INAUGURAL–DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich–Mathematischen Gesamtfakultät

der

RUPRECHT–KARLS–UNIVERSITÄT

HEIDELBERG

vorgelegt von

Dipl.-Math. Dörte Beigel

aus Gießen

Tag der mündlichen Prüfung

13. November 2012

Efficient goal-oriented global error estimation for BDF-type methods using discrete adjoints

Gutachter: Prof. Dr. Dr. h.c. Hans Georg Bock

Prof. Dr. Dr. h.c. Rolf Rannacher

Zusammenfassung

Die vorliegende Arbeit entwickelt Techniken zur Schätzung des globalen Fehlers, der bei der näherungsweisen Bestimmung von Lösungen von Anfangswertaufgaben (engl. *Initial Value Problems*, kurz IVPs) auf gegebenen Intervallen mit Mehrschrittverfahren basierend auf Rückwärtsdifferenzenformeln (engl. *Backward Differentiation Formulas*, kurz BDF) entsteht. Es werden dazu diskrete Adjungierte benutzt, die durch adjungierte Interne Numerische Differentiation (IND) des nominellen Integrationsschemas gewonnen werden. Zu diesem Zweck wird mit Hilfe einer neuen funktional-analytischen Formulierung die Brücke zwischen BDF-Verfahren und Petrov-Galerkin Finite-Elemente (FE)-Verfahren geschlagen. In Analogie zur Methodik der dual-gewichteten Residuen (engl. *Dual Weighted Residuals*) bei Galerkin-Verfahren für partielle Differentialgleichungen werden zielorientierte globale Fehlerschätzer entwickelt. Ihr asymptotisches Verhalten, ihre Genauigkeit bei BDF-Verfahren mit variabler Ordnung und Schrittweite sowie ihre Anwendbarkeit zur globalen Fehlersteuerung werden untersucht.

Die neuen Ergebnisse dieser Arbeit umfassen

- eine funktional-analytische Formulierung von IVPs bei gewöhnlichen Differentialgleichungen (engl. *Ordinary Differential Equations*, kurz ODEs) im Banachraum der stetig differenzierbaren Funktionen. Diese wird benötigt, da die klassische Hilbertraum-Formulierung es nicht erlaubt den Zusammenhang zwischen den diskreten Werten des adjungierten IND-Schemas und der Lösung des adjungierten IVP zu untersuchen. Die neue Formulierung führt zur Definition von schwachen Lösungen von adjungierten IVPs.

- eine Petrov-Galerkin FE-Diskretisierung der Funktionenräume, die es erlaubt die Banachraum-Formulierungen des IVP und seines Adjungierten in endlichdimensionale Probleme zu überführen. Es wird die Äquivalenz dieser endlichdimensionalen Probleme zu BDF-Verfahren mit variabler, aber vorgegebener Ordnung und Schrittweite und ihren adjungierten IND-Schemata gezeigt. Somit wird die FE-Näherung der schwachen Adjungierten aus den diskreten Werten des adjungierten IND-Schemas bestimmt und Diskretisierung und Differentiation kommutieren in der entwickelten Formulierung.

- einen Beweis dafür, dass die Werte des adjungierten IND-Schemas eines BDF-Verfahrens mit konstanter Ordnung und Schrittweite auf dem offenen Intervall gegen die Lösung des adjungierten IVP konvergieren. Des Weiteren wird ein Beweis dafür gegeben, dass die adjungierte FE-Näherung auf dem gesamten Intervall gegen die schwache Lösung des adjungierten IVP konvergiert.

- zielorientierte globale Fehlerschätzer für BDF-Verfahren, welche für jeden Integrationsschritt eine lokale Fehlergröße mit dem entsprechenden Wert des adjungierten IND-Schemas gewichten und in Summe eine genaue und effiziente Schätzung des tatsächlichen Fehlers liefern. Als lokale Fehlergröße kommen Defekt-Integrale und lokale Abschneidefehler zum Einsatz.

- Strategien zur zielorientierten globalen Fehlersteuerung für BDF-Verfahren, die entweder die lokal wirkende relative Toleranz oder – mit Hilfe der schrittweisen Fehlerindikatoren – das vorhandene Integrationsschema anpassen.

- ein ODE-Modell einer exothermen, selbst-beschleunigenden chemischen Reaktion mit Stoffübertragung, die in einem diskontinuierlichen Rührkessel durchgeführt wird. Mit Hilfe dieses Anwendungsbeispieles aus dem Chemieingenieurwesen werden Verwendbarkeit und Zuverlässigkeit der neuen Techniken zur näherungsweisen Bestimmung von schwachen Adjungierten und zur Simulation mit zielorientierter globaler Fehlersteuerung gezeigt.

Abstract

This thesis develops estimation techniques for the global error that occurs during the approximation of solutions of Initial Value Problems (IVPs) on given intervals by multistep integration methods based on Backward Differentiation Formulas (BDF). To this end, discrete adjoints obtained by adjoint Internal Numerical Differentiation (IND) of the nominal integration scheme are used. For this purpose, a bridge between BDF methods and Petrov-Galerkin Finite Element (FE) methods is built by a novel functional-analytic framework. Goal-oriented global error estimators are derived in analogy to the Dual Weighted Residual methodology in Galerkin methods for Partial Differential Equations. Their asymptotic behavior, their accuracy in BDF methods with variable order and stepsize as well as their applicability for global error control are investigated.

The novel results presented in this thesis include

- a functional-analytic framework for IVPs in Ordinary Differential Equations (ODEs) in the Banach space of continuously differentiable functions. This framework is needed since the classical Hilbert space setting is not suitable to analyze the relation between the discrete values of the adjoint IND scheme and the solution of the adjoint IVP. The new framework gives rise to the definition of weak solutions of adjoint IVPs.

- a Petrov-Galerkin FE discretization of the function spaces that allows to transform the variational formulations of the IVP and of its adjoint IVP into finite dimensional problems. The equivalence of these finite dimensional problems to BDF methods with variable but prescribed order and stepsize and their adjoint IND schemes is shown. Thus, the FE approximation of the weak adjoint is determined by the discrete values of the adjoint IND scheme and discretization and differentiation commute in the developed framework.

- a proof that the values of the adjoint IND scheme corresponding to a BDF method with constant order and stepsize converge to the solution of the adjoint IVP on the open interval. In addition, a proof is given that demonstrates the convergence of the FE approximation to the weak solution of the adjoint IVP on the entire interval.

- goal-oriented global error estimators for BDF methods that weight, for each integration step, a local error quantity with the corresponding value of the adjoint IND scheme and yields in sum an accurate and efficient estimate for the actual error. As local error quantity defect integrals and local truncation errors are employed, respectively.

- strategies for goal-oriented global error control in BDF methods that either adapt the locally acting relative tolerance or the given integration scheme using the stepwise error indicators.
- an ODE model of an exothermic, self-accelerating chemical reaction with mass transfer carried out in a discontinuous Stirred Tank Reactor. With this real-world example from chemical engineering the applicability and reliability of the novel techniques for the approximation of weak adjoints and for the simulation with goal-oriented global error control are shown.

Contents

Introduction

Dynamic processes are of great importance in numerous fields of scientific research such as engineering, physics, chemistry, biology, medicine, and economics. However, they appear not only in research but also in our daily life where we are surrounded by dynamic processes – for example, a flowing river or a bobsled in an ice channel. Mathematically, we can model dynamic processes by differential equations: The rate by which the state of a process changes over time is a function of the state itself, referred to as model function. Time is an independent variable whereas the state is a dependent variable defined by the differential equation.

To numerically solve a complicated differential equation which models a dynamic process, the continuous time interval is discretized using a finite number of time points. The differential equation is then solved at each of these points to give approximations of the state of the dynamic process. This procedure is called numerical integration. A typical integration method chooses the distance between the time points adaptively to keep an error quantity on the computed approximations small while the computational effort remains limited. However, the same error magnitude at different time points may have different effects on the evolution of the process state.

Imagine two situations: While a bobsledder is riding in a straight section of the channel, he suddenly rides over a small bump in the ice. This has nearly no effect on his arrival time at the end of the channel. Whereas if he is passing over from a curve to a straight section and suddenly hits a bump, he swerves and slows down which significantly influences his arrival time at the end. Mathematically, we can measure the effect of small intermediate changes – such as the bump which constitutes a change in the properties of the iced surface – on the final state by adjoint sensitivities. They are given as solution of an auxiliary adjoint differential equation. In the first situation, the adjoint sensitivity and hence the effect on the arrival time of the bobsledder is small, whereas in the second situation the adjoint sensitivity is huge and hence the small change caused by the bump shows a huge effect on the bobsledder's arrival time. These effects also occur in numerical integration and lead to different propagations of small local errors arising from discretization. Numerical integration methods usually estimate and control local errors, whereas global errors of computed approximations are the crucial quantities that should be estimated and controlled.

This doctoral thesis is devoted to the theoretical interpretation of adjoint information provided by differentiation of multistep integration methods and to the

development of global error estimators that can be used to control efficiency and accuracy in the solution of differential equations by multistep methods.

In the following we briefly indicate the significance of numerical integration and sensitivity generation in the field of applied mathematics and point out their current state of the art.

Simulation by integration methods

Our focus lies on Ordinary Differential Equations (ODEs) with initial conditions describing the process states at the starting time. They are called Initial Value Problems (IVPs) in ODEs. If the IVP is stiff and in particular stiff with a model function that is expensive to evaluate, the linear multistep Backward Differentiation Formula (BDF) method is the integrator of choice. In each integration step the BDF method reuses past approximations from former integration steps and evaluates the model function only once at the end of the current step. This yields an implicit equation which is then solved by efficient Newton-type methods, see e.g. Eich [53], Bauer [16], Shampine [109] and Brenan et al. [38].

Generally, IVPs also occur during the solution of Boundary Value Problems (BVPs) by shooting methods, see Osborne [100], Bulirsch [40] and Ascher and Petzold [8], and during the solution of instationary Partial Differential Equations (PDEs) by spatial discretization using a method of lines approach, see e.g. Ern and Guermond [57] and LeVeque [85]. They also appear as subtasks in the solution of parameter estimation problems with ODE constraints by multiple shooting, see Bock [28, 30], and in the solution of Optimal Control Problems (OCPs), see below.

Sensitivity generation

For a sufficiently smooth nominal solution, adjoint sensitivities are given by an adjoint IVP along the nominal solution. However, solving the adjoint IVP by integration just like the nominal IVP leads to a tremendous computational effort due to adaptive stepsize selection as well as to the appearance of non-differentiabilities. Instead, Bock's Internal Numerical Differentiation (IND) of the integration scheme used for the nominal IVP should be employed, see pioneering work of Bock [28, 30] as well as realizations for BDF methods by Bauer [16] and Albersmeyer [3]. Using (adjoint) IND means to differentiate (in adjoint mode) the integration scheme while keeping adaptive components fixed. This procedure gives the exact discrete adjoint sensitivities of the computed discrete IVP approximation at a computational cost directly related to the number of steps of the nominal integration scheme. Moreover, for one-step integration methods these discrete adjoints also approximate the exact continuous solution of the adjoint problem. Unfortunately, due to the use of past approximations in multistep integrators their adjoint IND schemes apparently do not provide approximations to the continuous adjoint solutions as recently discovered by Albersmeyer [3] and Sandu [106].

Error estimation in multistep integrators

So far, practical implementations of integration methods for IVPs typically use stepsize and further adaptive components to control only local error quantities for efficient integration, see Shampine [110]. But actually, the global error describes the quality of the computed approximation and hence should be controlled during integration. However, estimation techniques for global errors are still under development, see, for example, Johnson [77], Estep [58], Eriksson et al. [55], Böttcher and Rannacher [36], Moon et al. [96], Cao and Petzold [43], Lang and Verwer [84] as well as Tran and Berzins [118]. The crucial point is that global error estimation techniques require adjoint sensitivity information which could either be provided by solving adjoint IVPs or by applying adjoint IND to nominal integration schemes.

Optimal control

Several numerical approaches to solve OCPs involve the solution of differential equations as subtasks. In particular, Direct Single Shooting as well as Direct Multiple Shooting, proposed by Bock and Plitt [33], transform the OCP to a Nonlinear Program (NLP) using a control discretization and employ state-of-the-art integrators to solve IVPs on the shooting subintervals. Furthermore, nominal and adjoint IVPs have to be solved also in the solution of OCPs by indirect methods based on Pontryagin's Maximum Principle. Details can be found, for example, in Bock [26, 27, 29], Binder et al. [24], Gerdts [64] and Betts [23].

In particular, Direct Multiple Shooting together with IND has been successfully used to treat various problem classes involving OCPs such as optimum experimental design (see Bauer et al. [17] and Körkel et al. [80]), robust dynamic optimization (see Körkel et al. [80] and Diehl et al. [49]), nonlinear model predictive control (see Diehl [48] and Diehl et al. [50]), multi-level iterations (see Bock et al. [32], Albersmeyer et al. [4] and Kirches et al. [78]) as well as to treat OCPs in PDEs (see Schäfer [107] and Potschka [101]). In this solution approach the underlying dynamic process has to be solved many times on subintervals. Furthermore, the effort for sensitivity generation by IND is directly related to the number of nominal integration steps. Hence, an efficient choice of the nominal integration scheme based on global error control and its reuse in several optimization iterations promise a significant speed up in the overall solution procedure.

Aims and contributions of this thesis

The first aim of this thesis is to give an interpretation of the oscillating discrete adjoints of multistep BDF methods as they are observed by applying adjoint IND to the nominal BDF integration scheme and to relate these discrete adjoints to the solution of the adjoint IVP. The second aim of the thesis is to develop goal-oriented global error estimators for BDF methods with variable order and stepsize using discrete IND adjoints. For the first time, we build a bridge from BDF methods

and their adjoint IND schemes to Petrov-Galerkin Finite Element (FE) methods and carry over the Dual Weighted Residual (DWR) methodology for a posteriori error estimation, going back to Becker and Rannacher [19, 18], from FE methods for PDEs to BDF methods for ODEs. In the following the novel results of the thesis are described in detail.

A new variational formulation of IVPs giving rise to weak adjoints

Unfortunately, the special nature of multistep methods caused by the reuse of past approximations prohibits the analysis of BDF integration schemes and their adjoint IND schemes using the common variational formulation of IVPs in ODEs in Hilbert spaces. In this thesis we develop a new functional-analytic framework that is based on the duality pairing of continuous functions and normalized functions of bounded variation. This framework provides a well-posed variational formulation of IVPs in the more general Banach spaces of continuously differentiable functions and normalized functions of bounded variation. The application of this framework gives rise to the definition of weak adjoint solutions of adjoint IVPs. The weak adjoints are provided by the normalized integrals of the classical Hilbert space adjoints.

Petrov-Galerkin FE formulation of BDF methods and their adjoint IND schemes

We explicitly specify FE spaces that allow to transform the new formulation into finite dimensional Petrov-Galerkin equations. We show that they are equivalent to BDF methods with variable, but prescribed order and stepsize together with their discrete adjoint IND schemes. Hence, discretization and differentiation commute in the new framework. Thus, the BDF method represents an efficient formulation of the Petrov-Galerkin FE method and the oscillating discrete adjoint IND values are related to the classical solutions of adjoint IVPs via their weak adjoint solutions. The FE approximations to the weak adjoints are defined for any time point whereas the adjoint IND values are given only at discretization points. Furthermore, the FE approximations can be computed automatically by adjoint IND schemes without the explicit derivation of adjoint equations by the user.

Convergence of discrete adjoint IND values and FE weak adjoints

For BDF methods with constant order and stepsize in all integration steps except the starting steps, the adjoint IND schemes can be divided into three parts: the adjoint initialization steps, the adjoint main steps, and the adjoint termination steps caused by the nominal starting steps. The adjoint main steps are BDF steps that are consistent with a particular adjoint IVP, whereas the adjoint initialization and termination steps are always inconsistent. Nevertheless, using the strong stability of BDF methods we prove that the IND adjoints converge to the classical adjoint solutions on the main steps, which cover in the limit the open time interval. Then, we use this result to show the linear convergence of the FE approximations to the

weak adjoints in the total variation norm, and show that this implies the pointwise convergence on the entire time interval.

Goal-oriented global error estimation

We derive novel estimators for the global error in a criterion of interest evaluated at the final state of the IVP solution. These goal-oriented error estimators summate over all integration steps a nominal local error quantity multiplied by the adjoint IND value which exactly describes the sensitivity of the discrete final state on intermediate perturbations. Using the DWR methodology we derive that the nominal local error quantities are provided by the defect integrals of the nominal approximation. Combining the DWR methodology and the classical BDF convergence theory we additionally propose the local truncation errors as nominal local error quantities. We investigate both goal-oriented error estimators theoretically for BDF methods with constant order and stepsize and expose their relation to each other. Then, we further approximate them to evaluable versions in practical implementations, demonstrate their performance in terms of accuracy in fully adaptive BDF-type methods and show their superiority to an existing estimator proposed by Cao and Petzold [43].

Application of estimators for global error control

We employ the novel goal-oriented error estimates to obtain global error controlled approximations of IVP solutions. This is achieved by two different adaption strategies. The goal-oriented local tolerance adaption uses successively the goal-oriented error estimates to adapt the local relative tolerances for subsequent integrations with the standard selection mechanism for stepsize and order. The goal-oriented scheme adaption employs the error estimates and their local error indicators to directly adapt the integration schemes for subsequent integrations and completely replaces the standard selection mechanism for stepsize and order. It turns out that in this case the termination tolerances for the numerical solution of the nonlinear BDF equations are not fixed over all integration steps anymore, but have to be adjusted according to the local conditions.

Modeling and global error controlled simulation of a real-world example

The hydrolysis of propionic anhydride carried out in a discontinuous Stirred Tank Reactor (STR) is a representative for a wide class of strongly exothermic, self-accelerating reactions that are of great importance for the fine chemical industry. This particular reaction is realized in a laboratory-scale reactor and is used for research on detection and avoidance of thermal runaways, see e.g. Westerterp and Molga [125] and Molga and Cherbański [93, 94]. We build up a new ODE model of this dynamic process using validated subcomponents of previous work by Molga and Cherbański [93, 94] and Cherbański [44]. The resulting IVPs in ODEs are

highly nonlinear due to the mass transport term and the reaction rate coefficient of Arrhenius type. The newly composed model is able to describe experimental data which we have measured at the Warsaw University of Technology. Moreover, we show the applicability and reliability of the novel mathematical and numerical results of this thesis for the simulation of this real-world example with goal-oriented global error control.

Thesis overview

This thesis is organized in four parts: The status quo of BDF methods and their discrete adjoints, a novel interpretation for discrete adjoints of BDF methods, novel goal-oriented global error estimation for BDF methods, and numerical results.

Part I, which presents the status quo of BDF methods and their discrete adjoints, is divided into four chapters. In Chapter 1 we first introduce IVPs in ODEs and present the basic IVP theory including uniqueness and differentiability of the solution. Moreover, we introduce the adjoint IVP of the nominal IVP as adjoint problem of the forward variational IVP giving the sensitivities (derivatives) of the nominal solution with respect to the initial values. Conditioning and stiffness of IVPs are covered as well.

Chapter 2 describes multistep BDF methods to solve IVPs. The different error types appearing in multistep methods are defined and the convergence theory is presented. Finally, we sketch those aspects of practical realizations of efficient BDF-type methods that are of importance for the thesis.

In Chapter 3 the derivative generation for functions is briefly presented before we focus on the computation of derivatives of IVP solutions. We describe the two approaches of integrating the variational IVPs on the one hand or applying finite dimensional differentiation methods like Algorithmic Differentiation (AD) to the nominal integration scheme, i.e. Bock's IND approach, on the other hand. The adjoint versions of both approaches are examined in detail and we start to investigate the adjoint IND schemes of BDF methods in terms of integration methods applied to particular adjoint IVPs.

Chapter 4 summarizes basic concepts from real and functional analysis that are of great importance for Part II of this thesis.

Part II, which deals with a novel interpretation for discrete adjoints of BDF methods, is divided into three chapters. In Chapter 5 we first review the classical derivation of the adjoint IVP along the exact nominal solution in Hilbert spaces as part of the optimality conditions of a particular infinite dimensional Constrained Variational Problem (CVP). Since the Hilbert space formulation is not suitable to analyze BDF methods, we embed the CVP into the Banach space of all continuously differentiable functions and use the duality pairing between continuous functions and normalized functions of bounded variation. Using the new infinite dimensional optimality con-

ditions we define weak adjoint solutions, show their relation to the classical Hilbert space adjoints, and demonstrate the well-posedness of the new optimality conditions. Finally, we extend the setting to capture the space of all functions that are continuous and piecewise continuously differentiable.

Chapter 6 is devoted to the Petrov-Galerkin FE discretization of the infinite dimensional optimality conditions. We choose suitable FE spaces and demonstrate the equivalence between the discretized optimality conditions and the BDF scheme with variable, but prescribed order and stepsize together with its adjoint IND scheme. Finally, the commutativity of differentiation and discretization in the novel functional-analytic setting is elucidated as well as the so-called adjoint consistency of the adjoint IND scheme with the adjoint IVP.

We start Chapter 7 by proving the linear convergence of the discrete adjoint IND values of a BDF method with constant order and stepsize to the solution of the classical Hilbert space adjoint on the open time interval. Then, we show the convergence of the FE approximation to the weak adjoint solution on the entire interval using the former result.

Part III, which is about novel goal-oriented global error estimation for BDF methods, is divided into two chapters. In Chapter 8 we derive novel goal-oriented global error estimators for multistep BDF methods with variable order and stepsize. With the DWR methodology and a suitable approximation of the weights involving the unknown exact weak adjoints, approximations for the global error in a criterion of interest are developed. These approximations use the discrete adjoints provided by adjoint IND schemes. This is the first time that values generated by adjoint IND are used in a posteriori estimators for the goal-oriented global error. We can use defect integrals, local errors or local truncation errors as nominal local error quantities in the goal-oriented error approximations. After investigation of these error approximations for BDF methods with constant order and stepsize we derive, by further approximations, evaluable global error estimators for practical implementation.

With these goal-oriented error estimators at hand, Chapter 9 is dedicated to integrations by BDF methods with goal-oriented global error control. Two goal-oriented adaption strategies are proposed. The first one adapts the relative tolerance using the estimated global error and then uses the standard selection mechanism for the adaptive components. The second strategy adapts the integration scheme itself employing the local error indicators provided by the global error estimators and thus replaces the standard selection mechanism except the monitor strategy for matrix updates.

Part IV on numerical results is divided into three chapters. Chapter 10 starts with the numerical validation of the theoretical results of Part II. An academic nonlinear test case with analytic solutions is used to confirm numerically the convergence results of Chapter 7. Additionally, we give numerical evidence that the FE approximation serves as proper quantity to approximate the weak adjoint also in the case

of fully variable BDF-type methods as used in practice. Secondly, the goal-oriented global error approximations are investigated for BDF methods with constant order and stepsize and the error estimators are investigated for BDF-type methods with variable order and stepsize using IVPs with known analytic solutions. It turns out, that the estimators based on defect integrals and local truncation errors should be favored in practice and in fact are superior to an existing estimator.

In Chapter 11 we investigate numerically the goal-oriented global error control strategies. We use the goal-oriented global error estimator based on local truncation errors and IVP examples with analytic solutions. Both strategies, the local tolerance adaption and the scheme adaption, give approximations of the final state up to the desired accuracy. Depending on the local conditioning of the IVP, we point out the better of the two strategies.

Chapter 12 treats a real-world example, particularly its modeling and simulation. We describe the hydrolysis of propionic anhydride carried out in a discontinuous STR and model it by IVPs in ODEs. The new composed model reflects the real process which we have conducted at the Warsaw University of Technology. Moreover, using BDF methods with variable order and stepsize to solve this real-world IVP and determine its sensitivity in a safety function we confirm the reliability of the FE weak adjoints given via the adjoint IND values. We also use the goal-oriented global error control strategies to obtain approximations with a desired accuracy in the safety function.

