac{\tau_n}{2}\right)^{2i+1} \|v^{(i)}(t_n^-)\|_W^2 \leq C_2 \int_{I_n} \|v(t)\|_W^2 dt.$$

Since the constants are independent of b_j , they are also independent of \widetilde{W} and v , respectively. So, since $v \in P_m(I_n, W)$ was arbitrarily chosen, we are done. \square

Bibliography

- [1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. Department of Commerce, National Bureau of Standards, 1964. 10th Printing with corrections 1972.
- [2] V. Adolfsson. L^2 -integrability of second-order derivatives for Poisson’s equation in nonsmooth domains. *Math. Scand.*, 70(1):146–160, 1992.
- [3] N. Ahmed and V. John. Adaptive time step control for higher order variational time discretizations applied to convection-diffusion-reaction equations. *Comput. Methods Appl. Mech. Engrg.*, 285:83–101, 2015.
- [4] N. Ahmed and G. Matthies. Higher order continuous Galerkin–Petrov time stepping schemes for transient convection-diffusion-reaction equations. *ESAIM: M2AN*, 49(5):1429–1450, 2015.
- [5] G. Akrivis and C. Makridakis. Galerkin time-stepping methods for nonlinear parabolic equations. *ESAIM: M2AN*, 38(2):261–289, 2004.
- [6] G. Akrivis, C. Makridakis, and R. H. Nochetto. Galerkin and Runge–Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. *Numer. Math.*, 118(3):429–456, 2011.
- [7] I. Alonso-Mallo. Runge–Kutta methods without order reduction for linear initial boundary value problems. *Numer. Math.*, 91(4):577–603, 2002.
- [8] I. Alonso-Mallo and B. Cano. Avoiding order reduction of Runge–Kutta discretizations for linear time-dependent parabolic problems. *BIT*, 44(1):1–20, 2004.
- [9] M. Anselmann, M. Bause, S. Becher, and G. Matthies. Galerkin–collocation approximation in time for the wave equation and its post-processing. *ESAIM: M2AN*, 54(6):2099–2123, 2020.
- [10] W. Arendt and M. Kreuter. Mapping theorems for Sobolev spaces of vector-valued functions. *Studia Math.*, 240(3):275–299, 2018.
- [11] A. K. Aziz and P. Monk. Continuous finite elements in space and time for the heat equation. *Math. Comp.*, 52(186):255–274, 1989.
- [12] M. Bause, U. Köcher, F. A. Radu, and F. Schieweck. Post-processed Galerkin approximation of improved order for wave equations. *Math. Comp.*, 89(322):595–627, 2020.

- [13] S. Becher and G. Matthies. Unified analysis for variational time discretizations of higher order and higher regularity applied to non-stiff ODEs, 2021. Preprint arXiv:2105.06862v1.
- [14] S. Becher and G. Matthies. Variational time discretizations of higher order and higher regularity. *BIT*, 61(3):721–755, 2021.
- [15] S. Becher and G. Matthies. Variational time discretizations of higher order and higher regularity, 2021. Preprint arXiv:2003.04056v2.
- [16] S. Becher and G. Matthies. Unified analysis for variational time discretizations of higher order and higher regularity applied to non-stiff ODEs. *Numer. Algorithms*, 89(4):1533–1565, 2022.
- [17] S. Becher, G. Matthies, and D. Wenzel. Variational Methods for Stable Time Discretization of First-Order Differential Equations. In K. Georgiev, M. Todorov, and I. Georgiev, editors, *Advanced Computing in Industrial Mathematics: BGSIAM 2017*, volume 793 of *Studies in Computational Intelligence*, pages 63–75, Cham, 2019. Springer.
- [18] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. Julia: A fresh approach to numerical computing. *SIAM Rev.*, 59(1):65–98, 2017.
- [19] K. Burrage, W. H. Hundsdorfer, and J. G. Verwer. A study of B-convergence of Runge–Kutta methods. *Computing*, 36(1–2):17–34, 1986.
- [20] M. Calvo, S. González-Pinto, and J. I. Montijano. Runge–Kutta methods for the numerical solution of stiff semilinear systems. *BIT*, 40(4):611–639, 2000.
- [21] P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 40 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 2002.
- [22] M. C. Delfour and F. Dubeau. Discontinuous polynomial approximations in the theory of one-step, hybrid and multistep methods for nonlinear ordinary differential equations. *Math. Comp.*, 47(175):169–189, 1986.
- [23] E. Emmrich. Discrete versions of Gronwall’s lemma and their application to the numerical analysis of parabolic problems. Preprint 637-1999, Preprint series of the Institute of Mathematics, Technische Universität Berlin, 1999.
- [24] L. H. Encinas and J. M. Masqué. A short proof of the generalized Faà di Bruno’s formula. *Appl. Math. Lett.*, 16(6):975–979, 2003.
- [25] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
- [26] A. Ern and J.-L. Guermond. *Finite Elements III: First-Order and Time-Dependent PDEs*, volume 74 of *Texts in Applied Mathematics*. Springer, Cham, 2021.