In the last chapter we briefly summarize the results of this thesis and give some ideas for future research directions.

Appendix A starts with some frequently used definitions and theorems. Then, we prove some lemmas that are stated and used in Chapter 3 and 8, respectively. Furthermore, it contains the IVP test set as well as further data for the real-world example.

Part I

Status quo of BDF methods and their discrete adjoints

1 Theory of Initial Value Problems

This section reviews the basic theory of Initial Value Problems (IVPs) in Ordinary Differential Equations (ODEs). After the formal definition of an IVP, we recall the sufficient conditions for its well-posedness and the differentiability of its solution with respect to initial values. The sensitivity of the final state with respect to the initial values can be found in two different ways: the forward and the adjoint way. We close the chapter by some words on the stiffness of IVPs.

Definition 1.1 (Initial Value Problem) *Let be* $[t_\mathrm{s}, t_\mathrm{f}] \subset \mathbb{R}$. *An* Initial Value Problem (IVP) in Ordinary Differential Equations (ODEs) *is defined by a system of* d *first-order ODEs and* d *initial conditions*

$$\dot{\boldsymbol{y}}(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), \quad t \in [t_\mathrm{s}, t_\mathrm{f}] \tag{1.1a}$$

$$\boldsymbol{y}(t_\mathrm{s}) = \boldsymbol{y}_s \tag{1.1b}$$

where the right hand side $\boldsymbol{f}\colon [t_\mathrm{s}, t_\mathrm{f}] \times \mathbb{R}^d \to \mathbb{R}^d$ *and the unknown dynamic state* $\boldsymbol{y}\colon [t_\mathrm{s}, t_\mathrm{f}] \to \mathbb{R}^d$ *are vector-valued functions,* $t \in [t_\mathrm{s}, t_\mathrm{f}]$ *is the independent variable and* $\boldsymbol{y}_\mathrm{s} \in \mathbb{R}^d$ *the initial state vector (also called initial value). The componentwise derivative of* $\boldsymbol{y}$ *with respect to* t *is denoted by* $\dot{\boldsymbol{y}}$.

Definition 1.2 *For a matrix-valued function* $\boldsymbol{A}\colon [t_\mathrm{s}, t_\mathrm{f}] \to \mathbb{R}^{d\times d}$ *and a vector-valued function* $\boldsymbol{b}\colon [t_\mathrm{s}, t_\mathrm{f}] \to \mathbb{R}^d$ *the system*

$$\dot{\boldsymbol{y}}(t) = \boldsymbol{A}(t)\boldsymbol{y}(t) + \boldsymbol{b}(t)$$

is called a system of linear ODEs.

For this thesis, a functional output of IVP solutions is of great importance.

Definition 1.3 (Criterion of interest) *By a* criterion of interest *we mean a non-linear, sufficiently often differentiable functional* J *that is evaluated in the final state* $\boldsymbol{y}(t_\mathrm{f})$ *of the solution of IVP* (1.1).

Such a criterion is relevant whenever one is not interested in the whole solution $\boldsymbol{y}(t)$ of (1.1) or even the final state $\boldsymbol{y}(t_\mathrm{f})$, but only in a functional output of these quantities.

The settings of Definition 1.1 and 1.3 also capture the cases of a parameter-dependent right hand side $\boldsymbol{f}(t, \boldsymbol{y}, \boldsymbol{p})$ and a criterion of interest of Bolza type

$$J(\boldsymbol{y}) = \int_{t_\mathrm{s}}^{t_\mathrm{f}} J_1(\boldsymbol{y}(t), \boldsymbol{p})\mathrm{d}t + J_2(\boldsymbol{y}(t_\mathrm{f}))$$

due to standard reformulations, see Hartman [69] and Berkovitz [22].

1.1 Well-posedness of Initial Value Problems

This section focuses on the existence of a unique IVP solution and the well-posedness of the IVP. Due to Hadamard, a problem is *well-posed* if (i) a solution exists, (ii) the solution is unique, and (iii) the solution depends continuously on the input data. To investigate the well-posedness of an IVP we need the following property.

Definition 1.4 *The function $\boldsymbol{f}(t, \boldsymbol{y})$ defined on $\mathcal{D} \subset \mathbb{R} \times \mathbb{R}^d$ is said to be* Lipschitz continuous *on $\mathcal{D}$ with respect to $\boldsymbol{y}$, if a* Lipschitz constant *$L > 0$ exists such that*

$$\|\boldsymbol{f}(t, \boldsymbol{y}) - \boldsymbol{f}(t, \boldsymbol{y}^*)\| \leq L \|\boldsymbol{y} - \boldsymbol{y}^*\| \quad \forall (t, \boldsymbol{y}), (t, \boldsymbol{y}^*) \in \mathcal{D}.$$

Theorem 1.5 (Picard-Lindelöf) *Let $\boldsymbol{f}(t, \boldsymbol{y})$ be continuous on the region $\mathcal{R} = \{(t, \boldsymbol{y}) : t_\mathrm{s} \leq t \leq t_\mathrm{s} + a, \|\boldsymbol{y} - \boldsymbol{y_s}\| \leq b\} \subseteq \mathcal{D}$, Lipschitz continuous with respect to $\boldsymbol{y}$, and bounded by $\|\boldsymbol{f}(t, \boldsymbol{y})\| \leq M$ on $\mathcal{R}$. Then,*

$$\dot{\boldsymbol{y}}(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), \quad \boldsymbol{y}(t_\mathrm{s}) = \boldsymbol{y}_\mathrm{s} \tag{1.2}$$

has a unique solution $\boldsymbol{y}(t)$ on $[t_\mathrm{s}, t_\mathrm{s} + \alpha]$, where $\alpha = \min\{a, b/M\}$.

Proof *See Hartman [69].* □

The proof of Theorem 1.5 also shows that $\boldsymbol{y}(t)$ is continuously differentiable in t.

Remark 1.6 *If $\boldsymbol{f}(t, \boldsymbol{y})$ is differentiable with respect to $\boldsymbol{y}$, then L can be chosen as a bound on $\boldsymbol{f_y}(t, \boldsymbol{y})$ using any matrix norm, i.e. $L = \sup_{(t,\boldsymbol{y}) \in \mathcal{R}} \|\boldsymbol{f_y}(t, \boldsymbol{y})\|$.*

For Hadamard well-posedness of (1.2) it remains to guarantee the continuous dependency on the input data. For IVPs, the input data are given by the initial value $\boldsymbol{y}_\mathrm{s}$ and the right hand side $\boldsymbol{f}(t, \boldsymbol{y})$.

Theorem 1.7 *Let $\boldsymbol{f}(t, \boldsymbol{y})$, $\boldsymbol{g}(t, \boldsymbol{y})$ be continuous on the open set $\mathcal{D}$ and $\boldsymbol{f}(t, \boldsymbol{y})$ Lipschitz continuous in $\boldsymbol{y}$ with Lipschitz constant L. Suppose that*

$$\|\boldsymbol{f}(t, \boldsymbol{y}) - \boldsymbol{g}(t, \boldsymbol{y})\| \leq \varepsilon \quad \forall (t, \boldsymbol{y}) \in \mathcal{D}.$$

If $(t, \boldsymbol{y}(t))$ defined by the ODE of (1.2) and $(t, \boldsymbol{u}(t))$ defined by

$$\dot{\boldsymbol{u}}(t) = \boldsymbol{g}(t, \boldsymbol{u}(t))$$

lie in $\mathcal{D}$, then

$$\|\boldsymbol{y}(t) - \boldsymbol{u}(t)\| \leq \{\|\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{u}(t_\mathrm{s})\| + a\varepsilon\} \exp\left(L(t - t_\mathrm{s})\right).$$

Proof *See Shampine and Gordon [111].* □

1.2 Derivatives of IVP solutions with respect to initial values

By $\boldsymbol{y}(t; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ we denote explicitly the dependency of the solution $\boldsymbol{y}(t)$ of (1.1) on the initial condition $\boldsymbol{y}(t_\mathrm{s}) = \boldsymbol{y}_\mathrm{s}$. This section is devoted to the derivatives of the solution $\boldsymbol{y}(t; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ at the final time $t = t_\mathrm{f}$ with respect to $\boldsymbol{y}_\mathrm{s}$, or some subspace direction, and the derivatives of a functional output on $\boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ with respect to $\boldsymbol{y}_\mathrm{s}$.

1.2.1 Forward variational Initial Value Problem

Theorem 1.8 *Let $\boldsymbol{f}(t, \boldsymbol{y})$ be continuous on the open set $\mathcal{D}$ and exhibit a first-order partial derivative $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y})$ that is continuous on $\mathcal{D}$. Then, the unique solution $\boldsymbol{y}(t) = \boldsymbol{y}(t; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ of (1.1) is continuously differentiable in t and $(t_\mathrm{s}, \boldsymbol{y}_\mathrm{s}) \in \mathcal{D}$. Furthermore, the derivative $\boldsymbol{W}(t) = \partial \boldsymbol{y}(t; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s}) / \partial \boldsymbol{y}_\mathrm{s}$ of the solution $\boldsymbol{y}(t; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ with respect to the initial value $\boldsymbol{y}_\mathrm{s}$ solves the IVP in matrix form*

$$\dot{\boldsymbol{W}}(t) = \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t)) \boldsymbol{W}(t), \quad \boldsymbol{W}(t_\mathrm{s}) = \boldsymbol{I} \tag{1.3}$$

where $\boldsymbol{I}$ is the $d \times d$ unit matrix.

Proof *See Hartman [69].* □

Due to the assumptions of Theorem 1.8 on $\boldsymbol{f}(t, \boldsymbol{y})$, the right hand side of (1.3) and its partial derivative with respect to $\boldsymbol{W}$ are continuous in $(t, \boldsymbol{W})$. Hence, the solution $\boldsymbol{W}(t)$ exists uniquely and is continuously differentiable in t, cf. Theorem 1.5. It describes the dependency of $\boldsymbol{y}(t)$ at any time $t \in [t_\mathrm{s}, t_\mathrm{f}]$ on the whole initial state vector $\boldsymbol{y}_\mathrm{s}$. In case that the derivative of $\boldsymbol{y}(t)$ with respect to a subspace direction $\boldsymbol{v} \in \mathbb{R}^{d \times 1}$ of the whole initial state vector is of interest, the vector-valued derivative $\boldsymbol{w}(t) = \boldsymbol{W}(t)\boldsymbol{v}$ solves the so-called *forward variational IVP*

$$\dot{\boldsymbol{w}}(t) = \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t)) \boldsymbol{w}(t), \quad t \in [t_\mathrm{s}, t_\mathrm{f}], \quad \boldsymbol{w}(t_\mathrm{s}) = \boldsymbol{v}. \tag{1.4}$$

For a criterion of interest J on $\boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ (see Definition 1.3), the derivative $\partial J(\boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})) / \partial \boldsymbol{y}_\mathrm{s}$ of J with respect to the initial value $\boldsymbol{y}_s$ is given by

$$\partial J(\boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})) / \partial \boldsymbol{y}_\mathrm{s} = J'(\boldsymbol{y}(t_\mathrm{f})) \boldsymbol{W}(t_\mathrm{f}) \tag{1.5}$$

where $\boldsymbol{W}(t)$ solves (1.3). Alternatively, it can be obtained by solving a so-called adjoint IVP which will be in the focus of the following section.

1.2.2 Adjoint Initial Value Problem

In case that the derivative of a subspace direction $\boldsymbol{r} \in \mathbb{R}^{d \times 1}$ of the whole solution $\boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ with respect to the initial state vector $\boldsymbol{y}_\mathrm{s}$ is of interest, it is more efficient to solve the so-called *adjoint variational IVP*

$$\dot{\boldsymbol{\lambda}}(t) = -\boldsymbol{f}_{\boldsymbol{y}}^{\intercal}(t, \boldsymbol{y}(t)) \boldsymbol{\lambda}(t), \quad t \in [t_\mathrm{s}, t_\mathrm{f}], \quad \boldsymbol{\lambda}(t_\mathrm{f}) = \boldsymbol{r} \tag{1.6}$$

backwards in time. Solving (1.6) for $\boldsymbol{r} = \boldsymbol{e}_i$ with $i = 1, \ldots, d$ yields the solutions $\boldsymbol{\lambda}_i(t)$ and one obtains (rowwise) the derivative $\boldsymbol{\Lambda}^\intercal(t) := \partial \boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})/\partial \boldsymbol{y}(t)$ of the whole final state $\boldsymbol{y}(t_\mathrm{f})$ with respect to the solution $\boldsymbol{y}(t)$ at any time $t \in [t_\mathrm{s}, t_\mathrm{f}]$. Hence, $\boldsymbol{\Lambda}(t)$ solves the IVP in matrix form

$$\dot{\boldsymbol{\Lambda}}(t) = -\boldsymbol{f}_{\boldsymbol{y}}^\intercal(t, \boldsymbol{y}(t))\boldsymbol{\Lambda}(t), \quad t \in [t_\mathrm{s}, t_\mathrm{f}], \quad \boldsymbol{\Lambda}(t_\mathrm{f}) = \boldsymbol{I} \tag{1.7}$$

and describes the dependency of the final state $\boldsymbol{y}(t_\mathrm{f})$ on $\boldsymbol{y}(t)$ at any $t \in [t_\mathrm{s}, t_\mathrm{f}]$. Only in this chapter, $\boldsymbol{\Lambda}(t)$ denotes the matrix that is composed of the rows $\boldsymbol{\lambda}_i(t)$. Everywhere else in this thesis, $\boldsymbol{\Lambda}(t)$ will denote a vector-valued function of bounded variation.

Theorem 1.9 *With the assumptions of Theorem 1.8 and the solution $\boldsymbol{y}(t)$ of (1.1), the solutions $\boldsymbol{W}(t)$ of (1.3) on $[t_\mathrm{s}, t_\mathrm{f}]$ and $\boldsymbol{\Lambda}(t)$ of (1.7) are related by*

$$\boldsymbol{\Lambda}^\intercal(t)\boldsymbol{W}(t) = \boldsymbol{W}(t_\mathrm{f}), \quad t \in [t_\mathrm{s}, t_\mathrm{f}]$$

and, in particular, by $\boldsymbol{\Lambda}^\intercal(t_\mathrm{s}) = \boldsymbol{W}(t_\mathrm{f})$.

Proof *From the ODE of (1.3) we obtain for any $t \in [t_\mathrm{s}, t_\mathrm{f}]$*

$$\int_{t_\mathrm{s}}^{t} \boldsymbol{\Lambda}^\intercal(\tau) \left[\dot{\boldsymbol{W}}(\tau) - \boldsymbol{f}_{\boldsymbol{y}}(\tau, \boldsymbol{y}(\tau))\boldsymbol{W}(\tau)\right] \mathrm{d}\tau = \boldsymbol{0}.$$

Integration by parts yields

$$\boldsymbol{0} = \boldsymbol{\Lambda}^\intercal(t)\boldsymbol{W}(t) - \boldsymbol{\Lambda}^\intercal(t_\mathrm{s})\boldsymbol{W}(t_\mathrm{s}) - \int_{t_\mathrm{s}}^{t} \left[\dot{\boldsymbol{\Lambda}}(\tau) + \boldsymbol{f}_{\boldsymbol{y}}^\intercal(\tau, \boldsymbol{y}(\tau))\boldsymbol{\Lambda}(\tau)\right]^\intercal \boldsymbol{W}(\tau)\mathrm{d}\tau.$$

With the ODE of (1.7) and the initial conditions of (1.3) and (1.7) the assertions follow. □

Aside from (1.5), the derivative $\partial J(\boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s}))/\partial \boldsymbol{y}_\mathrm{s}$ of a criterion J on $\boldsymbol{y}(t_\mathrm{f}; t_\mathrm{s}, \boldsymbol{y}_\mathrm{s})$ is also given by the transposed solution $\boldsymbol{\lambda}^\intercal(t_\mathrm{s})$ of the so-called *adjoint IVP*

$$\dot{\boldsymbol{\lambda}}(t) = -\boldsymbol{f}_{\boldsymbol{y}}^\intercal(t, \boldsymbol{y}(t))\boldsymbol{\lambda}(t), \quad t \in [t_\mathrm{s}, t_\mathrm{f}] \tag{1.8a}$$

$$\boldsymbol{\lambda}(t_\mathrm{f}) = J'(\boldsymbol{y}(t_\mathrm{f}))^\intercal \tag{1.8b}$$

solved backwards in time. Note that by the adjoint approach, the derivative of a scalar criterion with respect to the initial state vector is given by solving a single IVP, whereas by the forward approach a system of d IVPs has to be solved.

From now on, we assume that the right hand side $\boldsymbol{f}(t, \boldsymbol{y})$ of IVP (1.1) satisfies at least the assumptions of Theorem 1.5 and that the final time t_f is given such that $[t_\mathrm{s}, t_\mathrm{f}] \subseteq \mathcal{R}$. From Chapter 3 on, we suppose additionally that the assumptions of Theorem 1.8 are fulfilled.

1.3 Conditioning of Initial Value Problems

A crucial property of IVPs is their inherent *conditioning*, or also called *stability*. It specifies the sensitivity of the IVP solution with respect to input perturbations, i.e. it describes how small changes in the input data affect the output of the IVP. Usually, for mathematical problems the term 'conditioning' is used whereas the term 'stability' refers to the corresponding property of numerical algorithms. In the context of differential equations, 'stability' is used for both, the problem and the numerical method for solving it. The following definition of the problem stability can be found, for example, in Heath [71] and Strehmel and Weiner [117].

Definition 1.10 *A solution $\boldsymbol{y}(t)$ of the ODE* (1.1a) *is said to be* (Liapunov-) stable *if for every $\varepsilon > 0$ there exists a $\delta > 0$ such that*

$$\boldsymbol{u}(t) \text{ solves (1.1a) and } \|\boldsymbol{u}(t_\mathrm{s}) - \boldsymbol{y}(t_\mathrm{s})\| \leq \delta \;\Rightarrow\; \|\boldsymbol{u}(t) - \boldsymbol{y}(t)\| \leq \varepsilon \quad \forall t \geq t_\mathrm{s}.$$

If additionally $\|\boldsymbol{u}(t) - \boldsymbol{y}(t)\| \to 0$ as $t \to \infty$, then $\boldsymbol{y}(t)$ is said to be asymptotically stable. *The solution $\boldsymbol{y}(t)$ is said to be* unstable, *if it is not stable.*

The stability of an IVP solution is, amongst the stability of the numerical algorithm used to solve the IVP, crucial for the accuracy of the computed approximation. Introduced errors during the computations are either reduced (asymptotically stable), maintained (stable) or accumulated (unstable).

The stability of an IVP solution can be determined in first order by the forward variational IVP (1.4): The solution $\boldsymbol{y}(t)$ of (1.1) is stable if, in Definition 1.10, $\|\boldsymbol{v}\| \leq \delta$ implies that $\|\boldsymbol{w}(t)\| \leq \varepsilon$ for the solution $\boldsymbol{w}(t)$ of (1.4). For a linear ODE with constant matrix $\boldsymbol{A}$, cf. Definition 1.2, the stability of its solution is characterized by the eigenvalues μ_i of $\boldsymbol{A}$, $1 \leq i \leq d$.

- If $\operatorname{Re}\mu_i < 0$ for all $i = 1, \ldots, d$, then the solution is asymptotically stable.
- If $\operatorname{Re}\mu_i \leq 0$ for all $i = 1, \ldots, d$ and $\operatorname{Re}\mu_i < 0$ for any μ_i that is not simple, then the solution is stable.
- If for any $i \in \{1, \ldots, d\}$ holds $\operatorname{Re}\mu_i > 0$, then the solution is unstable.

For general ODEs of the form (1.1a) an indication of the stability of a solution $\boldsymbol{y}(t)$ can be obtained by the time-varying eigenvalues $\mu_i(t)$ of the Jacobian $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))$. But the gained information is valid only locally in $(t, \boldsymbol{y}(t))$.

To determine the stability of the final solution $\boldsymbol{y}(t_\mathrm{f})$ of (1.1) in a criterion of interest J, i.e. in the output data $J(\boldsymbol{y}(t_\mathrm{f}))$, the adjoint IVP (1.8) can be used. According to Section 1.2 we have

$$\begin{aligned} J(\boldsymbol{u}(t_\mathrm{f})) - J(\boldsymbol{y}(t_\mathrm{f})) &\doteq J'(\boldsymbol{y}(t_\mathrm{f}))[\boldsymbol{u}(t_\mathrm{f}) - \boldsymbol{y}(t_\mathrm{f})] = J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f}) \\ &= J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{W}(t_\mathrm{f})\boldsymbol{v} = J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{\Lambda}^\intercal(t_\mathrm{f})\boldsymbol{v} = \boldsymbol{\lambda}^\intercal(t_\mathrm{f})\boldsymbol{v} \end{aligned}$$

such that if, in Definition 1.10, $\|\boldsymbol{v}\| \leq \delta$ implies that $\|\boldsymbol{\lambda}^{\mathsf{T}}(t_\mathrm{f})\boldsymbol{v}\| = |\boldsymbol{\lambda}^{\mathsf{T}}(t_\mathrm{f})\boldsymbol{v}| \leq \varepsilon$, then $\boldsymbol{y}(t_\mathrm{f})$ is stable in J with respect to perturbations in $\boldsymbol{y}_\mathrm{s}$. The effect of intermediate perturbations is described by $\boldsymbol{\lambda}(t)$ such that for $t \in [t_\mathrm{s}, t_\mathrm{f}]$ also $\|\boldsymbol{\lambda}(t)\| \leq \varepsilon$ should be satisfied, see description on page 48.

1.3.1 Condition number

The stability (conditioning) of an IVP solution in a criterion can be summarized by a scalar number limiting the ratio of changes in the output $J(\boldsymbol{y}(t_\mathrm{f}))$ and changes in the input $\boldsymbol{y}_\mathrm{s}$ and $\boldsymbol{f}(t, \boldsymbol{y})$. Using the L_1-norm we define the *condition number* by

$$\kappa := \|\boldsymbol{\lambda}(t_\mathrm{s})\|_1 + \|\boldsymbol{\lambda}\|_{L^1(t_\mathrm{s},t_\mathrm{f})^d} = \sum_{i=1}^{d} |\lambda_i(t_\mathrm{s})| + \sum_{i=1}^{d} \int_{t_\mathrm{s}}^{t_\mathrm{f}} |\lambda_i(t)| \,\mathrm{d}t.$$

The first term reflects the condition number with respect to changes in the initial values $\boldsymbol{y}_\mathrm{s}$ and the second with respect to changes in the right hand side $\boldsymbol{f}(t, \boldsymbol{y})$. Due to the norm, the condition number κ does not take into account the effects of error cancellations. Therefore, it is a worst-case quantity. This definition of the condition number is also used by Cao and Petzold [43] and Lang and Verwer [84].

1.4 Stiffness of Initial Value Problems

In many fields of application, for example in chemical engineering, as well as in the spatial discretization of instationary Partial Differential Equations (PDEs) using a method of lines approach ODEs appear that exhibit a certain property which is called *stiffness*. This property was first mentioned by Curtiss and Hirschfelder [46] who described it by the fact that the implicit Backward Differentiation Formula (BDF) methods work much better on these ODEs than explicit approaches. Unfortunately, there exists no general definition. Usual characterizations of stiffness along a solution $\boldsymbol{y}(t)$ of (1.1a) use the eigenvalues $\mu_i(t)$ of $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))$

- $\max_{\mathrm{Re}\,\mu_i(t)<0} |\mathrm{Re}\,\mu_i(t)| \gg \min_{\mathrm{Re}\,\mu_i(t)<0} |\mathrm{Re}\,\mu_i(t)|$
- $(t_\mathrm{f} - t_\mathrm{s}) \min_{1\leq i\leq d} \mathrm{Re}\,\mu_i(t) \ll -1$.

Stiffness is not a property of the ODE right hand side but of the IVP which can be stiff for certain initial values and/or certain time intervals. Phenomenological, a stiff IVP has slowly changing solution components and others that, in the transient phase, approach fastly a decaying steady state. For more aspects on stiff IVPs we refer to Hairer and Wanner [68], Shampine [109] and Strehmel and Weiner [117].

Due to Remark 1.6 stiff ODEs exhibit a large Lipschitz constant L, since for any matrix norm it holds that $\|\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\| \geq \varrho(\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t)))$ with spectral radius $\varrho(\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))) := \max_{1\leq i\leq d} |\mu_i(t)| \geq \max_{1\leq i\leq d} |\mathrm{Re}\,\mu_i(t)|$. Hence, explicit integration methods have to use very small stepsizes in regions of stiffness and are not recommended for practical use. Stiff IVPs call for another stability concept of numerical methods, the so-called *A-stability*, see end of Section 2.3.1.

2 Numerical solution of Initial Value Problems

This chapter focuses on the numerical solution of Initial Value Problems (IVPs) in Ordinary Differential Equations (ODEs) by Linear Multistep Methods (LMMs) based on Backward Differentiation Formulas (BDF). In the first part we state the BDF method. In the second section we define the different error types appearing in multistep methods before we come in the third part to the theoretical properties of BDF methods with constant order and stepsize and of BDF methods with variable order and stepsize. In the last part of this chapter we focus on practical aspects of BDF-type methods.

2.1 Backward Differentiation Formula method

The general form of LMMs with variable order and variable stepsize is defined below.

Definition 2.1 (Linear Multistep Method) *For a time grid* $t_\mathrm{s} = t_0 < \cdots < t_N = t_\mathrm{f}$*, the* Linear Multistep Method (LMM) *with start values* $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ *for fixed* m *determines successively aproximations* $\{\boldsymbol{y}_n\}_{n=m+1}^{N}$ *to the solution* $\boldsymbol{y}(t)$ *of the IVP* (1.1) *by*

$$\sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} = h_n \sum_{i=0}^{k_n} \beta_i^{(n)} \boldsymbol{f}(t_{n+1-i}, \boldsymbol{y}_{n+1-i}),\ n = m, \ldots, N-1 \tag{2.1}$$

where $\alpha_0^{(n)} \neq 0$ *and* $\left|\alpha_k^{(n)}\right| + \left|\beta_k^{(n)}\right| > 0$*. The LMM is called* explicit *if* $\beta_0^{(n)} = 0$ *and* implicit *if* $\beta_0^{(n)} \neq 0$*.*

The term 'linear' refers to the fact that $\boldsymbol{y}_n$ and $\boldsymbol{f}_n := \boldsymbol{f}(t_n, \boldsymbol{y}_n)$ enter the integration formula linearly. Most integration methods are linear, e.g. Runge-Kutta methods.

The LMMs based on Backward Differentiation Formulas were invented by Curtiss and Hirschfelder [46] in 1952 and became popular for stiff IVPs with the work of Gear [61] in 1971. The basic idea is to interpolate past approximations and an unknown new approximation by a polynomial of a certain order and to require that the polynomial satisfies the ODE in the new time point. We state here the BDF method in its general form as presented, for example, in Shampine [109]. This form uses the Lagrange representation of the interpolation polynomial and allows to use variable stepsizes and variable orders. It is particularly qualified for

analyzing purposes, whereas the Newton representation of interpolation polynomials is preferred for practical implementations, see Section 2.4.

Definition 2.2 (Backward Differentiation Formula method) *For a time grid* $t_\mathrm{s} = t_0 < \cdots < t_N = t_\mathrm{f}$, *the* Backward Differentiation Formula (BDF) method *with a self-starting procedure determines successively approximations* $\{\boldsymbol{y}_n\}_{n=1}^N$ *to the solution* $\boldsymbol{y}(t)$ *of IVP* (1.1) *by*

$$\boldsymbol{y}_0 = \boldsymbol{y}_\mathrm{s} \tag{2.2a}$$

$$\sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} = h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}),\ n = 0, \ldots, N-1 \tag{2.2b}$$

with stepsizes $h_n = t_{n+1} - t_n$ *and orders* k_n. *The coefficients* $\alpha_i^{(n)}$ *are determined by*

$$\alpha_i^{(n)} = h_n \dot{L}_i^{(n)}(t_{n+1}) \tag{2.3}$$

where the fundamental Lagrange polynomials are

$$L_i^{(n)}(t) = \prod_{j=0, j\neq i}^{k_n} \frac{t - t_{n+1-j}}{t_{n+1-i} - t_{n+1-j}}. \tag{2.4}$$

In each integration step, the BDF method provides a continuous approximation to the exact solution $\boldsymbol{y}(t)$ of (1.1) in a natural way using the interpolation polynomials

$$\boldsymbol{y}(t)\big|_{t\in[t_n, t_{n+1}]} \approx \boldsymbol{\mathcal{P}}_{n+1}(t) := \sum_{i=0}^{k_n} L_i^{(n)}(t)\ \boldsymbol{y}_{n+1-i}, \tag{2.5}$$

also known as *dense output*, see Section 2.4.2.

BDF methods are implicit LMMs since $\beta_0^{(n)} = 1$ and $\beta_i^{(n)} = 0$ for $i = 1, \ldots, k_n$ in Definition 2.1 and normalized with respect to $\beta_0^{(n)}$. In other presentations, e.g. Lambert [83], LMMs are normalized by assuming $\alpha_0^{(n)} = 1$.

In the solution of implicit LMMs two difficulties arise: They need appropriate start values and an approach to solve implicit, nonlinear equations.

During the starting procedure of BDF methods, the start values $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ (with m fixed) have to be determined, since the IVP (1.1) only provides $\boldsymbol{y}_0 = \boldsymbol{y}_\mathrm{s}$. The self-starting procedure, already mentioned in Definition 2.2, begins with $k_0 = 1$ (implicit Euler) and increases successively the order of the integration steps until the maximum order is reached. An alternative would be to use Runge-Kutta methods to determine $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$. In this thesis only self-starters are considered, for Runge-Kutta starters we refer to Bauer [16].

Since the BDF method is implicit and the right hand side $\boldsymbol{f}(t, \boldsymbol{y})$ is nonlinear, in each integration step (2.2b) a nonlinear system of equations has to be solved. The *BDF equation* (2.2b) possesses a unique solution $\boldsymbol{y}_{n+1}$ if stepsize h_n and order k_n are chosen such that

$$\left| h_n / \alpha_0^{(n)} \right| \cdot \| \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1}) \| < 1 \tag{2.6}$$

is satisfied, see e.g. Henrici [72]. The solution $\boldsymbol{y}_{n+1}$ of the BDF equation is usually approximated by a Newton-type method, see Section 2.4.3.

To state all crucial assumptions for BDF methods with variable order and variable stepsize at one place, we already postulate that beside (2.6)

- the stepsize ratios $\omega_i := h_i/h_{i-1}$ are bounded
- the coefficients $\alpha_i^{(n)}$ are bounded

due to appropriate choices of the stepsizes h_n and orders k_n, see also Section 2.3.2.

2.2 Errors in Linear Multistep Methods

The numerical time stepping method (2.1) approximates the solution of the IVP (1.1) by a finite number of calculations. The difference between the exact solution $\boldsymbol{y}(t)$ of (1.1) at t_n and its approximation $\boldsymbol{y}_n$ by the LMM is of great interest to quantify the reliability of the method. Following the presentation of Shampine and Gordon [111] we define the global and local error.

Definition 2.3 *The* global error **GE** *at* t_{n+1} *is defined by*

$$\mathbf{GE}(t_{n+1}) := \boldsymbol{y}(t_{n+1}) - \boldsymbol{y}_{n+1}$$

where $\boldsymbol{y}_{n+1}$ *is given by* (2.1) *and* $\boldsymbol{y}(t)$ *is the exact solution of* (1.1)*.*

Definition 2.4 *The* local error **LE** *at* t_{n+1} *is defined by*

$$\mathbf{LE}(t_{n+1}) := \boldsymbol{u}_n(t_{n+1}) - \boldsymbol{y}_{n+1} \tag{2.7}$$

where $\boldsymbol{y}_{n+1}$ *is given by* (2.1) *and* $\boldsymbol{u}_n(t)$ *is the exact solution of the local IVP*

$$\begin{aligned} \dot{\boldsymbol{u}}_n(t) &= \boldsymbol{f}(t, \boldsymbol{u}_n(t)), \quad t \in (t_n, t_{n+1}] \\ \boldsymbol{u}_n(t_n) &= \boldsymbol{y}_n. \end{aligned}$$

Analogous to one-step methods, the local error of multistep methods describes the error produced by a single integration step. For details on one-step integration methods and their errors we refer to Butcher [41], Hairer et al. [67] and Hairer and Wanner [68]. Unfortunately, there is no unique naming convention for errors in LMMs. For example, the global error of Definition 2.3 is called discretization error by Henrici [72, 73], global truncation error or accumulated truncation error by Lambert [83]. Another important term is that of the local truncation error. Its definition follows that of Lambert [83].

Definition 2.5 *For continuously differentiable functions* $\boldsymbol{y}(t)$*, the* linear difference operator *associated to the n-th step of the LMM* (2.1) *is defined by*

$$\boldsymbol{L}^{(n)}(\boldsymbol{y}; t_{n+1}, h_{n+1-k_n}, \ldots, h_n) := \sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}(t_{n+1-i}) - h_n \sum_{i=0}^{k_n} \beta_i^{(n)} \dot{\boldsymbol{y}}(t_{n+1-i})$$

with $t_{n+1-i} = t_{n+1} - \sum_{j=1}^{i} h_{n+1-j}$ *for* $i = 0, \ldots, k_n$*.*

Definition 2.6 *The* local truncation error **LTE** *at* t_{n+1} *is defined by*

$$\mathbf{LTE}(t_{n+1}) := \boldsymbol{L}^{(n)}(\boldsymbol{y}; t_{n+1}, h_{n+1-k_n}, \ldots, h_n) \tag{2.8}$$

where $\boldsymbol{y}(t)$ *is the exact solution of* (1.1)*.*

Thus, the local truncation error is given by inserting the exact solution into the difference equation. It measures how well the integration step of the LMM (2.1) models the ODE (1.1a) locally.

Definition 2.7 (Localizing Assumption) *Assume that the past values of the* n*-th step of the LMM* (2.1) *are exact, i.e.* $\boldsymbol{y}_{n+1-i} = \boldsymbol{y}(t_{n+1-i})$ *for* $i = 1, \ldots, k_n$*.*

With the uniqueness assumption on the solution of (1.1) and the Localizing Assumption (Definition 2.7) the local error definition by Hairer et al. [67] equal with that of Definition 2.4. The local error and the local truncation error are related in the following way.

Lemma 2.8 *Let* $\boldsymbol{f}(t, \boldsymbol{y})$ *be continuous in* t *and continuously differentiable in* $\boldsymbol{y}$*. Let* $\boldsymbol{y}(t)$ *be the exact solution of* (1.1) *and* $\boldsymbol{y}_{n+1}$ *determined by the* n*-th step of* (2.1) *under the Localizing Assumption. Then, the local error and the local truncation error at* t_{n+1} *are related by*

$$\left(\alpha_0^{(n)} \boldsymbol{I} - h_n \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{\eta})\right) \mathbf{LE}(t_{n+1}) = \mathbf{LTE}(t_{n+1}) \tag{2.9}$$

where $\boldsymbol{\eta}$ *lies in the segment between* $\boldsymbol{u}_n(t_{n+1}) = \boldsymbol{y}(t_{n+1})$ *and* $\boldsymbol{y}_{n+1}$*.*

Proof *Substracting* (2.1) *from* (2.8)*, using the Localizing Assumption (Definition 2.7), the Mean Value Theorem and the definition of the local error (Definition 2.4), the assertion is shown.* □

So far, we have assumed that the nonlinear equations of the LMM (2.1) are solved exactly. But in practice that is not the case and so we define another type of error.

Definition 2.9 *For an approximation* $\check{\boldsymbol{y}}_{n+1}$ *of the exact solution* $\boldsymbol{y}_{n+1}$ *of the* n*-th step of the LMM* (2.1) *the* residual *of the nonlinear equation is defined by*

$$\boldsymbol{\delta}_{n+1} := \alpha_0^{(n)} \check{\boldsymbol{y}}_{n+1} - h_n \boldsymbol{f}\left(t_{n+1}, \check{\boldsymbol{y}}_{n+1}\right) + \sum_{i=1}^{k_n} \left\{ \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} - h_n \beta_i^{(n)} \boldsymbol{f}\left(t_{n+1-i}, \boldsymbol{y}_{n+1-i}\right) \right\}.$$

For time-continuous approximations to the solution of (1.1) we define global error function and defect. Especially, the latter will play a cruicial role in this thesis.

Definition 2.10 *For any approximation* $\boldsymbol{z}(t)$ *to the solution of* (1.1)*, the* global error function *is defined by*

$$\boldsymbol{e}(t) := \boldsymbol{y}(t) - \boldsymbol{z}(t).$$

Definition 2.11 *For any approximation* $\boldsymbol{z}(t)$ *to the solution of* (1.1)*, the* defect *is defined by*

$$\boldsymbol{r}(t) := \dot{\boldsymbol{z}}(t) - \boldsymbol{f}(t, \boldsymbol{z}(t)).$$

This definition of the defect given by Gear [61] is more general than that of Shampine and Gordon [111]. Both use the term 'residual' instead of 'defect'. But, to avoid confusions with the residual of Definition 2.9, we will stay with the term 'defect' which is also used by Hairer et al. [67]. The defect is available at any point $t \in [t_\mathrm{s}, t_\mathrm{f}]$ and measures to which extent the approximation $\boldsymbol{z}(t)$ does not satisfy the ODE (1.1a).

If the approximation $\boldsymbol{z}(t)$ passes through $\{\check{y}_n\}_{n=1}^N$ we can include the error due to the time discretization and that due to the approximate solution of the nonlinear equation into a single quantity.

2.3 Theoretical foundations of BDF methods

This section recalls the theoretical properties of constant and variable LMMs, investigates the errors defined in the last section and pays a particular attention to BDF methods. It is assumed that the nonlinear equations of the LMM (2.1) are solved exactly, unless otherwise indicated. The term 'constant' means to use the same order and stepsize for all integration steps except for the starting steps, whereas the term 'variable' adverts to the use of changing orders and stepsizes during the integration. In practical implementations of BDF methods, of course, both order and stepsize are chosen adaptively to improve the performance, see Section 2.4.

2.3.1 Constant BDF methods

In this section, we investigate the asymptotic behavior of the errors defined in the previous section. To this end, we consider a so-called *constant LMM* with constant order k and constant stepsizes h

$$\sum_{i=0}^{k} \alpha_i \boldsymbol{y}_{n+1-i} = h \sum_{i=0}^{k} \beta_i \boldsymbol{f}(t_{n+1-i}, \boldsymbol{y}_{n+1-i}),\ n = m, \ldots, N-1 \tag{2.10}$$

where the start values $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ (with $m \geq k-1$ fixed) are given by a starting procedure. In this sense, a so-called *constant BDF method* reads

$$\sum_{i=0}^{k} \alpha_i \boldsymbol{y}_{n+1-i} = h \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}),\ n = m, \ldots, N-1. \tag{2.11}$$

To ease the notion in this section, we consider a scalar IVP.

Convergence of constant BDF methods

We now focus on how the approximations $\{y_n\}_{n=0}^N$ generated by a constant LMM converge to the exact solution $y(t)$ as the stepsize h tends to zero. This convergence analysis is with respect to the limit as $h \to 0$ and $n \to \infty$ while $nh = t - t_\mathrm{s}$ remains fixed. We follow mainly the presentation of Lambert [83], but use the coefficient numbering and corresponding definitions of Shampine [109].

Definition 2.12 *The constant LMM* (2.10) *is said to be* convergent*, if, for all IVPs* (1.1) *where the right hand side $f(t, y)$ is continuous in t and Lipschitz continuous in y, holds*

$$\lim_{\substack{h \to 0 \\ nh = t - t_\mathrm{s}}} y_n = y(t_n)$$

for all $t \in [t_\mathrm{s}, t_\mathrm{f}]$ and all start values $y_0, \ldots, y_m$ with $\lim_{h\to 0} y_n = y_\mathrm{s}$, $n = 0, \ldots, m$.

Definition 2.13 *The LMM* (2.10) *is said to be* convergent of order q*, if, for all IVPs* (1.1) *with sufficiently smooth right hand side, there exists a positive $\hat{h}$ such that*

$$|y(t_n) - y_n| \leq K h^q \quad \text{for } h \leq \hat{h}$$

whenever the start values satisfy

$$|y(t_n) - y_n| \leq K_\mathrm{s} h^q \quad \text{for } h \leq \hat{h},\ n = 0, \ldots, m.$$

The latter definition can be found e.g. in Hairer et al. [67]. As we will see later (cf. Theorem 2.20), the necessary and sufficient conditions for LMMs to be convergent are to be consistent and zero-stable.

Definition 2.14 *The* characteristic polynomials *of the LMM* (2.10) *are*

$$\rho(\xi) := \sum_{i=0}^{k} \alpha_i \xi^{k-i}, \quad \sigma(\xi) := \sum_{i=0}^{k} \beta_i \xi^{k-i}.$$

Definition 2.15 *The linear difference operator of Definition 2.5 and the associated LMM* (2.10) *are said to be of* (consistency) order q*, if*

$$\boldsymbol{L}(y; t_{n+1}, h) := \boldsymbol{L}^{(n)}(y; t_{n+1}, h, \ldots, h) = \mathcal{O}(h^{q+1})$$

holds for sufficiently smooth functions $y(t)$ and $h \to 0$.

Definition 2.16 *The LMM* (2.10) *is said to be* consistent*, if it has order $q \geq 1$.*

Definition 2.16 justifies that the order q in Definition 2.15 is called *consistency order*. For continuously differentiable functions $y(t)$ of sufficiently high order, one

obtains by Taylor series expansions around t_{n+1} of $y(t_{n+1-i})$, $\dot{y}(t_{n+1-i})$ for $i = 1, \ldots, k$ the linear difference operator as

$$\boldsymbol{L}(y; t_{n+1}, h) = C_0 y(t_{n+1}) + C_1 h y'(t_{n+1}) + \cdots + C_q h^q y^{(q)}(t_{n+1}) + \ldots$$

with the following coefficients being independent of h

$$C_0 = \sum_{i=0}^{k} \alpha_i, \quad C_1 = -\sum_{i=1}^{k} \{i\alpha_i + \beta_i\}, \quad C_q = (-1)^q \sum_{i=1}^{k} \left\{ \frac{i^q \alpha_i}{q!} + \frac{i^{q-1}\beta_i}{(q-1)!} \right\}.$$

Hence, the LMM is consistent if $C_0 = \rho(1) = 0$ and $C_1 = \rho'(1) - \sigma(1) = 0$ (since $\rho'(1) = k\rho(1) - \sum_{i=0}^{k} i\alpha_i$). By construction, constant BDF methods with order k are of *consistency order* $q = k$, since $C_0 = \cdots = C_k = 0$ and $C_{k+1} \neq 0$. Recalling Definition 2.6, the consistency of a method limits the magnitude of the local truncation error perpetrated in each integration step

$$\mathrm{LTE}(t_{n+1}) = \boldsymbol{L}(y; t_{n+1}, h) \doteq C_{q+1} h^{q+1} y^{(q+1)}(t_{n+1}). \tag{2.12}$$

Due to Lemma 2.8 the local error is limited to the same order $q + 1$ in h.

Definition 2.17 *For an LMM* (2.10) *of order q, the* error constant C *reads*

$$C := C_{q+1}/\sigma(1).$$

This definition of the error constant by Henrici [72] is invariant with respect to scaling of the formula (2.10).

Consistent multistep methods do not necessarily give good approximations to the exact solution of (1.1). To limit the error propagation by the multistep method, its zero-stability plays a crucial role.

Definition 2.18 *The LMM* (2.10) *is said to be* zero-stable*, if all roots of* $\rho(\xi)$ *lie on or inside the unit circle and those on the circle are simple.*

Zero-stability ensures that local inaccuracies do not propagate in a disastrous way. Henrici [73] and Hairer et al. [67] refer to zero-stable methods as stable methods. In 1972, Cryer [45] showed the following theorem.

Theorem 2.19 *Constant BDF methods* (2.11) *are zero-stable for* $k \leq 6$ *and unstable for* $k \geq 7$.

In 1956, Dahlquist [47] proved the following fundamental convergence theorem.

Theorem 2.20 *The necessary and sufficient conditions for the LMM* (2.10) *to be convergent are that it is consistent and zero-stable.*

The proof of the sufficiency gives rise to an a priori bound on the discrepancy of the approximation y_n to the exact solution $y(t_n)$. This bound was further improved by Henrici [72] and does not suppose that the nonlinear equations in (2.10) are solved exactly. Considering particularly the BDF method (2.11), each y_{n+1} is assumed to solve a perturbed equation

$$\sum_{i=0}^{k} \alpha_i y_{n+1-i} = h\beta_0 f(t_{n+1}, y_{n+1}) + \delta_{n+1},$$

instead of (2.11), see also Definition 2.9 and rename $\check{y}_{n+1}$ by y_{n+1}.

Theorem 2.21 *Let the exact solution $y(t)$ of (1.1) be $(k+1)$-times continuously differentiable and let the stepsize h satisfy $h\left|\beta_0\alpha_0^{-1}\right| L < 1$. Then, for $t = t_n \in [t_\mathrm{s}, t_\mathrm{f}]$ fixed, the a priori bound on the global error $\mathrm{GE}(t_n) = y(t_n) - y_n$ reads*

$$|y(t_n) - y_n| \leq \Upsilon^* \cdot \exp\left(L\Upsilon^* \left|\beta_0\right| (t_n - t_\mathrm{s})\right) \cdot \left\{ Ak\varepsilon + (t_n - t_\mathrm{s}) \left(\frac{\delta}{h} + \left|C_{k+1}\right| Y h^k \right) \right\}$$

with maximal error $\varepsilon := \max_{0\leq n\leq m} |y(t_n) - y_n|$ in the start values, maximal residual $\delta := max_{n=m,\ldots,N-1} |\delta_{n+1}|$ of the nonlinear equations and $Y := \max_{t\in[t_\mathrm{s},t_\mathrm{f}]} \left|y^{(k+1)}(t)\right|$, $A := \sum_{i=0}^{k} |\alpha_i|$, $\Upsilon^* := \Upsilon/(1 - h\left|\beta_0\alpha_0^{-1}\right| L)$ where $\Upsilon := \sup_{i=0,1,\ldots} |\upsilon_i| < \infty$ and

$$\upsilon_0 + \upsilon_1\xi + \upsilon_2\xi^2 + \cdots := \frac{1}{\rho(\xi)}.$$

Proof *See Henrici [72], also for the boundedness of Υ.* □

Thus, if the nonlinear equations in (2.11) are solved exactly, i.e. $\delta = 0$ holds, the BDF method (2.11) is convergent of order $p = k$ provided that the error ε in the start values is bounded by $K_\mathrm{s}h^k$, cf. Definition 2.13. Thus, the convergence order coincides with the consistency order.

The a priori bound on the global error given by Theorem 2.21 can also be written in terms of local truncation errors and residuals of the nonlinear equations. For a system of IVPs the global error $\mathbf{GE}(t_N) = \boldsymbol{y}(t_\mathrm{f}) - \boldsymbol{y}_N$ at the final state is bounded by

$$\|\mathbf{GE}(t_N)\| \leq K \left\{ \max_{0\leq n\leq m} \|\mathbf{GE}(t_n)\| + \frac{1}{h} \left(\max_{0\leq n\leq N} \|\boldsymbol{\delta}_n\| + \max_{0\leq n\leq N} \|\mathbf{LTE}(t_n)\| \right) \right\} \tag{2.13}$$

with constant $K := \Upsilon^* \cdot \exp\left(L\Upsilon^* \left|\beta_0\right| (t_\mathrm{f} - t_\mathrm{s})\right) \max\{Ak, t_\mathrm{f} - t_\mathrm{s}\}$.