- [27] A. Ern and F. Schieweck. Discontinuous Galerkin method in time combined with a stabilized finite element method in space for linear first-order PDEs. *Math. Comp.*, 85(301):2099–2129, 2016.
- [28] D. Estep. A posteriori error bounds and global error control for approximation of ordinary differential equations. *SIAM J. Numer. Anal.*, 32(1):1–48, 1995.
- [29] R. Frank, J. Schneid, and C. W. Ueberhuber. The concept of B-convergence. *SIAM J. Numer. Anal.*, 18(5):753–780, 1981.
- [30] F. C. Gao and M. J. Lai. A new H^2 regularity condition of the solution to Dirichlet problem of the Poisson equation and its applications. *Acta Math. Sin. (Engl. Ser.)*, 36(1):21–39, 2020.
- [31] K. R. Garren. Bounds for the eigenvalues of a matrix. Technical Report NASA-TN-D-4373, NASA Langley Research Center, Hampton, VA, 1968.
- [32] W. Gautschi. *Orthogonal Polynomials: Computation and Approximation*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2004.
- [33] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*, volume 69 of *Classics in Applied Mathematics*. SIAM, Philadelphia, 2011.
- [34] Ch. Grossmann, H.-G. Roos, and M. Stynes. *Numerical Treatment of Partial Differential Equations*. Springer-Verlag, Berlin, 2007.
- [35] W. Hackbusch. *Elliptic Differential Equations: Theory and Numerical Treatment*, volume 18 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 2017.
- [36] E. Hairer, G. Bader, and Ch. Lubich. On the stability of semi-implicit methods for ordinary differential equations. *BIT*, 22(2):211–232, 1982.
- [37] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 1993. Corrected 3rd printing 2008.
- [38] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, volume 14 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 1996. First softcover printing 2010.
- [39] P. Henrici. *Discrete variable methods in ordinary differential equations*. Wiley, New York, 1962.
- [40] B. L. Hulme. Discrete Galerkin and related one-step methods for ordinary differential equations. *Math. Comp.*, 26(120):881–891, 1972.
- [41] B. L. Hulme. One-step piecewise polynomial Galerkin methods for initial value problems. *Math. Comp.*, 26(118):415–426, 1972.