Strong stability of constant BDF methods

Another important property of LMMs is their strong stability.

Definition 2.22 *The LMM (2.10) is said to be* strongly stable, *if all roots of $\rho(\xi)$ lie inside the unit circle except for the principal root $\xi_1 = 1$.*

It can be verified by simple calculations that

Lemma 2.23 *The BDF method* (2.11) *is strongly stable.*

The following theorem will play an essential role in Section 7.1.

Theorem 2.24 *For a linear IVP of the form*

$$\dot{\boldsymbol{y}}(t) = \boldsymbol{G}(t)\boldsymbol{y}(t) + \boldsymbol{p}(t),\ \boldsymbol{y}(t_\mathrm{s}) = \boldsymbol{y}_\mathrm{s}$$

let the matrix $\boldsymbol{G}(t)$ *and the vector* $\boldsymbol{p}(t)$ *be continuously differentiable in* $t \in [t_\mathrm{s}, t_\mathrm{f}]$. *Consider the BDF method as a particular consistent and zero-stable method that is strongly stable. Furthermore, let the start values* $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{k-1}$ *generated by a starting procedure satisfy*

$$\boldsymbol{y}_n - \boldsymbol{y}_\mathrm{s} = \boldsymbol{\varepsilon}_n + \mathcal{O}(h), \quad n = 0, \ldots, k-1 \tag{2.14}$$

where the vectors $\boldsymbol{\varepsilon}_n$ *are arbitrary. Then, for* $t = t_n \in [t_\mathrm{s}, t_\mathrm{f}]$ *fixed, as* $h \to 0$,

$$\boldsymbol{y}_n = \boldsymbol{y}(t_n) + \boldsymbol{W}(t_n)\boldsymbol{\zeta} + \boldsymbol{\theta}\left(K_1 + \frac{K_2}{t_n + h - t_\mathrm{s}}\right) h$$

where $\|\boldsymbol{\theta}\| < 1$ *and* K_1, K_2 *are certain constants. The vector* $\boldsymbol{\zeta}$ *is*

$$\boldsymbol{\zeta} := \frac{1}{\rho'(1)} \sum_{i=0}^{k-1} \gamma_i\, \boldsymbol{\varepsilon}_{k-1-i},\ \text{where} \sum_{i=0}^{k-1} \gamma_i\, \xi^{k-1-i} := \frac{\rho(\xi)}{\xi - 1}$$

and $\boldsymbol{W}(t)$ *is the fundemental solution matrix of*

$$\dot{\boldsymbol{W}}(t) = \boldsymbol{G}(t)\boldsymbol{W}(t),\ \boldsymbol{W}(t_\mathrm{s}) = \boldsymbol{I}.$$

Proof *See Henrici [73].* □

This theorem describes the asymptotic behavior of the global error if the start values for a constant BDF method are error-prone independently of the stepsize h and the IVP at hand is linear.

Absolute stability of constant BDF methods

The concept of absolute stability takes care of the error propagation through the right-hand-side values $\boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})$ in (2.2b) which is not treated by the zero-stability but plays a crucial role for stiff IVPs described in Section 1.4. The absolute stability does not reduce the stepsize h to zero but rather examines the error propagation for increasing n. Absolute stability leads directly to the concepts of A- and $A(\alpha)$-stability which are crucial for stiff IVPs and guarantee moderate stepsizes also in regions of stiffness. For a general description we refer to Lambert [83] and Shampine [109].

2.3.2 Variable BDF methods

In this section we briefly consider so-called variable LMM (2.1). Their variable stepsizes and orders are the basis for the efficient solution of IVPs in practice. In the interest of brevity we just mention the most important issues and refer to the literature for more details, e.g. to Hairer et al. [67]. Firstly, we leave the order constant, i.e. $k_n = k$ in (2.1), and only vary the stepsize.

Definition 2.25 *The variable stepsize LMM* (2.1) *with constant order k is said to be of* (consistency) order q*, if*

$$\sum_{i=0}^{k} \alpha_i^{(n)} p(t_{n+1-i}) = h_n \sum_{i=0}^{k} \beta_i^{(n)} \dot{p}(t_{n+1-i})$$

holds for all polynomials $p(t)$ of degree $\leq q$ and for all partitions $\{t_n\}_{n=0}^N$.

For variable stepsize BDF methods the coefficients $\alpha_i^{(n)}$ depend on the stepsizes $h_{n+1-k}, \ldots, h_n$ according to (2.3). By construction, also the variable stepsize BDF methods are of consistency order $q = k$. If $\boldsymbol{y}(t)$ is sufficiently smooth and the stepsize ratios $\omega_i := h_i/h_{i-1}$ as well as the coefficients $\alpha_i^{(n)}$ are bounded, then the local truncation error of Definition 2.6 is limited due to the consistency order

$$\begin{aligned}\mathbf{LTE}(t_{n+1}) &= \boldsymbol{L}^{(n)}(\boldsymbol{y}; t_{n+1}, h_{n+1-k}, \ldots, h_n) \\ &= (-1)^{k+1} \frac{1}{(k+1)!} \left\{ \sum_{i=1}^{k} \left(\sum_{j=0}^{i-1} h_{n-i} \right)^{k+1} \alpha_i^{(n)} \right\} \boldsymbol{y}^{(k+1)}(t_{n+1}) + \mathcal{R} = \mathcal{O}(h_{max}^{k+1})\end{aligned} \tag{2.15}$$

for $h_{\max} := \max_n h_n$ and $h_{\max} \to 0$ due to Taylor series expansions. For constant stepsizes the above leading term coincides with (2.12).

Definition 2.26 *The variable stepsize LMM* (2.1) *with constant order k is said to be* zero-stable*, if*

$$\|\boldsymbol{A}_{n+l} \ldots \boldsymbol{A}_n\| \leq M$$

for all n and $l \geq 0$ where

$$\boldsymbol{A}_n = \begin{pmatrix} -\check{\alpha}_1^{(n)} & -\check{\alpha}_2^{(n)} & \cdots & \cdot & -\check{\alpha}_k^{(n)} \\ 1 & 0 & \cdots & \cdot & 0 \\ & 1 & & \cdot & 0 \\ & & \ddots & \vdots & \vdots \\ & & & 1 & 0 \end{pmatrix}$$

with $\check{\alpha}_i^{(n)} := \alpha_i^{(n)}/\alpha_0^{(n)}$.

A convergence theorem for variable stepsize LMMs can be found in Hairer et al. [67]. It assumes, apart from consistency, zero-stability and suitable start values, that the coefficients $\alpha_i^{(n)}$, $\beta_i^{(n)}$ and the ratios ω_n are bounded. Conditions on the variable stepsizes that guarantee zero-stability and boundedness have been studied by various authors, see e.g. Hairer et al. [67] and references therein. The coefficients $\alpha_i^{(n)}$, $\beta_i^{(n)}$ are also influenced by the order k_n such that order changes may improve the zero-stability of the variable BDF method. Further aspects and investigations can be found e.g. in Shampine [109] and references therein as well as in Gear and Watanabe [63]. The theoretical foundation in this area is still not satisfactory, but practical experiences provide suitable selection mechanisms for h_n and k_n.

2.4 Practical aspects of BDF-type methods

In this section, we describe several strategies that are important for practical implementations of BDF-type methods. We will do this by means of the variable order variable stepsize BDF method `DAESOL-II`, see Albersmeyer and Bock [5] and Albersmeyer [3]. This BDF integrator for IVPs in linear-implicit Differential Algebraic Equations (DAEs) of index one is programmed in C++ and part of the SolvIND integrator suite, see Albersmeyer and Kirches [6]. The strategies concerning the solution of IVPs in ODEs go back to Enke [54], Bleser [25] and Eich [52], the DAE-extensions to Eich [52, 53] and the derivative generation to Bauer [16] and Albersmeyer [2, 3]. In this section, we describe only those strategies concerning the solution of IVPs in ODEs that are important for the thesis at hand. The issues related to derivative generation will be addressed in Section 3.4.2. The implementation is based on the Newton representation of the interpolation polynomial and is designed to guarantee that the local truncation error is smaller than a prescribed tolerance while the computational effort is as low as possible.

2.4.1 Estimation of the local truncation error

Generally, the local truncation error of a multistep method with consistency order q can be estimated by approximating the derivative $\boldsymbol{y}^{(q+1)}$ in the leading term of $\mathbf{LTE}(t_{n+1})$ using divided differences. For BDF methods, the divided differences are in terms of $\boldsymbol{y}$.

We spend here some words on the estimation of the local truncation error as realized in `DAESOL-II` since it will play an important role in Part III of this thesis. However, the ideas have already been described several times, see e.g. Bleser [25], Eich [53], Albersmeyer [3] and Brenan et al. [38]. In the realization `DAESOL-II` of a variable order variable stepsize BDF method the local truncation error of Defini-

tion 2.6 is approximated using its two leading terms

$$\mathbf{LTE}(t_{n+1}) = -h_n \left(\frac{1}{\psi_{k_n+1}(n+1)} \mathbf{\Phi}^{\mathrm{ex}}_{k_n+2}(n+1) + \prod_{i=1}^{k_n+1} \psi_i(n+1) \nabla^{k_n+2}[\boldsymbol{y}(t_{n+1}), \boldsymbol{y}(t_{n+1}), \boldsymbol{y}(t_n), \ldots, \boldsymbol{y}(t_{n-k_n})] \right) \tag{2.16}$$

with $\psi_i(n+1) := t_{n+1} - t_{n+1-i} = h_n + \cdots + h_{n+1-i} = \psi_{i-1}(n) + h_n$, divided differences

$$\nabla^0 \boldsymbol{y}(t_n) := \boldsymbol{y}(t_n)$$

$$\nabla^{i+1}[\boldsymbol{y}(t_n), \ldots, \boldsymbol{y}(t_{n-i-1})] := \frac{\nabla^i[\boldsymbol{y}(t_n), \ldots, \boldsymbol{y}(t_{n-i})] - \nabla^i[\boldsymbol{y}(t_{n-1}), \ldots, \boldsymbol{y}(t_{n-i-1})]}{t_n - t_{n-i-1}}$$

and additionally $\nabla^1[\boldsymbol{y}(t), \boldsymbol{y}(t)] := \dot{\boldsymbol{y}}(t)$ as well as modified divided differences

$$\mathbf{\Phi}^{\mathrm{ex}}_1(n) := \boldsymbol{y}(t_n)$$
$$\mathbf{\Phi}^{\mathrm{ex}}_i(n) := \psi_1(n) \cdot \cdots \cdot \psi_{i-1}(n) \nabla^{i-1}[\boldsymbol{y}(t_n), \ldots, \boldsymbol{y}(t_{n-i+1})]$$

for $i > 1$. The equivalence of the leading terms in (2.15) and (2.16) is due to the standard interpolation theory, see e.g. Bleser [25] or Stoer and Bulirsch [116]. In practice, the exact solution values $\boldsymbol{y}(t_{n+1-i})$ for $i = 0, \ldots, k_n - 1$ in (2.16) have to be replaced by their approximations $\boldsymbol{y}_{n+1-i}$ to obtain an estimate $\widehat{\mathbf{LTE}}(t_{n+1})$ of $\mathbf{LTE}(t_{n+1})$. For constant stepsize the estimate $\mathbf{\Phi}_{k_n+2}(n+1)$ of $\mathbf{\Phi}^{\mathrm{ex}}_{k_n+2}(n+1)$ is asymptotically correct as shown by Gear [62].

In each integration step, it is guaranteed by a suitable choice of stepsize h_n and order k_n that

$$\left\| \widehat{\mathbf{LTE}}(t_{n+1}) \right\| \leq \texttt{RelTol} \tag{2.17}$$

for a user given relative tolerance `RelTol`. The selection of stepsize and order is based on a sophisticated control strategy that has its origin in Bleser [25] and has been improved by Eich [52]. This strategy uses the formula of the local truncation error for variable stepsizes to check if the proposed stepsize after a step acceptance might result in an acceptable next step. In the case of a step rejection due to a solution failure of the Newton-type method applied to the nonlinear BDF equation the control strategy incorporates the convergence behavior of the Newton-type method. Its superior performance in step numbers, order changes, step rejections, matrix updates and rebuilds has been demonstrated by Bleser [25] and Eich [52, 53]. A detailed description can also be found in Bauer [16] and Albersmeyer [3].

Remark 2.27 *In practice, a weighted norm is used instead of the Euclidean norm to regard possibly different orders of magnitude in the solution components. The weighted norm reads*

$$\|\boldsymbol{v}\|_{\boldsymbol{s}} = \sqrt{\frac{1}{d} \sum_{i=1}^{d} \left(\frac{v_i}{s_i} \right)^2}$$

where the scaling vector $\boldsymbol{s} \in \mathbb{R}^d$ *depends on the particular scaling method chosen by the user. For example, the DASSL scaling (cf. Brenan et al. [38]) realized in* `DAESOL-II` *uses* $s_i = |(\boldsymbol{y}_n)_i| + \texttt{aTol}_i/\texttt{RelTol}$ *on the subinterval* $[t_n, t_{n+1}]$ *where* `aTol` *is a user given vector of absolute tolerances and* $\boldsymbol{y}_n$ *the last accepted trajectory value. For further scaling methods of* `DAESOL-II` *we refer to Albersmeyer [3].*

Controlling the local truncation error also limits the local error of the n-th integration step since, under the Localizing Assumption, Lemma 2.8 yields

$$\|\mathbf{LE}(t_{n+1})\| \doteq \frac{1}{\alpha_0^{(n)}} \|\mathbf{LTE}(t_{n+1})\| \tag{2.18}$$

based on the Neumann series (see Theorem A.4) which theoretically supposes that $h_n/\alpha_0^{(n)} \|\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{\eta})\| < 1$. Moreover, for constant stepsize it is $\alpha_0^{(n)} = \sum_{j=1}^{k_n} 1/j$ which implies $\alpha_0^{(n)} \in [1, 2.45]$ since $k_n \leq 6$ due to Theorem 2.19. Hence, we obtain $\|\mathbf{LE}(t_{n+1})\| \leq \|\mathbf{LTE}(t_{n+1})\|$ for all orders k_n and in particular for $k_n > 1$ we gain $\|\mathbf{LE}(t_{n+1})\| \leq 2/3 \|\mathbf{LTE}(t_{n+1})\|$ since $\alpha_0^{(n)} \geq 3/2$ in this case.

2.4.2 Continuous representation

A continuous representation of the approximate IVP solution is provided by the composition of the interpolation polynomials of each BDF integration step. This section focuses on the error of the interpolation polynomial $\boldsymbol{\mathcal{P}}_{n+1}(t)$ between integration points $t \in [t_n, t_{n+1}]$. They arise from two sources: the interpolation error of the polynomial through $\boldsymbol{y}(t_{n+1-k_n}), \ldots, \boldsymbol{y}(t_{n+1})$ and the error due to the approximation of $\boldsymbol{y}(t_{n+1})$ by $\boldsymbol{y}_{n+1}$ generated by the BDF method.

Lemma 2.28 *Let* $\boldsymbol{f}(t, \boldsymbol{y})$ *be continuous in* t *and continuously differentiable in* $\boldsymbol{y}$. *Let* $\boldsymbol{y}(t)$ *be the exact solution of* (1.1) *and* $\boldsymbol{y}_{n+1}$ *determined by the* n*-th step of* (2.2) *under the Localizing Assumption. Then, it holds for* $t \in [t_n, t_{n+1}]$

$$\|\boldsymbol{u}_n(t) - \boldsymbol{\mathcal{P}}_{n+1}(t)\| \leq \left(\frac{\alpha_0^{(n)}}{4} + 1\right) \|\mathbf{LE}(t_{n+1})\| \doteq \left(\frac{1}{4} + \frac{1}{\alpha_0^{(n)}}\right) \|\mathbf{LTE}(t_{n+1})\|$$

where $\boldsymbol{u}_n(t)$ *is the exact solution of local IVP of Definition 2.4.*

Proof *Due to the Localizing Assumption and the uniqueness of the solution of* (1.1) *it is* $\boldsymbol{u}_n(t) = \boldsymbol{y}(t)$ *on* $[t_n, t_{n+1}]$, *and in particular* $\mathbf{LE}(t_{n+1}) = \boldsymbol{u}_n(t_{n+1}) - \boldsymbol{y}_{n+1} = \boldsymbol{y}(t_{n+1}) - \boldsymbol{y}_{n+1}$. *This yields*

$$\begin{aligned}
\|\boldsymbol{u}_n(t) - \boldsymbol{\mathcal{P}}_{n+1}(t)\| &= \left\| \boldsymbol{y}(t) - L_0^{(n)}(t)\, \boldsymbol{y}_{n+1} - \sum_{i=1}^{k_n} L_i^{(n)}(t)\, \boldsymbol{y}(t_{n+1-i}) \right\| \\
&= \left\| \boldsymbol{y}(t) - \sum_{i=0}^{k_n} L_i^{(n)}(t)\, \boldsymbol{y}(t_{n+1-i}) + L_0^{(n)}(t)(\boldsymbol{y}(t_{n+1}) - \boldsymbol{y}_{n+1}) \right\| \\
&\leq \left\| \boldsymbol{y}(t) - \sum_{i=0}^{k_n} L_i^{(n)}(t)\, \boldsymbol{y}(t_{n+1-i}) \right\| + \left| L_0^{(n)}(t) \right| \cdot \|\mathbf{LE}(t_{n+1})\|
\end{aligned}$$

The interpolation error is given by Theorem A.2 and bounded by

$$\begin{aligned}
&\left\| \boldsymbol{y}(t) - \sum_{i=0}^{k_n} L_i^{(n)}(t)\, \boldsymbol{y}(t_{n+1-i}) \right\| \\
&\quad = \left\| \prod_{i=0}^{k_n} (t - t_{n+1-i})\, \nabla^{k_n+1}[\boldsymbol{y}(t), \boldsymbol{y}(t_{n+1}), \ldots, \boldsymbol{y}(t_{n+1-k_n})] \right\| \\
&\quad \leq \left\| \frac{\psi_1(n+1)^2}{4} \psi_2(n+1) \cdot \cdots \cdot \psi_{k_n}(n+1)\, \nabla^{k_n+1}[\boldsymbol{y}(t), \boldsymbol{y}(t_{n+1}), \ldots, \boldsymbol{y}(t_{n+1-k_n})] \right\| \\
&\quad \approx \frac{1}{4} \left\| \mathbf{LTE}(t_{n+1}) \right\|
\end{aligned}$$

where we followed Eich [53]. Thus, the interpolation error is bounded by the local truncation error. Furthermore, $L_0^{(n)}(t)$ is strictly monotonically increasing on $[t_n, t_{n+1}]$ with maximum $L_0^{(n)}(t_{n+1}) = 1$ at t_{n+1}. Together with (2.18) *the assertions are shown.* □

Hence, the BDF polynomials provide a continuous representation of the exact solution that meets the concept of the *natural interpolation* as introduced in Bock and Schlöder [34]. At least for orders $k_n > 1$ and constant stepsize, the error of the continuous representation within the interval $[t_n, t_{n+1}]$ of the n-th integration step is bounded by $\|\mathbf{LTE}(t_{n+1})\|$ since $1/4 + 1/\alpha_0^{(n)} < 1$, cf. end of Section 2.4.1.

2.4.3 Solution of the nonlinear BDF equations

In each integration step of a BDF method (2.2), the solution $\boldsymbol{y}_{n+1}$ of the nonlinear BDF equations (2.2b), i.e.

$$\boldsymbol{\mathcal{F}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1}) := \alpha_0^{(n)} \boldsymbol{y}_{n+1-i} - h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) + \sum_{i=1}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} = \mathbf{0}, \tag{2.19}$$

has to be found for given past values $\boldsymbol{y}_{n+1-k_n}, \ldots, \boldsymbol{y}_n$. This system of equations possesses a unique solution $\boldsymbol{y}_{n+1}$ if stepsize h_n and order k_n are chosen such that (2.6) holds. In practical implementations, the nonlinear BDF equation (2.19) is solved iteratively. Hence, a start value for the iteration has to be *predicted* and then *corrected* to approximate the solution of (2.19). The correction can either be done by fix point iteration or by a Newton-type method. The fix point iteration imposes less computational effort per integration step, but may take very small steps especially for stiff IVPs. Although the computational effort of a Newton-type method is bigger, its convergence does not directly depend on stepsize h_n and stiffness.

To predict a start value $\boldsymbol{y}_{n+1}^{(0)}$ for the Newton-type method the interpolation polynomial $\boldsymbol{\mathcal{P}}_{n+1}^{\mathrm{P}}$ of degree k_n through the past values $\boldsymbol{y}_{n-k_n}, \ldots, \boldsymbol{y}_n$ is evaluated at t_{n+1}

$$\boldsymbol{y}_{n+1}^{(0)} = \boldsymbol{y}_{n+1}^{\mathrm{P}} := \boldsymbol{\mathcal{P}}_{n+1}^{\mathrm{P}}(t_{n+1}).$$

This predicted value $\boldsymbol{y}_{n+1}^{\mathrm{P}}$ is, for sufficiently small h_n, near the exact solution $\boldsymbol{y}_{n+1}^{*}$ of (2.19), hence the start value $\boldsymbol{y}_{n+1}^{(0)} = \boldsymbol{y}_{n+1}^{\mathrm{P}}$ should lie inside the local convergence region of the Newton-type method. The iterates then are $\boldsymbol{y}_{n+1}^{(i+1)} = \boldsymbol{y}_{n+1}^{(i)} + \Delta\boldsymbol{y}_{n+1}^{(i)}$ with increments

$$\Delta\boldsymbol{y}_{n+1}^{(i)} = -\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}\left(\boldsymbol{y}_{n+1}^{(i)}\right)^{-1}\boldsymbol{\mathcal{F}}_{\mathrm{BDF}}^{(n)}\left(\boldsymbol{y}_{n+1}^{(i)}\right)$$

for $i = 0, \ldots, s_n - 1$ where the Jacobian $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}$ of $\boldsymbol{\mathcal{F}}_{\mathrm{BDF}}^{(n)}$ is given by

$$\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}) = \alpha_0^{(n)}\boldsymbol{I} - h_n\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}).$$

In practice, the performance of a few iterations per integration step using an approximation of the inverse of $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\cdot)$ has turned out to be sufficient to get efficiently an approximation to the solution of IVP (1.1), cf. Gear [61], Enke [54] and Eich [52].

The Newton-type method implemented in `DAESOL-II` is based on a sophisticated *monitor strategy* that guarantees the convergence of the method while the efficiency is controlled using a hierarchical update procedure for the iteration matrix approximating the BDF Jacobian $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\cdot)$. This monitor strategy is based on the *Local Contraction Theorem* of Bock [31] (see Theorem 2.29). It goes back to Enke [54] and has been improved by Eich [52] who also demonstrated its particular efficiency if the Jacobian of $\boldsymbol{f}$ varies only slowly and is expensive to evaluate. It is also described by Eich [53], Bauer [16] and Albersmeyer [3].

In each integration step, the Newton-type method perfoms at most three iterations with constant approximation $\boldsymbol{M}_n$ of $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\cdot)$. The method is considered as converged if the increment fulfills $||\Delta\boldsymbol{y}_{n+1}^{(s_n-1)}|| <$ `NTol` for a prescribed Newton tolerance `NTol`. If two iterations are performed, the convergence rate δ_0 of the Newton-type method can be estimated with (2.20) by $\hat{\delta}_0 := ||\Delta\boldsymbol{y}_{n+1}^{(1)}||/||\Delta\boldsymbol{y}_{n+1}^{(0)}||$. If $\delta_0 < 0.25$, we have gained in the last iterate $\boldsymbol{y}_{n+1}^{(2)}$ more than one digit compared to the first increment since

$$\left\|\boldsymbol{y}_{n+1}^{(2)} - \boldsymbol{y}_{n+1}^{*}\right\| \leq \frac{\delta_0^2}{1-\delta_0}\left\|\Delta\boldsymbol{y}_{n+1}^{\mathrm{P}}\right\| \leq \frac{1}{12}\left\|\Delta\boldsymbol{y}_{n+1}^{\mathrm{P}}\right\|$$

due to the a priori estimate of the Local Contraction Theorem. The third iteration is performed if $0.25 \leq \delta_0 < 0.3$ and also gives one digit more accuracy. Von Schwerin [121] gave a formula for the bounds on δ_0 as a function of desired digits and iterations. Overall, the final approximation of $\boldsymbol{y}_{n+1}^{*}$ is provided by $\boldsymbol{y}_{n+1}^{(s_n)} = \boldsymbol{y}_{n+1}^{(s_n-1)} + \Delta\boldsymbol{y}_{n+1}^{(s_n-1)}$ with $s_n \in \{1, 2, 3\}$. The Newton tolerance `NTol` is chosen to be `NTol` $= 0.08 \cdot$ `RelTol`.

If this Newton-type method does not converge, a hierarchical update procedure for the iteration matrix $\boldsymbol{M}_n$ is used. The cheapest way to improve $\boldsymbol{M}_n \approx \alpha_0^{(n)}\boldsymbol{I} - h_n\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1}^{\mathrm{P}})$ is to insert $\alpha_0^{(n)}$, h_n and to *decompose* the resulting matrix. If still no convergence is achieved, then the whole matrix is *rebuilt*, including the evaluation of $\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1}^{\mathrm{P}})$ and the subsequent decomposition. The last option is to reject the

step and reduce the stepsize to improve the first iterate $\boldsymbol{y}_{n+1}^{\mathrm{P}}$.

Now, we state the Local Contraction Theorem of Bock [31] that forms the basis of the monitor strategy described above. Furthermore, it will be used in Section 8.5.

Theorem 2.29 (Local Contraction Theorem) *Let be $\boldsymbol{f} : \mathcal{D} \to \mathbb{R}^n$, $\mathcal{D} \subset \mathbb{R}^n$ and $\boldsymbol{f} \in C^1(\mathcal{D})$. The Jacobian of $\boldsymbol{f}$ is denoted by $\boldsymbol{J}(\boldsymbol{y}) = \boldsymbol{f}_{\boldsymbol{y}}(\boldsymbol{y})$ and $\boldsymbol{A}^{-1}$ is an approximation of $\boldsymbol{J}^{-1}$. A root $\boldsymbol{y}^*$ of $\boldsymbol{f}$ is in demand.*
If for all $\boldsymbol{y}', \boldsymbol{y} \in \mathcal{D}$, $\tau \in [0,1]$ and $\boldsymbol{y}' - \boldsymbol{y} = -\boldsymbol{A}^{-1}(\boldsymbol{y})\boldsymbol{f}(\boldsymbol{y}) = \Delta\boldsymbol{y}$ there exist $\omega < \infty$ and $\kappa < 1$ such that

1. *The generalized Lipschitz condition on $\boldsymbol{J}$ and $\boldsymbol{A}^{-1}$ holds*

$$\left\|\boldsymbol{A}^{-1}(\boldsymbol{y}')\left[\boldsymbol{J}(\boldsymbol{y}+\tau(\boldsymbol{y}'-\boldsymbol{y})) - \boldsymbol{J}(\boldsymbol{y})\right](\boldsymbol{y}'-\boldsymbol{y})\right\| \leq \omega\tau\left\|\boldsymbol{y}'-\boldsymbol{y}\right\|^2.$$

2. *The compatibility condition on $\boldsymbol{A}^{-1}$ holds*

$$\left\|\boldsymbol{A}^{-1}(\boldsymbol{y}')\left[\boldsymbol{I} - \boldsymbol{J}(\boldsymbol{y})\boldsymbol{A}^{-1}(\boldsymbol{y})\right]\boldsymbol{f}(\boldsymbol{y})\right\| \leq \kappa\left\|\boldsymbol{y}'-\boldsymbol{y}\right\|.$$

3. *The start value $\boldsymbol{y}^{(0)}$ of the iteration fulfills $\delta_0 < 1$ where*

$$\delta_i := \kappa + \frac{\omega}{2}\left\|\Delta\boldsymbol{y}^{(i)}\right\|.$$

4. *The closed ball $\mathcal{D}_0 := \left\{\boldsymbol{y} \in \mathbb{R}^n : \left\|\boldsymbol{y}-\boldsymbol{y}^{(0)}\right\| \leq \left\|\Delta\boldsymbol{y}^{(0)}\right\|/(1-\delta_0)\right\}$ lies in $\mathcal{D}$.*

Then, the iterates $\boldsymbol{y}^{(i+1)} = \boldsymbol{y}^{(i)} + \Delta\boldsymbol{y}^{(i)}$ with $\Delta\boldsymbol{y}^{(i)} = -\boldsymbol{A}^{-1}(\boldsymbol{y}^{(i)})\boldsymbol{f}(\boldsymbol{y}^{(i)})$ satisfy

1. *$\boldsymbol{y}^{(i)}$ are well-defined and $\boldsymbol{y}^{(i)} \in \mathcal{D}_0$.*

2. *There exists $\boldsymbol{y}^* \in \mathcal{D}_0$ and the sequence $\{\boldsymbol{y}^{(i)}\}_{i=0}^{\infty}$ converges to $\boldsymbol{y}^*$ with the rate*

$$\left\|\Delta\boldsymbol{y}^{(i+1)}\right\| \leq \delta_i\left\|\Delta\boldsymbol{y}^{(i)}\right\| = \kappa\left\|\Delta\boldsymbol{y}^{(i)}\right\| + \frac{\omega}{2}\left\|\Delta\boldsymbol{y}^{(i)}\right\|^2. \tag{2.20}$$

3. *The a priori error estimate holds*

$$\left\|\boldsymbol{y}^{(i+j)} - \boldsymbol{y}^*\right\| \leq \frac{(\delta_i)^j}{1-\delta_i}\left\|\Delta\boldsymbol{y}^{(i)}\right\| \leq \frac{(\delta_0)^{i+j}}{1-\delta_0}\left\|\Delta\boldsymbol{y}^{(0)}\right\|.$$

4. *The limit $\boldsymbol{y}^*$ fulfills $\boldsymbol{A}^{-1}(\boldsymbol{y}^*)\boldsymbol{f}(\boldsymbol{y}^*) = \boldsymbol{0}$. Moreover, if $\boldsymbol{A}^{-1}(\boldsymbol{y})$ is continuous and nonsingular in $\boldsymbol{y}^*$, then $\boldsymbol{f}(\boldsymbol{y}^*) = \boldsymbol{0}$.*

Proof *See Bock [31].* □

We refer the reader to Potschka [101] for a detailed presentation of theorem and proof.

3 Computing adjoint derivatives of IVP solutions

This chapter starts with a description of methods to evaluate derivatives of functions. In the second part we describe two ways to obtain derivatives of Initial Value Problem (IVP) solutions: Firstly, the variational IVPs are solved, and secondly the Internal Numerical Differentiation (IND) of integrators as invented by Bock [28] is presented. The system of equations resulting from adjoint IND of Backward Differentiation Formula (BDF) methods is derived explicitly before we briefly describe the efficient realization of adjoint IND in `DAESOL-II`. Thereafter, we compare the two approaches from a conceptional point of view before we focus on the discrete adjoint IND values and their relation to the solution of the adjoint IVP.

3.1 Derivative generation for functions

In this section, we briefly review different ways to obtain derivatives of differentiable functions. Derivatives are required, e.g., for the solution of the nonlinear BDF equation (2.2b) where the derivative $\boldsymbol{f_y}(t, \boldsymbol{y})$ of the right hand side with respect to $\boldsymbol{y}$ is needed, cf. Section 2.4.3. The function itself can be available as a computer-evaluated function or as analytical expression. Derivatives can be obtained analytically, numerically or algorithmically.

3.1.1 Analytical differentiation

Analytical derivatives can either be determined by hand or by computer algebra systems. Although this procedure gives derivative values that are exact up to machine precision, it has several drawbacks. The differentiation by hand is time-consuming and error-prone. The usage of computer algebra systems like Maple [90] and Mathematica [102] may lead to derivatives that are expensive to evaluate since common subexpressions in different terms are not exploited appropriately.

3.1.2 Approximation using numerical schemes

Derivatives of functions $\boldsymbol{g}\colon \mathbb{R}^{n_{\text{in}}} \to \mathbb{R}^{n_{\text{out}}}$ in direction $\boldsymbol{d} \in \mathbb{R}^{n_{\text{in}}}$ with $\|\boldsymbol{d}\| = 1$ can be approximated by one-sided finite differences

$$\boldsymbol{g}'(\boldsymbol{x})\ \boldsymbol{d} = \frac{\boldsymbol{g}(\boldsymbol{x} + s\boldsymbol{d}) - \boldsymbol{g}(\boldsymbol{x})}{s} + \mathcal{O}(s) \tag{3.1}$$

based on Taylor series expansions. Unfortunately, the approximations of the derivatives are subject to truncation errors for large increments s and subject to cancellation errors for tiny increments. And even the optimal increment size leads, under standard assumptions on $\boldsymbol{g}$, to derivative approximations that have lost half of the significant digits of the function evaluation. To overcome the cancellation errors a complex step approximation of the first-order derivative can be used, see Lyness and Moler [89]. A short comparison of the numerical schemes can be found, for example, in Albersmeyer [3].

3.1.3 Algorithmic Differentiation

Another way to evaluate derivatives of computer-evaluated functions is to use Algorithmic Differentiation (AD) techniques. In this section, we briefly review the main ideas of AD and refer to the textbook by Griewank [65] for a comprehensive description of the topic.

The basis of AD is the decomposition of a computer-evaluated function into a sequence of certain elemental functions like $+, -, *, /, \exp$, etc. that are continuously differentiable. Thus, the function evaluation can be described by a computational graph of the elemental functions. The edges of the computational graph represent the elemental functions whereas the nodes represent the intermediate results or intermediate values. Then, the principle of AD is to systematically apply the Chain Rule of Calculus to the elemental functions of the computational graph. There exist two distinct ways of AD: The *forward mode* traverses the graph from the input variables towards the output variables whereas the *adjoint mode* traverses the other way round. Both modes give the derivative up to machine precision since the exact derivatives of the elemental functions are known.

Forward mode of AD This mode computes efficiently the forward directional derivative of the computer-evaluated function $\boldsymbol{g}\colon \mathbb{R}^{n_{\text{in}}} \to \mathbb{R}^{n_{\text{out}}}$ in forward direction $\dot{\boldsymbol{x}} \in \mathbb{R}^{n_{\text{in}}}$ evaluated at $\boldsymbol{x} \in \mathbb{R}^{n_{\text{in}}}$

$$\boldsymbol{g}'(\boldsymbol{x})\, \dot{\boldsymbol{x}} \in \mathbb{R}^{n_{\text{out}} \times 1}$$

where $\mathbb{R}^{n_{\text{out}} \times 1}$ denotes that it is a column vector. The numerical effort for evaluating the function and computing p_{fwd} directional derivatives by the forward mode is theoretically bounded by $(1 + 1.5 p_{\text{fwd}})$ times the effort for function evaluation, see Griewank [65]. Thus, it is recommended particularly when $p_{\text{fwd}} \ll n_{\text{out}}$. In this mode, each intermediate value describes the derivative of the corresponding intermediate function value with respect to the input variables given by the direction.

Adjoint mode of AD This mode is also called reverse or backward mode. It computes the adjoint directional derivative of $\boldsymbol{g}$ in adjoint direction $\bar{\boldsymbol{y}} \in \mathbb{R}^{n_{\text{out}}}$ evaluated at $\boldsymbol{x} \in \mathbb{R}^{n_{\text{in}}}$

$$\bar{\boldsymbol{y}}^{\intercal}\, \boldsymbol{g}'(\boldsymbol{x}) \in \mathbb{R}^{1 \times n_{\text{in}}}$$

where $\mathbb{R}^{1\times n_{\text{in}}}$ denotes that it is a row vector with n_{in} entries. Since the adjoint mode traverses backwards through the graph it has to be preceded by a function evaluation with storage of the intermediate results. The overall numerical effort (including the storage) for evaluating the function and computing p_{bwd} directional derivatives by the adjoint mode is theoretically bounded by $(1.5 + 2.5p_{\text{bwd}})$ times the effort for function evaluation, see Griewank [65]. Thus, it is recommended particularly when $p_{\text{bwd}} \ll n_{\text{in}}$. In this mode, each intermediate value describes the derivative of the output variables given by the direction with respect to the particular intermediate function value.

3.2 Solution of variational Initial Value Problems

Let the function $\tilde{\boldsymbol{y}}(t)$ be an approximation to the solution of IVP (1.1). With the techniques described in Section 3.1 the variational IVPs (1.4) and (1.6) along $\tilde{\boldsymbol{y}}(t)$ can be set up and solved by any integration method. This gives an approximation to the solution of the perturbed variational IVP and hence an approximation to the derivative of the nominal IVP solution $\boldsymbol{y}(t)$. This procedure is called *continuous* or *differentiate-then-discretize* approach. It assumes that a continuous approximation $\tilde{\boldsymbol{y}}(t)$ to the solution $\boldsymbol{y}(t)$ already exists such that the matrix $\boldsymbol{f_y}(t, \tilde{\boldsymbol{y}}(t))$ is continuous in t. The derivative $\boldsymbol{f_y}(t, \boldsymbol{y})$ is obtained by one of the approaches of Section 3.1. There exist several codes following this procedure of integrating the variational IVPs, e.g. the multistep integrators IDAS and CVODES of the SUNDIALS suite, cf. Hindmarsh et al. [74]. In the forward mode the mentioned implementations solve simultaneously the IVPs (1.1) and (1.4), see Li et al. [86] as well as Serban and Hindmarsh [108]. In the adjoint mode they first solve (1.1) and then separately the adjoint IVP (1.6) along a piecewise interpolant through the discrete nominal approximations, see Cao et al. [42] as well as Serban and Hindmarsh [108].

3.3 Internal Numerical Differentiation

Another way to approximate derivatives of the solution $\boldsymbol{y}(t)$ of IVP (1.1) is based on the level of integrators. One might think in approximating the derivative of the solution of (1.1) with respect to initial values by solving (1.1) also for perturbed initial values using any integrator and finite differences (cf. Section 3.1.2) afterwards. This approach is known as External Numerical Differentiation (END) and uses the integrator as a "black box". It implicitly assumes that both integration outputs are computed sufficiently accurate such that (3.1) is accurate enough. Hence, either the integration effort is very high or the derivative is not that accurate. To overcome these difficulties IND was proposed by Bock [28].

Principle of Internal Numerical Differentiation A more sophisticated way to approximate the required derivatives is provided by IND that was first presented by Bock [28]. The *basic principle of IND* is to differentiate the calculation rule used to

obtain the approximations $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ where the calculation rule itself is generated by any adaptive integrator. Hence, after the nominal integration of (1.1), the adaptive components of the integrator are kept constant and the *fixed discretization scheme* is differentiated.

There exist different variants to realize the IND principle. To describe them we restrict ourselves to linear integration methods. The first variant, called *varied trajectories*, uses the fixed discretization scheme to solve (1.1) for perturbed initial values. The perturbed solutions are then used in finite differences (3.1) to give first-order approximations of the derivative, cf. Bock [28]. The (theoretical) performance of the limit in the perturbation size of the initial values would yield the same output as one would obtain by using the fixed discretization scheme to solve the forward variational IVP (1.4) provided that the partial derivative $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y})$ is available up to machine precision (Section 3.1.1 and 3.1.3), cf. Section 3.2 and Remark A.5. This is the so-called *analytical limit of IND* shown by Bock [30].

IND can also be applied in *adjoint* mode which was first described by Bock [31] for Runge-Kutta integrators and later on in Bock et al. [35] for BDF methods. Adjoint IND differentiates the fixed discretization scheme backwards in time starting at the final time.

In the case of the implicit BDF method, IND exists in different variants. One variant is *direct* IND which neglects the residuals caused by the approximate solution of the nonlinear BDF equations (2.2b) and makes use of the Implicit Function Theorem, cf. Section 3.4.2. Hence, it can be understood to assume that (2.2b) are solved exactly. We describe it in-depth in Section 3.4 for the adjoint IND mode. Another variant of IND is the following: it also differentiates the iterations of the Newton-type method and reuses the iteration matrices, cf. Section 2.4.3. This *iterative* IND applies AD techniques to the fixed discretization scheme and gives the exact derivatives of the nominal approximations $\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_N$ (up to machine precision). It is crucial not to apply AD techniques to the control mechanism determining the adaptive components, see also Eberhard and Bischof [51]. For algorithmical details and a comparison of the computational effort of the different forward IND variants of BDF methods we refer to Bauer [16], whereas details and comparisons of the different adjoint IND variants are given in Albersmeyer and Bock [5] and Albersmeyer [3].

Overall, the concept of IND is to differentiate the discretization scheme that was used to solve the nominal IVP. Hence, this procedure belongs to the so-called *discrete* or *discretize-then-differentiate* approaches.

3.4 Adjoint IND of BDF methods

3.4.1 Direct adjoint IND of BDF methods

In this section, we specify the direct adjoint IND of BDF methods. For given adaptive components the direct IND approach can be understood to assume that the

nonlinear BDF equations (2.2b) are solved exactly. Hence, it coincides with applying adjoint differentiation to the nominal BDF method (2.2) with prescribed stepsizes $\{h_n\}_{n=0}^{N-1}$ and orders $\{k_n\}_{n=0}^{N-1}$. Adjoint IND for BDF methods was first described by Bock et al. [35] and later by Sandu [106] as *reverse automatic differentiation* of BDF methods.

Lemma 3.1 *For a variable BDF method* (2.2) *with self-starting procedure, after freezing the adaptive components* $\{h_n\}_{n=0}^{N-1}$ *and* $\{k_n\}_{n=0}^{N-1}$*, the discrete adjoint IND scheme in adjoint direction* $\boldsymbol{r} = J'(\boldsymbol{y}_N)$ *is given by*

$$\alpha_0^{(N-1)} \boldsymbol{\lambda}_N - J'(\boldsymbol{y}_N)^\intercal = h_{N-1} \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t_N, \boldsymbol{y}_N) \boldsymbol{\lambda}_N \tag{3.2a}$$

$$\sum_{\substack{0 \le i \le N-1-n \\ i \le k_{\max}}} \alpha_i^{(n+i)} \boldsymbol{\lambda}_{n+1+i} = h_n \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t_{n+1}, \boldsymbol{y}_{n+1}) \boldsymbol{\lambda}_{n+1}, \ n = N-2, \dots, 0 \tag{3.2b}$$

$$\boldsymbol{\lambda}_0 + \alpha_1^{(0)} \boldsymbol{\lambda}_1 = \boldsymbol{0} \tag{3.2c}$$

with the convention that $\alpha_i^{(n)} = 0$ *for* $i > k_n$ *and* $k_{\max} = \max_n\{k_n\}$.

Proof *See Section A.2.1.* □

Note that due to Theorem 2.19 it is always $k_{\max} \le 6$. The variables $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ are the derivatives of $J(\boldsymbol{y}_N)$ with respect to each intermediate integration step (2.2b) due to AD, see also Section 3.1.3. Thus, they describe the sensitivity of the finite dimensional system of equations (2.2). The entity $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ is also called *discrete stability* in contrast to the continuous stability $\boldsymbol{\lambda}(t)$ of the IVP (1.1) explained in Section 1.3 or on page 48.

3.4.2 Practical aspects of adjoint IND of BDF-type methods

A more efficient version of *direct* adjoint IND is presented in Albersmeyer and Bock [5] as well as in Albersmeyer [3]. It can be derived from the following equivalent domain space formulation of the n-th integration step (2.2b) of the BDF method

$$\boldsymbol{y}_{n+1} = \boldsymbol{\theta}_{n+1}(\boldsymbol{y}_n, \dots, \boldsymbol{y}_{n+1-k_n}) \tag{3.3a}$$

where $\boldsymbol{\theta}_{n+1}$ is defined implicitly as solution of the nonlinear root finding problem

$$\begin{aligned} &\boldsymbol{F}(\boldsymbol{y}_{n+1-k_n}, \dots, \boldsymbol{y}_n, \boldsymbol{\theta}_{n+1}(\boldsymbol{y}_n, \dots, \boldsymbol{y}_{n+1-k_n})) \\ &\quad := \alpha_0^{(n)} \boldsymbol{\theta}_{n+1} + \sum_{i=1}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} - h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{\theta}_{n+1}) = \boldsymbol{0} \end{aligned} \tag{3.3b}$$

for $n = 0, \dots, N-1$. Using the Implict Function Theorem, the associated adjoint IND scheme reads as follows.

Lemma 3.2 *For a variable BDF method with self-starting procedure written in the domain space form* (3.3)*, after freezing the adaptive components* $\{h_n\}_{n=0}^{N-1}$ *and* $\{k_n\}_{n=0}^{N-1}$*, the discrete adjoint IND scheme in adjoint direction* $\boldsymbol{r} = J'(\boldsymbol{y}_N)$ *is given by*

$$\bar{\boldsymbol{y}}_N = J'(\boldsymbol{y}_N)^\intercal \tag{3.4a}$$

$$\bar{\boldsymbol{y}}_{n+1} = -\sum_{\substack{1\le i\le N-1-n \\ i\le k_{\max}}} \alpha_i^{(n+i)} \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n+i)}(\boldsymbol{y}_{n+1+i})^{-\intercal}\, \bar{\boldsymbol{y}}_{n+1+i},\ n = N-2,\ldots,0 \tag{3.4b}$$

$$\bar{\boldsymbol{y}}_0 = -\alpha_1^{(0)} \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(0)}(\boldsymbol{y}_1)^{-\intercal} \bar{\boldsymbol{y}}_1 \tag{3.4c}$$

with the convention that $\alpha_i^{(n)} = 0$ *for* $i > k_n$ *and* $k_{\max} = \max_n\{k_n\} \le 6$*.*

Proof *See Section A.2.1.* □

The relation of the adjoint IND values $\bar{\boldsymbol{y}}_{n+1}$ and $\boldsymbol{\lambda}_{n+1}$ is described by the following lemma.

Lemma 3.3 *Let* $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ *be generated by* (3.2) *and* $\{\bar{\boldsymbol{y}}_n\}_{n=0}^N$ *by* (3.4)*. Then, they are related by*

$$\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})^\intercal \boldsymbol{\lambda}_{n+1} = \bar{\boldsymbol{y}}_{n+1},\ n = N-1,\ldots,0$$
$$\boldsymbol{\lambda}_0 = \bar{\boldsymbol{y}}_0.$$

with the Jacobian $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1}) = \alpha_0^{(n)} \boldsymbol{I} - h_n \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})$ *of the nonlinear BDF equation* (2.2b)*.*

Proof *See Section A.2.1.* □

The *iterative* adjoint IND scheme as presented in Albersmeyer and Bock [5] and Albersmeyer [3] uses also the above formulation (3.3) of the BDF integration step. Hence, it is the iterative analogon to (3.4). The iterative variant is more efficient than the direct variant since it gets along without building and decomposing the BDF Jacobian $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})$ in every integration step. It just needs adjoint directional derivatives of $\boldsymbol{f}(t, \boldsymbol{y}_{n+1})$.

The iterative adjoint IND scheme realized in `DAESOL-II` is based on the univariate Taylor Coefficient propagation of AD (see Griewank [65]). This propagation is also used to generate efficiently directional forward and forward/adjoint derivatives of arbitrary order of IVP solutions. For details we refer to Albersmeyer [3].