- [42] W. H. Hundsdorfer. *The numerical solution of nonlinear stiff initial value problems: an analysis of one step methods*, volume 12 of *CWI Tracts*. Centre for Mathematics and Computer Science, Amsterdam, 1985.
- [43] T. Hytönen, J. van Neerven, M. Veraar, and L. Weis. *Analysis in Banach spaces. Volume I: Martingales and Littlewood-Paley theory*, volume 63 of *Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge / A Series of Modern Surveys in Mathematics*. Springer, Cham, 2016.
- [44] H. Joulak and B. Beckermann. On Gautschi’s conjecture for generalized Gauss–Radau and Gauss–Lobatto formulae. *J. Comput. Appl. Math.*, 233(3):768–774, 2009.
- [45] C. Makridakis and R. H. Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numer. Math.*, 104(4):489–514, 2006.
- [46] G. Matthies and F. Schieweck. Higher order variational time discretizations for nonlinear systems of ordinary differential equations. Preprint 23/2011, Fakultät für Mathematik, Otto-von-Guericke-Universität Magdeburg, 2011.
[https://www.math.ovgu.de/Forschung/Veröffentlichungen/Preprints_Technical+Reports+\(alte+Version\)/Preprints/2011/11_23.html](https://www.math.ovgu.de/Forschung/Veröffentlichungen/Preprints_Technical+Reports+(alte+Version)/Preprints/2011/11_23.html).
- [47] R. L. Mishkov. Generalization of the formula of Faà di Bruno for a composite function with a vector argument. *Internat. J. Math. & Math. Sci.*, 24(7):481–491, 2000.
- [48] G. Petrova. Generalized Gauss–Radau and Gauss–Lobatto formulas with Jacobi weight functions. *BIT*, 57(1):191–206, 2017.
- [49] A. Prothero and A. Robinson. On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Math. Comp.*, 28(125):145–162, 1974.
- [50] B. Schweizer. *Partielle Differentialgleichungen: Eine anwendungsorientierte Einführung*. Springer-Verlag, Berlin, 2nd edition, 2018.
- [51] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*, volume 12 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 3rd edition, 2002.
- [52] V. Thomée. *Galerkin Finite Element Methods for Parabolic Problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2nd edition, 2006.
- [53] M. Vlasák and F. Roskovec. On Runge–Kutta, collocation and discontinuous Galerkin methods: Mutual connections and resulting consequences to the analysis. In *Programs and Algorithms of Numerical Mathematics 17*, pages 231–236, Prague, 2015. Institute of Mathematics AS CR.
- [54] I. Voulis. *A space-time approach to two-phase stokes flow: well-posedness and discretization*. Dissertation, RWTH Aachen University, 2019. <https://doi.org/10.18154/RWTH-2019-04874>.

- [55] J. Weidmann. *Linear Operators in Hilbert Spaces*, volume 68 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1980.
- [56] J. Wloka. *Partielle Differentialgleichungen: Sobolevräume und Randwertaufgaben*. Teubner, Stuttgart, 1st edition, 1982.
- [57] E. Zeidler. *Nonlinear Functional Analysis and its Applications I: Fixed-Point Theorems*. Springer-Verlag, New York, 1986.

Confirmation

- 1) I herewith declare that I have produced this paper without the prohibited assistance of third parties and without making use of aids other than those specified; notions taken over directly or indirectly from other sources have been identified as such. This paper has not previously been presented in identical or similar form to any other German or foreign examination board.
- 2) The present thesis has been produced at the Institute of Numerical Mathematics, Faculty of Mathematics, School of Science, TU Dresden under scientific supervision of Prof. Dr. Gunar Matthies.
- 3) There have been no prior attempts to obtain a doctoral degree at any university.
- 4) I recognize the Doctorate Regulations (Promotionsordnung) of the School of Science of the TU Dresden dated 23rd February 2011, last amended by resolutions of the Faculty Council dated 15th June 2011 and 18th June 2014 as well as of the School Committee dated 23rd May 2018.

Versicherung

- 1) Hiermit versichere ich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde vorgelegt.
- 2) Die vorliegende Arbeit wurde am Institut für Numerische Mathematik, Fakultät Mathematik, Bereich Mathematik und Naturwissenschaften, TU Dresden unter wissenschaftlicher Betreuung von Prof. Dr. Gunar Matthies angefertigt.
- 3) Es wurden zuvor keine Promotionsvorhaben unternommen.
- 4) Ich erkenne die Promotionsordnung des Bereichs Mathematik und Naturwissenschaften der TU Dresden vom 23.02.2011, zuletzt geändert durch Beschlüsse des Fakultätsrates vom 15.06.2011 und 18.06.2014 sowie des Bereichsrates vom 23.05.2018, an.

Dresden, 19.05.2022