Both adjoint IND values $\boldsymbol{\lambda}_0$ and $\bar{\boldsymbol{y}}_0$ at t_s are the exact adjoint derivatives of J at the computed final solution $\boldsymbol{y}_N$ (up to machine precision). The adjoint value $\bar{\boldsymbol{y}}_0$ is successfully used in efficient direct methods for the solution of Optimal Control Problems (OCPs), cf. Albersmeyer [3]. Moreover, both values $\boldsymbol{\lambda}_0$ and $\bar{\boldsymbol{y}}_0$ at t_s converge to the exact adjoint solution $\boldsymbol{\lambda}(t_\mathrm{s})$ with the same rate as $\boldsymbol{y}_N$ to $\boldsymbol{y}(t_\mathrm{f})$ due to Remark A.6 and Lemma 3.3, or see also Bock [31] for general linear integration methods.

3.5 Discretize-then-differentiate approach vs. Differentiate-then-discretize approach

In this section we focus on the advantages and disadvantages of the continuous and discrete approaches described in Section 3.2 and 3.3, 3.4 to obtain derivatives of IVP solutions with respect to initial values. We also highlight the desirable property that both approaches are in agreement. We concentrate on adjoint differentiation since the case of forward differentiation is already treated satisfactorily by the analytical limit of IND in forward mode, cf. Bock [30] or Section 3.3 and Remark A.5.

Discretize-then-differentiate approach The advantages of this approach include its straightforward and generic applicability to any IVP. Moreover, after the problem definition, the procedure processes automatically. The generated adjoint derivatives are the exact derivatives of the nominal approximations, and hence they are the proper quantities from the discrete point of view. If used in efficient direct methods for OCPs based on shooting, this is crucial for the convergence of the inexact Sequential Quadratic Programming (SQP) method to solve the Nonlinear Program (NLP). If, on the other hand, approximations to the adjoint solutions are required, the disadvantage of this approach is that the generated adjoint derivatives may not provide an adequate approximation to the adjoint IVP in a straightforward manner.

Differentiate-then-discretize approach These methods first differentiate the IVP at hand, and then discretize the resulting (nominal and adjoint) problems to approximate their solutions. For the numerical solution of the combined unstable problem (cf. Section 1.3), one needs good initial guesses for quantities that may not have apparent physical interpretations, i.e. the adjoints. The quality of the initial guess highly influences the convergence behavior (if there is convergence at all) of the numerical procedure. These disadvantages are confronted by the advantage that the numerical procedure approximates the (nominal and adjoint) solutions up to its inherent order of accuracy.

Commutativity as desirable property In order to benefit from the advantages of both approaches, it is desirable that they lead to the same discrete systems of equations. In the case of Runge-Kutta methods with non-zero weights, the discrete adjoint scheme generated by adjoint IND is itself a Runge-Kutta scheme for the adjoint IVP (1.8), and thus gives a convergent approximation to the adjoint solution as shown by Bock [28] and later by Walther [122] and Sandu [104]. In the case of continuous and discontinuous Galerkin methods applied to (1.1), the discrete adjoint schemes yield approximations to the solution of (1.8), see e.g. Johnson [77]. The situation becomes significantly more complex in the case of multistep methods, as the discrete adjoint IND schemes of Linear Multistep Methods (LMMs) are generally *not* consistent with the adjoint IVP (1.8). We will discuss this in the next section.

3.6 Adjoint IND vs. solution of adjoint IVP

In this section we focus on the adjoint IND values $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ generated by the adjoint IND scheme (3.2) and compare them to the solution $\boldsymbol{\lambda}(t)$ of the adjoint IVP (1.8). The discrete adjoint IND value $\boldsymbol{\lambda}_0$ shows the same convergence behavior towards its counterpart $\boldsymbol{\lambda}(t_\mathrm{s})$ like the approximate solution $\boldsymbol{y}_N$ generated by the nominal BDF method towards $\boldsymbol{y}(t_\mathrm{s})$, cf. Remark A.6 or Bock [31] or Sandu [106] for constant LMMs. This does not hold for $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ compared to $\boldsymbol{\lambda}(t)$ at intermediate points. It has been treated phenomenologically by Albersmeyer [3] and to a certain theoretical extent by Sandu [106].

To investigate not only the adjoint at t_s but rather on the whole interval $[t_\mathrm{s}, t_\mathrm{f}]$, we define a perturbed adjoint IVP along a sufficiently smooth approximation $\tilde{\boldsymbol{y}}(t)$ of the IVP solution $\boldsymbol{y}(t)$ by

$$\dot{\tilde{\boldsymbol{\lambda}}}(t) = -\boldsymbol{f}_{\boldsymbol{y}}^{\intercal}\left(t, \tilde{\boldsymbol{y}}(t)\right) \tilde{\boldsymbol{\lambda}}(t), \quad t \in [t_\mathrm{s}, t_\mathrm{f}] \tag{3.5a}$$

$$\tilde{\boldsymbol{\lambda}}(t_\mathrm{f}) = J'\left(\tilde{\boldsymbol{y}}(t_\mathrm{f})\right)^{\intercal}. \tag{3.5b}$$

The exact solutions of (1.8) and (3.5) can be given explicitly since both IVPs are linear. Hence, it can be shown that their distance in the $C^0[t_\mathrm{s}, t_\mathrm{f}]^d$-norm (see Section 4.2) is

$$\left\| \boldsymbol{\lambda}(t) - \tilde{\boldsymbol{\lambda}}(t) \right\|_{C^0[t_\mathrm{s}, t_\mathrm{f}]^d} \leq K \left\| \boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t) \right\|_{C^0[t_\mathrm{s}, t_\mathrm{f}]^d} \tag{3.6}$$

for a constant K. To investigate the relation between $\boldsymbol{\lambda}_n$ and $\boldsymbol{\lambda}(t_n)$ for $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ generated by (3.2), we first consider the relation between $\boldsymbol{\lambda}_n$ and $\tilde{\boldsymbol{\lambda}}(t_n)$ and then make use of (3.6). To this end, recall the general form of an LMM given in Definition 2.1.

Lemma 3.4 *For a variable BDF method with self-starting procedure, the associated adjoint IND scheme* (3.2) *is an LMM applied to the perturbed adjoint IVP* (3.5) *provided that* $\tilde{\boldsymbol{y}}(t)$ *satisfies* $\tilde{\boldsymbol{y}}(t_n) = \boldsymbol{y}_n$ *for* $n = 0, \ldots, N$.

Proof *For* $N - 2 \geq n \geq 0$, *the* n*-th step of* (3.2b) *can be written as*

$$\sum_{\substack{0 \leq i \leq N-1-n \\ i \leq k_{\max}}} \alpha_i^{(n+i)} \boldsymbol{\lambda}_{n+1+i} = -h_{n+1} \frac{h_n}{h_{n+1}} \left[-\boldsymbol{f}_{\boldsymbol{y}}^{\intercal}(t_{n+1}, \boldsymbol{y}_{n+1}) \boldsymbol{\lambda}_{n+1} \right].$$

It proceeds from t_{n+2} *to* t_{n+1} *with stepsize* $-h_{n+1}$ *and determines the new approximation* $\boldsymbol{\lambda}_{n+1}$ *using the past values* $\boldsymbol{\lambda}_{n+2}, \cdots, \boldsymbol{\lambda}_{n+1+k_n}$. *The right hand side of* (3.5a) *is evaluated at* $(t_{n+1}, \boldsymbol{\lambda}_{n+1})$ *and multiplied by* $\beta_0^{(n)} = h_n / h_{n+1}$, *hence* $\beta_i^{(n)} = 0$ *for* $i > 0$. *Equation* (3.2a) *proceeds from* t_{N+1} *to* t_N *with stepsize* $-h_{N-1}$ *and uses* $\boldsymbol{\lambda}_{N+1} := J'(\boldsymbol{y}_N)$ *and* $\beta_0^{(N)} = 1$. *Equation* (3.2c) *proceeds from* t_1 *to* t_0 *with stepsize* $-h_1$ *and* $\beta_0^{(-1)} = 0$ *(explicit LMM step). Hence,* (3.2) *is an implicit LMM applied to* (3.5) *with an explicit last step, cf. Definition 2.1.* □

As seen in Section 2.3 consistency and zero-stability of LMMs are the essential properties for their convergence provided that the start errors are small.

Lemma 3.5 *Consider a constant BDF method with order k, stepsize h and sufficiently accurate self-starting procedure for $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ and $m \geq k$ fixed. Then, for the associated adjoint IND scheme* (3.2) *holds that*

1. *Adjoint initialization steps:* (3.2a) *and* (3.2b) *with $n = N-2, \ldots, N-k+1$ are inconsistent.*
2. *Adjoint main steps:* (3.2b) *with $n = N-k, \ldots, m$ are consistent of order k with* (3.5)*, and asymptotically consistent with the adjoint IVP* (1.8)*.*
3. *Adjoint termination steps:* (3.2b) *with $n = m-1, \ldots, 0$ and* (3.2c) *are inconsistent.*

Proof *Due to the consistency with order k of the nominal constant BDF method, the coefficients $\alpha_i^{(n+1)} = \alpha_i$ of* (3.2b) *with $n = N-k, \ldots, m$ satisfy the requirements for consistency order k with* (3.5)*, cf. Section 2.3.1. Taking $\tilde{\boldsymbol{y}}(t)$ to be the continuous representation resulting from* (2.5) *it is $\|\boldsymbol{y}(t) - \tilde{\boldsymbol{y}}(t)\|_{C^0[t_\mathrm{s}, t_\mathrm{f}]^d} \leq ch^k$, cf. Theorem 2.21 or Shampine and Zhang [112] and additionally Section 2.4.2. Hence, the solution $\tilde{\boldsymbol{\lambda}}(t)$ of* (3.5) *converges with order k to the solution $\boldsymbol{\lambda}(t)$ of* (1.8)*. The α-coefficients of all other steps do not sum up to zero and hence the formulas are inconsistent, cf. Section 2.3.1.* □

Remark 3.6 *If the stepsizes in the self-starting procedure of a constant BDF method vary, then the adjoint main steps are given by* (3.2b) *for $n = N-k, \ldots, m+k-1$.*

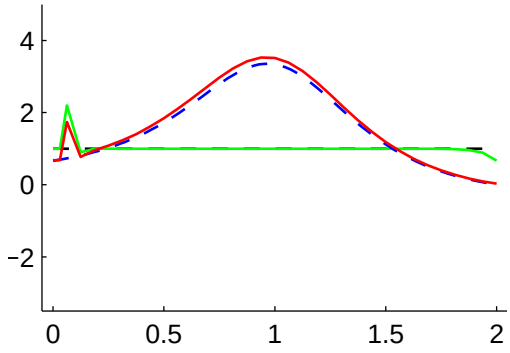
Lemma 3.7 *For a variable BDF method the associated adjoint IND scheme* (3.2) *is inconsistent with* (3.5)*.*

Proof *The α_i-coefficients of the LMM* (3.2) *do not sum up to zero and hence* (3.2) *is inconsistent with* (3.5)*, cf. Section 2.3.2.* □

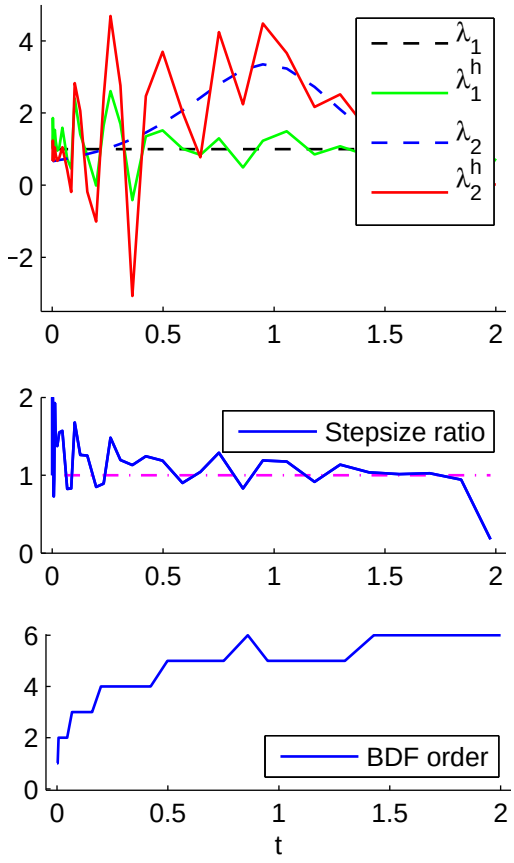
This consistency behavior of adjoint IND schemes has been observed also by Sandu [105, 106]. However, zero-stability (see Definition 2.26) of the nominal BDF method (2.2) with constant order k implies zero-stability of the adjoint IND scheme (3.2) as shown by Sandu [105, 106].

But note that also for constant BDF methods the inconsistency of the adjoint IND initialization steps results in start errors that are of order zero in h, see end of Section 7.1. Hence, the convergence Theorem 2.21 is not applicable. Nevertheless, in Section 7.1 we will demonstrate convergence on the open interval $(t_\mathrm{s}, t_\mathrm{f})$.

The properties of the discrete adjoint IND scheme (3.2) as highlighted above in theory can also be observed numerically. We use the Catenary problem as a nonlinear test case with analytic nominal and adjoint solution, see Section 10.1. Furthermore, we use the adjoint IND scheme of a constant BDF method with order 2 and that of the variable BDF method `DAESOL-II`. The results are depicted in Figure 3.1.



(a) Constant BDF method



(b) Variable BDF method

Figure 3.1: Comparison of the discrete adjoint IND values $\boldsymbol{\lambda}^h = [\lambda_1^h, \lambda_2^h]^\intercal$ and the analytic solution $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]^\intercal$ of the adjoint IVP on the Catenary test case. Stepsize ratio (penultimate row) and BDF order (bottom) of the variable BDF method.

4 Elements of real and functional analysis

This chapter reviews some basic concepts of real and functional analysis, respectively, that are of significance in the progress of this thesis. At this point, we assume that the reader is familiar with the concepts of Riemann- and Lebesgue-integrals, see, for example, Rudin [103] and Kolmogorov and Fomin [79].

4.1 Functions of bounded variation and the Riemann-Stieltjes integral

This section is devoted to the definition of the Riemann-Stieltjes integral. For this, we first need the notion of a function of bounded variation. We follow the presentations of Kolmogorov and Fomin [79] and Natanson [97].

4.1.1 Functions of bounded variation

Definition 4.1 *A* partition *of the interval $[a,b]$ is a finite set $\mathcal{T}$ of $m+1$ points such that $a=\tau_0<\cdots<\tau_m=b$. The* size *of partition $\mathcal{T}$ is defined by $|\mathcal{T}|:=m$ and the* fineness *of $\mathcal{T}$ by $h(\mathcal{T}):=\max_{1\le j\le m}(\tau_j-\tau_{j-1})$. The set of all partitions on $[a,b]$ is denoted by $\mathcal{T}([a,b])$.*

Definition 4.2 *A function Φ defined on $[a,b]$ is said to be* of bounded variation *if there exists a constant $C>0$ such that*

$$\sum_{j=1}^{|\mathcal{T}|}|\Phi(\tau_j)-\Phi(\tau_{j-1})|\le C$$

for every partition $\mathcal{T}\in\mathcal{T}([a,b])$.

For later use, we already define the total variation.

Definition 4.3 *The* total variation *of a function Φ on $[a,b]$ is given by*

$$V_a^b(\Phi):=\sup_{\mathcal{T}\in\mathcal{T}([a,b])}\sum_{j=1}^{|\mathcal{T}|}|\Phi(\tau_j)-\Phi(\tau_{j-1})|\,.$$

Lemma 4.4 *For an integrable function φ on $[a,b]$, the indefinite integral*

$$\Phi(t)=\int_a^t\varphi(\tau)\,\mathrm{d}\tau$$

is a function of bounded variation on $[a,b]$.

Proof *The constant C limiting the variation of Φ is given by the total variation of Φ*

$$\begin{aligned} V_a^b(\Phi) &= \sup_{\mathcal{T}\in\mathcal{T}([a,b])} \sum_{j=1}^{|\mathcal{T}|} \left| \int_{\tau_{j-1}}^{\tau_j} \varphi(\tau)\,\mathrm{d}\tau \right| \\ &\le \sup_{\mathcal{T}\in\mathcal{T}([a,b])} \sum_{j=1}^{|\mathcal{T}|} \left\{ (\tau_j - \tau_{j-1}) \operatorname*{ess\,sup}_{\tau\in[\tau_{j-1},\tau_j]} |\varphi(\tau)| \right\} \\ &\le (b-a) \operatorname*{ess\,sup}_{\tau\in[a,b]} |\varphi(\tau)| =: C < \infty \end{aligned}$$

since φ is integrable on $[a,b]$. □

We use here the definition of the jump function in such a way that it is continuous from the right. This will be of importance later on. But generally one could also assume continuity from the left for the considerations of the current section.

Definition 4.5 *Let $h_1, h_2, \ldots, h_n, \ldots$ be numbers corresponding to at most countably many discontinuity points $t_1, t_2, \ldots, t_n, \ldots$ in $[a,b]$ that satisfy*

$$\sum_n |h_n| < \infty.$$

Then, the function

$$\Phi(t) = \sum_{\{n\,:\,t_n \ge t\}} h_n$$

is called a jump function. *Moreover, if $t_1 < t_2 < \cdots < t_n < \cdots$, then Φ is called a* step function.

4.1.2 Riemann-Stieltjes integral

Definition 4.6 *Let f, Φ be two functions defined on $[a,b]$ with Φ being of bounded variation. Furthermore, let*

$$\mathcal{T}^1 \subset \mathcal{T}^2 \subset \cdots \subset \mathcal{T}^k \subset \cdots$$

be a sequence of refined partitions $\mathcal{T}^k \in \mathcal{T}([a,b])$ such that $h(\mathcal{T}^k) \to 0$ as $k \to \infty$, and let $\theta_j^k \in [\tau_{j-1}^k, \tau_j^k]$ be arbitrary, $j = 1, \ldots, |\mathcal{T}^k|$. If the sum

$$\sum_{j=1}^{|\mathcal{T}^k|} f(\theta_j^k)[\Phi(\tau_j^k) - \Phi(\tau_{j-1}^k)]$$

approaches for $k \to \infty$ a limit independently of the choice of the partition $\mathcal{T}^k$ and the points θ_j^k, then this limit is called the Riemann-Stieltjes integral *of the integrand f with respect to the generating function Φ and is denoted by*

$$\int_a^b f(t)\,\mathrm{d}\Phi(t).$$

Theorem 4.7 *If f is continuous on $[a,b]$, then its Riemann-Stieltjes integral exists.*

Proof *See Kolmogorov and Fomin [79].* □

The Riemann-Stieltjes integral is linear in both, the integrand $f = f_1 + f_2$ and the integrator $\Phi = \Phi_1 + \Phi_2$, provided that each integral $\int_a^b f_i(t)\,\mathrm{d}\Phi_j(t)$ exists.

Lemma 4.8 *Let be $c \in (a,b)$.*

1. *If the integral $\int_a^b f(t)\,\mathrm{d}\Phi(t)$ exists, then also the integrals $\int_a^c f(t)\,\mathrm{d}\Phi(t)$ and $\int_c^b f(t)\,\mathrm{d}\Phi(t)$ exist.*
2. *If $\int_a^c f(t)\,\mathrm{d}\Phi(t)$, $\int_c^b f(t)\,\mathrm{d}\Phi(t)$ and $\int_a^b f(t)\,\mathrm{d}\Phi(t)$ exist, then it holds*

$$\int_a^b f(t)\,\mathrm{d}\Phi(t) = \int_a^c f(t)\,\mathrm{d}\Phi(t) + \int_c^b f(t)\,\mathrm{d}\Phi(t).$$

Proof *See Natanson [97].* □

Note that Assertion 1 of Lemma 4.8 can not be inverted. To overcome this obstacle we will introduce in Section 5.3.1 an appropriate extension of the Riemann-Stieltjes integral.

Theorem 4.9 *If f is continuous on $[a,b]$, and Φ possesses an integrable and bounded derivative $\Phi'(t)$ in every $t \in [a,b]$, then it holds*

$$\int_a^b f(t)\,\mathrm{d}\Phi(t) = \int_a^b \Phi'(t) f(t)\,\mathrm{d}t.$$

Proof *See Natanson [97].* □

Lemma 4.10 *If f is continuous on $[a,b]$ and Φ is a jump function given by Definition 4.5, then the Riemann-Stieltjes integral reduces to a sum*

$$\int_a^b f(t)\,\mathrm{d}\Phi(t) = \sum_n h_n f(t_n).$$

Proof *For Φ having a single jump at t_1 of height h_1, the Riemann-Stieltjes integral is*

$$\begin{aligned}\int_a^b f(t)\,\mathrm{d}\Phi(t) &= \lim_{k\to\infty} \sum_{j=1}^{|\mathcal{T}^k|} f(\theta_j^k)[\Phi(\tau_j^k) - \Phi(\tau_{j-1}^k)] \\ &= f(t_1)[\Phi(t_1^+) - \Phi(t_1^-)] = f(t_1)h_1\end{aligned}$$

since the only remaining addend is $f(\theta_j^k)[\Phi(\tau_j^k) - \Phi(\tau_{j-1}^k)]$ with $\tau_{j-1}^k < t_1 \le \tau_j^k$. If Φ exhibits countably many jumps, then the Riemann-Stieltjes integral is the sum of these jump heights h_n multiplied by the corresponding integrand values $f(t_n)$. □

4.2 Function spaces and their properties

This section first defines normed, Banach and Hilbert spaces, respectively, and focuses subsequently on those function spaces that are of importance for this thesis. Details concerning the first part of this section can be found, for example, in Wloka [126], Gajewski et al. [60] and Alt [7].

Definition 4.11 *A pair* $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ *of a linear space* $\mathbb{X}$ *and a norm* $\|\cdot\|_{\mathbb{X}}$ *on* $\mathbb{X}$ *is called a* normed space*.*

Definition 4.12 *A subset* A *of a normed space* $\mathbb{X}$ *is said to be* dense *in* $\mathbb{X}$ *if each element* $x \in \mathbb{X}$ *is the limit of a sequence of elements in* A*.*

Definition 4.13 *For normed spaces* $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ *and* $(\mathbb{Y}, \|\cdot\|_{\mathbb{Y}})$ *the space* $\mathcal{L}(\mathbb{X}, \mathbb{Y})$ *consists of all continuous linear operators* $\mathfrak{A}$ *from* $\mathbb{X}$ *to* $\mathbb{Y}$*. It is a normed space with*

$$\|\mathfrak{A}\|_{\mathcal{L}(\mathbb{X},\mathbb{Y})} := \sup_{\|x\|_{\mathbb{X}}=1} \|\mathfrak{A}(x)\|_{\mathbb{Y}}\,.$$

If $\mathfrak{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ *is bijective, then* $\mathfrak{A}^{-1} \in \mathcal{L}(\mathbb{Y}, \mathbb{X})$ *and* $\mathfrak{A}$ *is said to be an* isomorphism*. If* $\|\mathfrak{A}(x)\|_{\mathbb{Y}} = \|x\|_{\mathbb{X}}$ *for all* $x \in \mathbb{X}$*, then* $\mathfrak{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ *is said to be an* isometry*.*

The spaces $\mathbb{X}$ and $\mathbb{Y}$ are isometrically isomorphic, in symbols $\mathbb{X} \cong \mathbb{Y}$, if there exists an isometric isomorphism between $\mathbb{X}$ and $\mathbb{Y}$. Such spaces are of great interest since they have identical structures and only the nature of their elements differs.

If $\mathbb{X}_1, \ldots, \mathbb{X}_d$ are finitely many normed spaces with norms $\|\cdot\|_{\mathbb{X}_1}, \ldots, \|\cdot\|_{\mathbb{X}_d}$, then the finite Cartesian product space $\mathbb{X} := \mathbb{X}_1 \times \cdots \times \mathbb{X}_d$ is a normed space, see Wloka [126]. We always choose $\|\boldsymbol{x}\|_{\mathbb{X}} = \max_{1\le i\le d} \|x_i\|_{\mathbb{X}_i}$ as norm except for the definition of the condition number in Section 1.3.1 where we have taken $\|\boldsymbol{x}\|_{\mathbb{X}} = \sum_{i=1}^{d} \|x_i\|_{\mathbb{X}_i}$.

Definition 4.14 *A normed space* $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ *is called a* Banach space*, if it is complete with respect to* $\|\cdot\|_{\mathbb{X}}$*, i.e. every Cauchy sequence in* $\mathbb{X}$ *has a limit in* $\mathbb{X}$*.*

Definition 4.15 *A pair* $(\mathbb{X}, (\cdot,\cdot)_{\mathbb{X}})$ *of a linear space* $\mathbb{X}$ *and a scalar product* $(\cdot,\cdot)_{\mathbb{X}}$ *on* $\mathbb{X}$ *is called a* pre-Hilbert space*.*

In a pre-Hilbert space a norm can be introduced by $\|x\|_{\mathbb{X}} = \sqrt{(x,x)_{\mathbb{X}}}$ such that it is always a normed space.

Definition 4.16 *A pre-Hilbert space* $(\mathbb{X}, (\cdot,\cdot)_{\mathbb{X}})$ *is called a* Hilbert space*, if it is complete with respect to the introduced norm* $\|\cdot\|_{\mathbb{X}} = \sqrt{(\cdot,\cdot)_{\mathbb{X}}}$*.*

In the second part of this section we consider some particular function spaces that are for importance for Part II of this thesis.

The space $C^0[a,b]$ of all continuous functions on $[a,b]$ equipped with the norm $\|f\|_{C^0[a,b]} = \max_{t\in[a,b]} |f(t)|$ is a Banach space, cf. Wloka [126]. The space $C^1[a,b]$

of all continuously differentiable functions on $[a,b]$ is a Banach space with respect to the norm $\|f\|_{C^1[a,b]} = \max_{t\in[a,b]} |f(t)| + \max_{t\in[a,b]} |f'(t)|$, cf. Gajewski et al. [60]. The space $C^1_{\mathrm{b}}(a,b]$ of all continuously differentiable and bounded functions with bounded derivatives is a Banach space with respect to the norm $\|f\|_{C^1_{\mathrm{b}}(a,b]} = \sup_{t\in(a,b]} |f(t)| + \sup_{t\in(a,b]} |f'(t)|$, cf. Adams and Fournier [1].

We next consider the spaces of functions of bounded variation. The space of all functions of bounded variation on $[a,b]$ is denoted by $\mathrm{BV}[a,b]$ and can be equipped with the norm $\|\Phi\|_{\mathrm{BV}[a,b]} = |\Phi(a)| + V_a^b(\Phi)$ where $V_a^b(\Phi)$ is the total variation of Definition 4.3, cf. Luenberger [88]. But of more importance for this thesis will be the following space.

Definition 4.17 *The* normalized space of all functions of bounded variation *is denoted by* $\mathrm{NBV}[a,b]$ *and consists of all functions of bounded variation on* $[a,b]$ *that vanish at the point* a *and are continuous from the right on* (a,b)*. It is equipped with the norm* $\|\Phi\|_{\mathrm{NBV}[a,b]} = V_a^b(\Phi)$*.*

Remark 4.18 *We have chosen here the normalization of* $\mathrm{BV}[a,b]$ *with respect to continuity from the right. Generally, continuity from the left could also be assumed. But for our purpose continuity from the right is more convenient. This will become clear in Chapter 5 and 6.*

The space $\mathrm{NBV}[a,b]$ with norm $\|\cdot\|_{\mathrm{NBV}[a,b]}$ is a Banach space, cf. Kolmogorov and Fomin [79].

Finally, we come to the spaces of Lebesgue-integrable functions. The space $L^2(a,b)$ of all quadratically Lebesgue-integrable functions is a Hilbert space with respect to the scalar product

$$(f,g)_{L^2(a,b)} = \int_{(a,b)} f(t)g(t)\,\mathrm{d}t.$$

The Sobolev space $H^1(a,b)$ of all $L^2(a,b)$-functions with weak derivative in $L^2(a,b)$ is also a Hilbert space with the appropriate scalar product. For details we refer to Adams and Fournier [1].

4.3 Dual spaces and linear functionals

This section starts with dual spaces and linear functionals in general and gives representations for dual spaces of particular function spaces.

Definition 4.19 $\mathbb{X}' := \mathcal{L}(\mathbb{X},\mathbb{R})$ *is called the* dual space *of a normed space* $\mathbb{X}$*. The elements* $\mathfrak{L}$ *of* $\mathbb{X}'$ *are called* linear functionals*.*

The norm $\|\cdot\|_{\mathbb{X}'}$ of the dual space $\mathbb{X}'$ is given by

$$\|\mathfrak{L}\|_{\mathbb{X}'} = \|\mathfrak{L}\|_{\mathcal{L}(\mathbb{X},\mathbb{R})} = \sup_{\|x\|_{\mathbb{X}}=1} |\mathfrak{L}(x)| .$$

For the dual of a finite Cartesian product space the following theorem holds.

Theorem 4.20 *Let $\mathbb{X} = \mathbb{X}_1 \times \cdots \times \mathbb{X}_d$ be a finite Cartesian product space of normed spaces $(\mathbb{X}_i, \|\cdot\|_{\mathbb{X}_i})$ with norm $\|\boldsymbol{x}\|_{\mathbb{X}} = \max_{1\le i\le d} \|x_i\|_{\mathbb{X}_i}$. Then, the continuous linear functionals $\boldsymbol{\mathfrak{L}}$ on $\mathbb{X}$ are given by*

$$\boldsymbol{\mathfrak{L}}(\boldsymbol{x}) = \sum_{i=1}^{d} \mathfrak{L}_i(x_i)$$

where $\mathfrak{L}_i$ are the continuous linear functionals on $\mathbb{X}_i$, $i = 1, \ldots, d$. In other words, the dual space of $\mathbb{X}$ is $\mathbb{X}' = \mathbb{X}_1' \times \cdots \times \mathbb{X}_d'$ with norm

$$\|\boldsymbol{\mathfrak{L}}\|_{\mathcal{L}(\mathbb{X},\mathbb{R})} = \max_{1\le i\le d} \|\mathfrak{L}_i\|_{\mathcal{L}(\mathbb{X}_i,\mathbb{R})} .$$

Proof *See Wloka [126].* □

For linear functionals the following important extension theorem holds.

Theorem 4.21 (Hahn-Banach Extension Theorem) *Let $(\mathbb{X}, \|\cdot\|_{\mathbb{X}})$ be a normed space and $\mathbb{G} \subset \mathbb{X}$ be a closed linear subspace of $\mathbb{X}$ with the same norm $\|\cdot\|_{\mathbb{X}}$. Furthermore, let $\mathfrak{L} \in \mathbb{G}'$ be a linear functional on $\mathbb{G}$. Then, $\mathfrak{L}$ can be extended to a linear functional $\widehat{\mathfrak{L}}$ on $\mathbb{X}$ preserving the norm, i.e. $\widehat{\mathfrak{L}}\big|_{\mathbb{G}} = \mathfrak{L}$ and $||\widehat{\mathfrak{L}}||_{\mathbb{X}'} = \|\mathfrak{L}\|_{\mathbb{G}'} < \infty$.*

Proof *See Wloka [126].* □

On the other hand, functionals can also be restricted to subspaces. Consider two Banach spaces $\mathbb{X}$ and $\mathbb{Y}$ with $\mathbb{X} \subset \mathbb{Y}$, $\mathbb{X}$ is dense in $\mathbb{Y}$ and $\|x\|_{\mathbb{Y}} \le c\,\|x\|_{\mathbb{X}}$ for all $x \in \mathbb{X}$ and c constant. Then, it holds that $\mathbb{Y}' \subset \mathbb{X}'$ (but not necessarily dense) and $\|\mathfrak{L}\|_{\mathbb{X}'} \le c\,\|\mathfrak{L}\|_{\mathbb{Y}'}$ for all $\mathfrak{L} \in \mathbb{Y}'$, see Gajewski et al. [60].

Definition 4.22 *For a normed space $\mathbb{X}$, the mapping $\langle\cdot,\cdot\rangle_{\mathbb{X}',\mathbb{X}} : \mathbb{X}' \times \mathbb{X} \to \mathbb{R}$ given by*

$$\langle x', x\rangle_{\mathbb{X}',\mathbb{X}} := \mathfrak{L}(x)$$

is called duality pairing *of $\mathbb{X}'$ and $\mathbb{X}$.*

From Chapter 5 on the dual spaces of the continuous and the quadratically Lebesgue-integrable functions, respectively, are of fundamental importance.

Theorem 4.23 (Riesz Representation Theorem) *Let $\mathfrak{L}$ be a continuous linear functional on $C^0[a,b]$, i.e. $\mathfrak{L} \in \big(C^0[a,b]\big)' = \mathcal{L}\big(C^0[a,b],\mathbb{R}\big)$. Then, there exists a unique function $\Phi \in \mathrm{NBV}[a,b]$ such that for all $f \in C^0[a,b]$ holds*

$$\mathfrak{L}(f) = \int_a^b f(t)\,\mathrm{d}\Phi(t) \tag{4.1}$$

and moreover $\|\mathfrak{L}\|_{\mathcal{L}(C^0[a,b],\mathbb{R})} = \|\Phi\|_{\mathrm{NBV}[a,b]}$.

Proof *See Luenberger [88].* □

Remark 4.24 *The uniqueness of Φ in Theorem 4.23 only holds if the normalized space $\mathrm{NBV}[a,b]$ of $\mathrm{BV}[a,b]$ is used.*

Thus, the dual of $C^0[a,b]$ is isometrically isomorphic to the normalized space of all functions of bounded variation, i.e. $\big(C^0[a,b]\big)' = \mathcal{L}\big(C^0[a,b],\mathbb{R}\big) \cong \mathrm{NBV}[a,b]$. The duality pairing takes the form

$$\langle \Phi, f\rangle_{\mathrm{NBV}[a,b],C^0[a,b]} = \int_a^b f(t)\,\mathrm{d}\Phi(t).$$

The duals of Hilbert spaces exhibit a canonical structure.

Theorem 4.25 *Let $(\mathbb{X},(\cdot,\cdot)_{\mathbb{X}})$ be a Hilbert space. Then, for every continuous linear functional $\mathfrak{L} \in \mathbb{X}'$ there exists an element $g \in \mathbb{X}$ such that for all $f \in \mathbb{X}$ holds*

$$\mathfrak{L}(f) = (f,g)_{\mathbb{X}}$$

and moreover $\|\mathfrak{L}\|_{\mathbb{X}'} = \|g\|_{\mathbb{X}}$.

Proof *See Wloka [126].* □

Hence, the dual of a Hilbert space is isometrically isomorphic to the Hilbert space itself and the duality pairing coincides with the scalar product. For the quadratically Lebesgue-integrable functions on (a,b) we have $\big(L^2(a,b)\big)' \cong L^2(a,b)$ and

$$\langle g, f\rangle_{L^2(a,b),L^2(a,b)} = (g,f)_{L^2(a,b)} = \int_{(a,b)} g(t) f(t)\,\mathrm{d}t.$$

For the duals of the finite Cartesian products $C^0[a,b]^d$ and $L^2(a,b)^d$ we introduce the following notation

$$\int_a^b \boldsymbol{f}(t)\,\mathrm{d}\boldsymbol{\Phi}(t) := \sum_{i=1}^d \int_a^b f_i(t)\,\mathrm{d}\Phi_i(t),$$

$$\int_a^b \boldsymbol{g}^{\mathsf{T}}(t)\boldsymbol{f}(t)\,\mathrm{d}t := \sum_{i=1}^d \int_a^b g_i(t) f_i(t)\,\mathrm{d}t,$$

which is in accordance with Theorem 4.20.

4.4 Differentiability in Banach spaces

In this section let $\mathbb{X}$ and $\mathbb{Y}$ be Banach spaces, $\mathbb{U} \subset \mathbb{X}$ be open and $\mathfrak{M} : \mathbb{U} \to \mathbb{Y}$ be a given mapping. Details on this section can be found, for example, in Zeidler [130].

Definition 4.26 *If for two elements $x \in \mathbb{U}$ and $\delta x \in \mathbb{X}$ the limit*

$$\mathfrak{M}'(x)(\delta x) := \lim_{s \searrow 0} \frac{\mathfrak{M}(x + s\delta x) - \mathfrak{M}(x)}{s} \in \mathbb{Y}$$

exists, then $\mathfrak{M}'(x)(\delta x)$ is called directional derivative *of $\mathfrak{M}$ at x in direction δx and the mapping $\delta x \mapsto \mathfrak{M}'(x)(\delta x)$* first variation *of $\mathfrak{M}$ at x. If the limit exists for all $\delta x \in \mathbb{X}$, then $\mathfrak{M}$ is called* directionally differentiable *at x.*

Note that the directional derivative, if it exists, is not necessarily linear in the direction.

Definition 4.27 *A directionally differentiable mapping $\mathfrak{M} : \mathbb{U} \to \mathbb{Y}$ is called* Gâteaux differentiable *at $x \in \mathbb{U}$, if $\mathfrak{M}'(x)$ is a continuous linear mapping from $\mathbb{X}$ to $\mathbb{Y}$, i.e. $\mathfrak{M}'(x) \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$.*

Thus, the Gâteaux derivative of a functional $\mathfrak{M} : \mathbb{X} \to \mathbb{R}$ is an element of the dual space $\mathbb{X}' = \mathcal{L}(\mathbb{X}, \mathbb{R})$, i.e.

$$\mathfrak{M}'(x)(\delta x) = \left\langle \mathfrak{M}'(x), \delta x \right\rangle_{\mathbb{X}', \mathbb{X}}.$$

Definition 4.28 *A Gâteaux differentiable mapping $\mathfrak{M} : \mathbb{U} \to \mathbb{Y}$ is called* Fréchet differentiable *at $x \in \mathbb{U}$, if $\mathfrak{M}'(x)$ satisfies*

$$\lim_{\|\delta x\|_{\mathbb{X}} \to 0} \frac{\|\mathfrak{M}(x + \delta x) - \mathfrak{M}(x) - \mathfrak{M}'(x)(\delta x)\|_{\mathbb{Y}}}{\|\delta x\|_{\mathbb{X}}} = 0.$$

Thus, the Fréchet differentiability of $\mathfrak{M}$ at $x \in \mathbb{U}$ states that

$$\mathfrak{M}(x + \delta x) - \mathfrak{M}(x) = \mathfrak{M}'(x)(\delta x) + o(\|\delta x\|_{\mathbb{X}}), \quad \delta x \to 0,$$

i.e. it reflects the concept of linear approximations and agrees with the (total) differentiability of functions on finite dimensional spaces. Like in the finite dimensional case, the Fréchet differentiability of $\mathfrak{M}$ at x implies continuity of $\mathfrak{M}$ at x, cf. Zeidler [130]. Throughout the whole thesis, we consider Fréchet differentiability of mappings between Banach spaces. Hence, we restrict the subsequent considerations to Fréchet derivatives.

Higher-order Fréchet derivatives are constructed successively. For example, the second Fréchet derivative of a functional $\mathfrak{M} : \mathbb{X} \to \mathbb{R}$ is constructed as follows: The mapping $x \mapsto \mathfrak{M}'(x)$ going from $\mathbb{X}$ to $\mathcal{L}(\mathbb{X}, \mathbb{R})$ is differentiated at x in direction $\overline{\delta x}$ to give the second Fréchet derivative $\mathfrak{M}''(x) \in \mathcal{L}(\mathbb{X}, \mathcal{L}(\mathbb{X}, \mathbb{R}))$.

The basic theorems of finite dimensional differential calculus can be generalized to Fréchet differentiable mappings between Banach spaces, see Zeidler [130], as well as the concept of partial derivatives. For Banach spaces $\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y}$ let the mapping $\mathfrak{M} : \mathbb{D} \subset \mathbb{X}_1 \times \mathbb{X}_2 \to \mathbb{Y}$ be given by $(x_1, x_2) \mapsto \mathfrak{M}(x_1, x_2)$.

Definition 4.29 *If, for fixed x_2, the mapping $\mathfrak{N}(x_1) = \mathfrak{M}(x_1, x_2)$ has a Fréchet derivative at $x_1 \in \mathbb{D}$, then $\mathfrak{M}_{x_1}(x_1, x_2) = \mathfrak{N}'(x_1)$ is called the* partial Fréchet derivative *of $\mathfrak{M}$ at (x_1, x_2) with respect to x_1.*

The Fréchet derivative with respect to x_2 is defined similarly such that for the Fréchet differentiability of $\mathfrak{M}$ the following lemma holds.

Lemma 4.30 *If $\mathfrak{M} : \mathbb{D} \subset \mathbb{X}_1 \times \mathbb{X}_2 \to \mathbb{Y}$ is Fréchet differentiable at (x_1, x_2), then the partial Fréchet derivatives $\mathfrak{M}_{x_1}$ and $\mathfrak{M}_{x_2}$ exist at (x_1, x_2) and it holds for all $\delta x_1 \in \mathbb{X}_1$ and $\delta x_2 \in \mathbb{X}_2$ that*

$$\mathfrak{M}'(x_1, x_2)(\delta x_1, \delta x_2) = \mathfrak{M}_{x_1}(x_1, x_2)(\delta x_1) + \mathfrak{M}_{x_2}(x_1, x_2)(\delta x_2).$$

Conversely, if $\mathfrak{M}_{x_1}$ and $\mathfrak{M}_{x_2}$ exist in a neighborhood of (x_1, x_2) and are continuous at (x_1, x_2), then $\mathfrak{M}$ is Fréchet differentiable and the above equality holds.

Proof *See Zeidler [130].* □

Throughout this thesis the mapping $\boldsymbol{g}\colon C^1[a,b]^d \to C^0[a,b]^d$ defined by

$$\boldsymbol{g}(\boldsymbol{y}(\cdot)) := \dot{\boldsymbol{y}}(\cdot) - \boldsymbol{f}(\cdot, \boldsymbol{y}(\cdot)),$$

where $\boldsymbol{f}(t, \boldsymbol{y})$ is the right hand side of (1.1a), plays a central role. Since we generally require that $\boldsymbol{f}(t, \boldsymbol{y})$ is continuous in t and continuously differentiable in $\boldsymbol{y}$ (cf. Section 1.2), $\boldsymbol{g}$ is Fréchet differentiable at $\boldsymbol{y}(\cdot)$ in direction $\boldsymbol{v}(\cdot)$ with Fréchet derivative

$$\boldsymbol{g}'(\boldsymbol{y}(\cdot))(\boldsymbol{v}(\cdot)) = \dot{\boldsymbol{v}}(\cdot) - \boldsymbol{f}_{\boldsymbol{y}}(\cdot, \boldsymbol{y}(\cdot))\boldsymbol{v}(\cdot),$$

see, for example, Ioffe and Tihomirov [75]. If the mapping $\boldsymbol{g}$ is defined on Lebesgue spaces, i.e. $\boldsymbol{g}\colon H^1(a,b)^d \to L^2(a,b)^d$, it is also Fréchet differentiable with the above derivative.

Part II

A novel interpretation for discrete adjoints of BDF methods

5 Weak adjoint solutions

In this chapter we derive a novel functional-analytic framework for Initial Value Problems (IVPs) in Ordinary Differential Equations (ODEs). With this framework and results of Chapter 6 and Chapter 7 we shed light on the unknown relation between the discrete adjoint Internal Numerical Differentiation (IND) values of Backward Differentiation Formula (BDF) methods and the solution of the adjoint IVP, see Section 3.6. For that purpose, we set up a particular Constrained Variational Problem (CVP), investigate its infinite dimensional optimality conditions in different function spaces and introduce the notion of weak adjoint solutions. For our theoretical investigations, we embed the IVP (1.1) into an artificial optimization framework and derive the adjoint IVP as part of the first-order necessary optimality conditions. To this end, we consider the CVP

$$\min_{\boldsymbol{y}} \; J(\boldsymbol{y}(t_\mathrm{f})) \tag{5.1a}$$

$$\text{s. t. } \; \dot{\boldsymbol{y}}(t) = \boldsymbol{f}(t, \boldsymbol{y}(t)), \quad t \in [t_\mathrm{s}, t_\mathrm{f}] \tag{5.1b}$$

$$\boldsymbol{y}(t_\mathrm{s}) = \boldsymbol{y}_\mathrm{s} \tag{5.1c}$$

which is equivalent to evaluating $J(\boldsymbol{y}(t_\mathrm{f}))$ in the solution of (1.1). The feasible set of (5.1) consists of a single element, namely the unique solution of the nominal IVP (1.1), cf. Section 1.1.

This chapter is organized in three parts, where each part is dedicated to the solution of the CVP in a particular function space. The first part identifies the adjoint given by (1.8) with the Lagrange multiplier of the CVP in a functional-analytic setting in a Hilbert space. This part describes the main ideas of the procedure. Secondly, we carry over the procedure in a more general way to the solution of CVP (5.1) in the space of continuously differentiable functions. In this setting, the Lagrange multiplier is an element of the space of normalized functions with bounded variation. This setting is still not enough to analyze BDF methods and their discrete adjoint IND schemes since BDF methods give continuous, piecewise continuously differentiable approximations to the solution of IVP (1.1). To capture this case we finally extend the trial space to the continuous, piecewise continuously differentiable functions. Solving (5.1) on the latter space requires an appropriate extension of the Riemann-Stieltjes integral.

For reasons of completeness we include here large parts of Beigel et al. [21]. Modifications are conducted to refer to other parts of this thesis and to keep the unified structure of the thesis.

5.1 Classical adjoint as Lagrange multiplier in $L^2(t_\mathrm{s}, t_\mathrm{f})^d$

The core of this section is the identification of the adjoint as the Lagrange multiplier of the CVP in a functional-analytic setting. The basic ideas described in this section are of course not new. However, the setting for the case of ODEs is fundamental for this contribution. Since we have not found a comprehensive presentation in the literature, we include here a detailed derivation.

Solving the CVP (5.1) in the space $H^1(t_\mathrm{s}, t_\mathrm{f})^d$, we note that $\dot{\boldsymbol{y}}(\cdot) - \boldsymbol{f}(\cdot, \boldsymbol{y}(\cdot))$ is an element of $L^2(t_\mathrm{s}, t_\mathrm{f})^d$. Thus, the Lagrangian $\mathcal{L} : H^1(t_\mathrm{s}, t_\mathrm{f})^d \times L^2(t_\mathrm{s}, t_\mathrm{f})^d \to \mathbb{R}$ of (5.1) in $H^1(t_\mathrm{s}, t_\mathrm{f})^d$ using the L^2-scalar product (see Section 4.2) is

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}) := J(\boldsymbol{y}(t_\mathrm{f})) - \int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\lambda}^\intercal(t) \left[\dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t))\right] \mathrm{d}t - \boldsymbol{\lambda}^\intercal(t_\mathrm{s}) \left[\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] \tag{5.2}$$

where $\boldsymbol{\lambda} \in L^2(t_\mathrm{s}, t_\mathrm{f})^d$ is the Lagrange multiplier in the dual space of $L^2(t_\mathrm{s}, t_\mathrm{f})^d$, cf. Section 4.3. The optimality condition of (5.1) is based on the Fréchet derivative of $\mathcal{L}$ at $(\boldsymbol{y}, \boldsymbol{\lambda})$ in direction $(\boldsymbol{w}, \boldsymbol{\chi})$ which exists due to Section 4.4 and the Fréchet differentiability of J

$$\begin{aligned}
&\mathcal{L}'(\boldsymbol{y}, \boldsymbol{\lambda})(\boldsymbol{w}, \boldsymbol{\chi}) = \mathcal{L}_{\boldsymbol{y}}(\boldsymbol{y}, \boldsymbol{\lambda})(\boldsymbol{w}) + \mathcal{L}_{\boldsymbol{\lambda}}(\boldsymbol{y}, \boldsymbol{\lambda})(\boldsymbol{\chi}) \\
&= \left\{ J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f}) - \int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\lambda}^\intercal(t) \left[\dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)\right] \mathrm{d}t - \boldsymbol{\lambda}^\intercal(t_\mathrm{s})\boldsymbol{w}(t_\mathrm{s}) \right\} \\
&\quad + \left\{ - \int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\chi}^\intercal(t) \left[\dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t))\right] \mathrm{d}t - \boldsymbol{\chi}^\intercal(t_\mathrm{s}) \left[\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] \right\}.
\end{aligned}$$

The necessary condition for a stationary point $(\boldsymbol{y}, \boldsymbol{\lambda}) \in H^1(t_\mathrm{s}, t_\mathrm{f})^d \times L^2(t_\mathrm{s}, t_\mathrm{f})^d$ of (5.1) is that $\mathcal{L}'(\boldsymbol{y}, \boldsymbol{\lambda})(\boldsymbol{w}, \boldsymbol{\chi}) = 0$ holds for all directions $(\boldsymbol{w}, \boldsymbol{\chi}) \in H^1(t_\mathrm{s}, t_\mathrm{f})^d \times L^2(t_\mathrm{s}, t_\mathrm{f})^d$, see e.g. Luenberger [88] or Ioffe and Tihomirov [75]. Choosing $\boldsymbol{w} = \boldsymbol{0} \in H^1(t_\mathrm{s}, t_\mathrm{f})^d$ and only varying $\boldsymbol{\chi} \in L^2(t_\mathrm{s}, t_\mathrm{f})^d$ the necessary condition reads

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\chi}^\intercal(t) \left[\dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t))\right] \mathrm{d}t + \boldsymbol{\chi}^\intercal(t_\mathrm{s}) \left[\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] = 0, \ \forall \boldsymbol{\chi} \tag{5.3}$$

which possesses the same unique solution $\boldsymbol{y} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ as (1.1). For $\boldsymbol{\chi} = \boldsymbol{0} \in L^2(t_\mathrm{s}, t_\mathrm{f})^d$ and variable $\boldsymbol{w} \in H^1(t_\mathrm{s}, t_\mathrm{f})^d$ one obtains by using integration by parts

$$\left[J'(\boldsymbol{y}(t_\mathrm{f})) - \boldsymbol{\lambda}^\intercal(t_\mathrm{f})\right] \boldsymbol{w}(t_\mathrm{f}) - \int_{t_\mathrm{f}}^{t_\mathrm{s}} \left[\dot{\boldsymbol{\lambda}}(t) + \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t, \boldsymbol{y}(t))\boldsymbol{\lambda}(t)\right]^\intercal \boldsymbol{w}(t) \mathrm{d}t = 0, \ \forall \boldsymbol{w}$$

which possesses the same solution as (1.8). Under the assumptions of Section 1.2, the unique solution $\boldsymbol{\lambda}(t)$ of (1.8) is continuously differentiable on $[t_\mathrm{s}, t_\mathrm{f}]$ and depends continuously on the input data, cf. Theorem 1.7.

Interpretation of the adjoint If the constraints of (5.1), i.e. the nominal IVP (1.1), hold exactly, then the Lagrangian defined by (5.2) takes the value $\mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda}) = J(\boldsymbol{y}(t_\mathrm{f}))$.

But if, for example, the ODE constraint (5.1b) is perturbed by some function $\boldsymbol{r}(t)$, then the adjoint solution $\boldsymbol{\lambda}(t)$ of (1.8) describes the effect of this perturbation on the value of $\mathcal{L}$. Interpreting $\mathcal{L}$ as a function of $\boldsymbol{g}(t) := \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t))$, i.e. $\widetilde{\mathcal{L}}(\boldsymbol{g}, \boldsymbol{\lambda}) := \mathcal{L}(\boldsymbol{y}, \boldsymbol{\lambda})$, the differentiation with respect to $\boldsymbol{g}$ in direction $\boldsymbol{r}$ gives

$$\widetilde{\mathcal{L}}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\lambda})(\boldsymbol{r}) = -\int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\lambda}^\intercal(t)\boldsymbol{r}(t)\mathrm{d}t$$

such that the value of $\widetilde{\mathcal{L}}$ changes in first order to

$$\begin{aligned}\widetilde{\mathcal{L}}(\boldsymbol{g}+\boldsymbol{r}, \boldsymbol{\lambda}) =& \widetilde{\mathcal{L}}(\boldsymbol{g}, \boldsymbol{\lambda}) + \widetilde{\mathcal{L}}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\lambda})(\boldsymbol{r}) + \mathcal{O}\left(\|\boldsymbol{r}\|^2_{C^0[t_\mathrm{s}, t_\mathrm{f}]^d}\right) \\ =& J(\boldsymbol{y}(t_\mathrm{f})) - \int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\lambda}^\intercal(t)\boldsymbol{r}(t)\mathrm{d}t + \mathcal{O}\left(\|\boldsymbol{r}\|^2_{C^0[t_\mathrm{s}, t_\mathrm{f}]^d}\right).\end{aligned}$$

Analogously, the effect of a perturbed initial condition is described by $\boldsymbol{\lambda}(t_\mathrm{s})$. Hence, the adjoint $\boldsymbol{\lambda}(t)$ describes the shadow prices in J for violating the initial condition or the ODE constraint during the solution of the IVP (1.1). Or, in the terminology of Section 1.3, it describes the *(continuous) stability* of the IVP solution in J. The perturbation in J resulting from input perturbations $\|\boldsymbol{r}\|_{C^0[t_\mathrm{s}, t_\mathrm{f}]^d} \leq \varrho$ and $\|\boldsymbol{r}_\mathrm{s}\|_\infty \leq \varrho$ is bounded by

$$|J(\boldsymbol{y}(t_\mathrm{f})) - J(\bar{\boldsymbol{y}}(t_\mathrm{f}))| \leq \left|\int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\lambda}^\intercal(t)\boldsymbol{r}(t)\mathrm{d}t\right| + |\boldsymbol{\lambda}^\intercal(t_\mathrm{s})\boldsymbol{r}_\mathrm{s}| \leq \kappa\varrho$$

where $\bar{\boldsymbol{y}}(t)$ solves the perturbed IVP $\dot{\bar{\boldsymbol{y}}}(t) = \boldsymbol{f}(t, \bar{\boldsymbol{y}}(t)) - \boldsymbol{r}(t)$, $\bar{\boldsymbol{y}}(t_\mathrm{s}) = \boldsymbol{y}_\mathrm{s} + \boldsymbol{r}_\mathrm{s}$, $\boldsymbol{\lambda}(t)$ solves (1.8) and κ is the condition number of Section 1.3.1. This again clarifies the worst-case character of the condition number already described in Section 1.3.1.

Here we derived once again the adjoint IVP (1.8) of Section 1.2.2 but with the help of the standard functional-analytic setting based on Hilbert spaces. This setting is also the basis for the construction of one-step Finite Element (FE) Galerkin methods for ODEs, see e.g. Eriksson et al. [55, 56] and Böttcher and Rannacher [36]. Nevertheless, Section 3.6 showed that adjoint IND schemes of variable multistep BDF methods can not be used to integrate the adjoint IVP.

Formulating the evaluation $J(\boldsymbol{y}_N)$ in the solution of the BDF method (2.2) equivalently as a Nonlinear Program (NLP), analogously as done in the beginning of this chapter for the infinite dimensional case, the resulting NLP is a discretization of (5.1). The first-order optimality conditions are then given by the BDF method (2.2) and its discrete adjoint IND scheme (3.2), for details see Section 6.3. As shown above, the infinite dimensional optimality conditions of (5.1) in $H^1(t_\mathrm{s}, t_\mathrm{f})^d$ are given by the nominal IVP (1.1) and the adjoint IVP (1.8). Hence, the finite dimensional optimality conditions are no discretization of the infinite dimensional conditions due to the adjoints, cf. Section 3.6 and Figure 3.1.

In the remaining of this chapter we will derive infinite dimensional optimality conditions in a more general functional-analytic setting of particular Banach spaces. Their discretization using the FE spaces of Chapter 6 will finally yield in the BDF method together with its adjoint IND scheme.

5.2 Weak adjoint as Lagrange multiplier in $\mathrm{NBV}(t_\mathrm{s}, t_\mathrm{f})^d$

In this section we solve the CVP (5.1) on the space $C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ of all continuously differentiable functions. To this end, we need a variational formulation of the ODE constraint of (5.1). For $\boldsymbol{y} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ the constraint $\dot{\boldsymbol{y}}(\cdot) - \boldsymbol{f}(\cdot, \boldsymbol{y}(\cdot)) = \boldsymbol{0}$ is an element of $C^0[t_\mathrm{s}, t_\mathrm{f}]^d$ and we have to use the duality pairing of $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ and $C^0[t_\mathrm{s}, t_\mathrm{f}]^d$, cf. Section 4.3. Thus, the variational formulation of the IVP (1.1), being the constraints of (5.1), reads: Find $\boldsymbol{y} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ with $\boldsymbol{y}(t_\mathrm{s}) = \boldsymbol{y}_\mathrm{s}$ such that

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t)) \,\mathrm{d}\boldsymbol{\Gamma}(t) = 0 \quad \forall \boldsymbol{\Gamma} \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d. \tag{5.4}$$

This problem possesses at least one solution which is the classical solution given by (1.1). The uniqueness follows from the fact that for continuous functions $g \in C^0[t_\mathrm{s}, t_\mathrm{f}]$ it holds

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} g(t) \,\mathrm{d}\Psi(t) = 0 \quad \forall \Psi \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}] \quad \Rightarrow \quad g = 0.$$

Thus, both formulations (1.1) and (5.4) give the same solution $\boldsymbol{y}(t)$ and (5.4) is well-posed according to the well-posedness of (1.1) described in Section 1.1.

Solving the CVP (5.1) on the function space $C^1[t_\mathrm{s}, t_\mathrm{f}]^d$, the Lagrangian $\mathcal{L} : C^1[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d \to \mathbb{R}$ is given by

$$\mathcal{L}(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l}) := J(\boldsymbol{y}(t_\mathrm{f})) - \int_{t_\mathrm{s}}^{t_\mathrm{f}} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t)) \,\mathrm{d}\boldsymbol{\Lambda}(t) - \boldsymbol{l}^\intercal [\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}] \tag{5.5}$$

where the Lagrange multipliers $\boldsymbol{l}$ and $\boldsymbol{\Lambda}$ lie in the corresponding dual spaces $\mathbb{R}^d$ and $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$, cf. Section 4.3. The Lagrangian is based on the variational formulation (5.4) and includes the initial condition using an additional Lagrange multiplier. We first state the central theorem of this section and defer the proof for the end of the section.

Theorem 5.1 *The optimality conditions of the CVP* (5.1) *on* $C^1[t_\mathrm{s}, t_\mathrm{f}]^d$*, i.e.*

$$J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f}) - \int_{t_\mathrm{s}}^{t_\mathrm{f}} \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t) \,\mathrm{d}\boldsymbol{\Lambda}(t) - \boldsymbol{l}^\intercal \boldsymbol{w}(t_\mathrm{s}) = 0, \tag{5.6a}$$

$$- \int_{t_\mathrm{s}}^{t_\mathrm{f}} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t)) \,\mathrm{d}\boldsymbol{\Gamma}(t) = 0, \tag{5.6b}$$

$$-\boldsymbol{r}^\intercal [\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}] = 0, \tag{5.6c}$$

$$\forall (\boldsymbol{w}, \boldsymbol{\Gamma}, \boldsymbol{r}) \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d,$$

possess a unique solution $(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l})$ *in* $C^1[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$*. Moreover,* $\boldsymbol{y}(t)$ *is the solution of* (1.1)*, and* $\boldsymbol{l}$ *and* $\boldsymbol{\Lambda}(t)$ *are given in terms of the adjoint solution* $\boldsymbol{\lambda}(t)$ *of* (1.8)

$$\boldsymbol{l} = \boldsymbol{\lambda}(t_\mathrm{s}), \quad \boldsymbol{\Lambda}(t) = \int_{t_\mathrm{s}}^{t} \boldsymbol{\lambda}(\tau) \mathrm{d}\tau, \tag{5.7}$$

with componentwise integration.

The necessary optimality condition for a stationary point $(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l})$ of the Lagrangian (5.5) is given by

$$\begin{pmatrix} \mathcal{L}_{\boldsymbol{y}}(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{w}) \\ \mathcal{L}_{\boldsymbol{\Lambda}}(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{\Gamma}) \\ \mathcal{L}_{\boldsymbol{l}}(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{r}) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad \forall \boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d, \ \boldsymbol{\Gamma} \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d, \ \boldsymbol{r} \in \mathbb{R}^d$$

which is exactly (5.6). As equations (5.6b)-(5.6c) are already given by (5.4) and discussed over there, we now focus on equation (5.6a) of the optimality conditions. Provided that $\boldsymbol{y}(t)$ is known, the adjoint problem in variational formulation reads: Find $(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ such that (5.6a) holds for all $\boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$.

Lemma 5.2 *For the solution $\boldsymbol{y}(t)$ of* (5.6b)-(5.6c)*, a corresponding adjoint solution $(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ of* (5.6a) *is provided by* (5.7)*.*

Proof *Recall that the adjoint IVP* (1.8) *has a unique solution $\boldsymbol{\lambda} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$, cf. Section 1.2.2. Multiplying the transposed of* (1.8a) *from the right by any $\boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$, integrating over $[t_\mathrm{s}, t_\mathrm{f}]$ and adding the transposed of* (1.8b) *multiplied by $\boldsymbol{w}(t_\mathrm{f})$ yields*

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} \left[\dot{\boldsymbol{\lambda}}(t) + \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t, \boldsymbol{y}(t))\boldsymbol{\lambda}(t)\right]^\intercal \boldsymbol{w}(t)\mathrm{d}t - \left[\boldsymbol{\lambda}(t_\mathrm{f}) - J'(\boldsymbol{y}(t_\mathrm{f}))^\intercal\right]^\intercal \boldsymbol{w}(t_\mathrm{f}) = 0. \tag{5.8}$$

Integration by parts gives for all $\boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{\lambda}^\intercal(t) \left[\dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)\right] \mathrm{d}t - \boldsymbol{\lambda}^\intercal(t_\mathrm{s})\boldsymbol{w}(t_\mathrm{s}) + J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f}) = 0.$$

Consequently, (5.7) *provides a solution $(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ of* (5.6a)*, since the indefinite integral $\Lambda_i(t) = \int_{t_\mathrm{s}}^t \lambda_i(\tau)\mathrm{d}\tau$ is a normalized function of bounded variation and it holds $\int_{t_\mathrm{s}}^{t_\mathrm{f}} g(t)\mathrm{d}\Lambda_i(t) = \int_{t_\mathrm{s}}^{t_\mathrm{f}} \Lambda_i'(t)g(t)\mathrm{d}t = \int_{t_\mathrm{s}}^{t_\mathrm{f}} \lambda_i(t)g(t)\mathrm{d}t$, cf. Section 4.1.1 and 4.1.2.* □

The next lemma proves the uniqueness of the adjoint solution.

Lemma 5.3 *For the solution $\boldsymbol{y}(t)$ of* (5.6b)-(5.6c)*, the corresponding adjoint solution $(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ of* (5.6a) *is unique.*

Proof *Equation* (5.6a) *is equivalent to*

$$\underbrace{\int_{t_\mathrm{s}}^{t_\mathrm{f}} \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) + \boldsymbol{l}^\intercal \boldsymbol{w}(t_\mathrm{s})}_{=:\boldsymbol{A}(\boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{w})} = \underbrace{J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f})}_{=:B(\boldsymbol{w})} \quad \forall \boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$$

where B and $\boldsymbol{A}(\boldsymbol{\Lambda}, \boldsymbol{l})$ are linear functionals on $C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ and $\boldsymbol{A} : \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d \to \left(C^1[t_\mathrm{s}, t_\mathrm{f}]^d\right)'$ is linear in $(\boldsymbol{\Lambda}, \boldsymbol{l})$. We have to show that $\mathcal{N}(\boldsymbol{A}) = \{(\boldsymbol{0}, \boldsymbol{0})\}$, where the nullspace of $\boldsymbol{A}$ is given by

$$\mathcal{N}(\boldsymbol{A}) = \left\{(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d : \ \boldsymbol{A}(\boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{w}) = 0 \quad \forall \boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d\right\}.$$

Due to Section 1.2, for every initial value $\boldsymbol{w}_1(t_\mathrm{s}) \in \mathbb{R}^d$ *there exists a function* $\boldsymbol{w}_1 \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ *that satisfies the ODE of* (1.4)*. Inserting* $\boldsymbol{w}_1$ *in* $\boldsymbol{A}(\boldsymbol{\Lambda}, \boldsymbol{l})$ *then gives*

$$\boldsymbol{A}(\boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{w}_1) = \int_{t_\mathrm{s}}^{t_\mathrm{f}} \mathbf{0}\,\mathrm{d}\boldsymbol{\Lambda}(t) + \boldsymbol{l}^\intercal \boldsymbol{w}_1(t_\mathrm{s}) = 0 + \boldsymbol{l}^\intercal \boldsymbol{w}_1(t_\mathrm{s}).$$

Thus, $\boldsymbol{l}$ *has to vanish in order to ensure* $\boldsymbol{A}(\boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{w}) = 0\ \forall \boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$*. Now, we search for functions* $\boldsymbol{\Lambda} \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ *with*

$$\boldsymbol{A}(\boldsymbol{\Lambda}, \mathbf{0})(\boldsymbol{w}) = \int_{t_\mathrm{s}}^{t_\mathrm{f}} \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) = 0 \quad \forall \boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d.$$

With $\boldsymbol{g}(t) := \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)$*, it is the same to vary either* $\boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ *or* $\boldsymbol{g} \in C^0[t_\mathrm{s}, t_\mathrm{f}]^d$*, since the inhomogeneous ODE possesses a unique solution* $\boldsymbol{w}(t)$ *for every* $\boldsymbol{g}(t)$*. According to the uniqueness of* Ψ *in* (4.1) *it holds*

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{g}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) = 0 \quad \forall \boldsymbol{g} \in C^0[t_\mathrm{s}, t_\mathrm{f}]^d \quad \Rightarrow \quad \boldsymbol{\Lambda} = \mathbf{0}.$$

Thus, $\mathcal{N}(\boldsymbol{A}) = \{(\mathbf{0}, \mathbf{0})\}$ *which proves the uniqueness of the solution of* (5.6a)*.* □

With this knowledge at hand we can now come to the proof of Theorem 5.1.

Proof (of Theorem 5.1) *As seen in the beginning of the section, the equations* (5.6b)-(5.6c) *have the same unique solution* $\boldsymbol{y}(t)$ *as* (1.1) *which implies their well-posedness. According to Lemma 5.2, a solution of* (5.6a) *is provided by* (5.7)*. Furthermore, it is the only solution of* (5.6a) *according to Lemma 5.3. Since* $\boldsymbol{\lambda}(t)$ *depends continuously on* $J'(\boldsymbol{y}(t_\mathrm{f}))^\intercal$ *(cf. Section 1.2.2) this still holds for* $\boldsymbol{\Lambda}(t)$ *and* $\boldsymbol{l}$*. Thus,* (5.6a) *together with* (5.6b)-(5.6c) *is well-posed.* □

With the concept of weak solutions from Partial Differential Equations (PDEs), see e.g. Johnson [76], the triple $(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l})$ is a weak solution of (1.1) and (1.8) since it solves the variational formulation (5.6) of (1.1) and (1.8). Thus, we will call $\boldsymbol{\Lambda}$ a *weak adjoint solution* of (1.8) or shortly *weak adjoint*. Note that for the nominal solution, the weak solution $\boldsymbol{y}$ defined by (5.6c)-(5.6b) is directly given by the *classical solution* of (1.1). Whereas for the adjoint, the weak solution $\boldsymbol{\Lambda}$ is sufficiently regular such that a classical solution of (1.8) is provided by $\boldsymbol{\Lambda}' = \boldsymbol{\lambda}$.

Interpretation of the weak adjoint For the weak adjoint the same interpretation like for the classical adjoint holds, cf. page 48. Perturbations $\boldsymbol{r}(t)$ of the ODE constraint affect the value of $J(\boldsymbol{y}(t_\mathrm{f}))$ by

$$\tilde{\mathcal{L}}_{\boldsymbol{g}}(\boldsymbol{g}, \boldsymbol{\Lambda}, \boldsymbol{l})(\boldsymbol{r}) = -\int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{r}(t)\mathrm{d}\boldsymbol{\Lambda}(t)$$

and the effect of a perturbed initial condition is described by $\boldsymbol{l} = \boldsymbol{\lambda}_\mathrm{s}$. The weak adjoint $\boldsymbol{\Lambda}(t)$ itself accumulates the classical adjoint from t_s to t due to (5.7). Moreover,

the adjoint pair $(\boldsymbol{\Lambda}, \boldsymbol{l})$ can be used to get a lower bound on the condition number κ defined in Section 1.3.1 since

$$\begin{aligned}\|\boldsymbol{l}\|_1 + \|\boldsymbol{\Lambda}(t_\mathrm{f})\|_1 &= \sum_{i=1}^{d} |\lambda_i(t_\mathrm{s})| + \sum_{i=1}^{d} \left| \int_{t_\mathrm{s}}^{t_\mathrm{f}} \lambda_i(\tau) \mathrm{d}\tau \right| \\ &\leq \sum_{i=1}^{d} |\lambda_i(t_\mathrm{s})| + \sum_{i=1}^{d} \int_{t_\mathrm{s}}^{t_\mathrm{f}} |\lambda_i(\tau)| \, \mathrm{d}\tau = \kappa.\end{aligned}$$

The bound is sharp if for every $i = 1, \ldots, d$ either $\lambda_i(t) \geq 0$ or $\lambda_i(t) \leq 0$ holds for all $t \in [t_\mathrm{s}, t_\mathrm{f}]$.

5.3 Weak adjoint as Lagrange multiplier in $(Y[t_\mathrm{s}, t_\mathrm{f}]^d)'$

Most integrators, including the BDF method of Chapter 2, give approximations to the solution of (1.1) that are not continuously differentiable on the whole interval $[t_\mathrm{s}, t_\mathrm{f}]$ but rather continuous and piecewise continuously differentiable. To capture this case, an appropriate extension of the trial space $C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ is required. To this end, we employ a time grid $t_\mathrm{s} = t_0 < t_1 < \cdots < t_N = t_\mathrm{f}$ and a partition of $[t_\mathrm{s}, t_\mathrm{f}]$ using subintervals $I_n := (t_n, t_{n+1}]$ of length $h_n = t_{n+1} - t_n$ such that $[t_\mathrm{s}, t_\mathrm{f}] = \{t_\mathrm{s}\} \cup I_0 \cup \cdots \cup I_{N-1}$. Choosing the trial space as

$$Y[t_\mathrm{s}, t_\mathrm{f}]^d := \left\{ \boldsymbol{y} \in C^0[t_\mathrm{s}, t_\mathrm{f}]^d : \boldsymbol{y}\big|_{I_n} \in C_b^1(I_n)^d \right\}, \tag{5.9}$$

where $C_b^1(I_n)$ is the space of all continuously differentiable and bounded functions with bounded derivative, cf. Section 4.2.

Solving the CVP (5.1) on $Y[t_\mathrm{s}, t_\mathrm{f}]^d$, the ODE constraint $\dot{\boldsymbol{y}}(\cdot) - \boldsymbol{f}(\cdot, \boldsymbol{y}(\cdot)) = \boldsymbol{0}$ is piecewise continuous and bounded on $[t_\mathrm{s}, t_\mathrm{f}]$, i.e. continuous on I_n. Due to the latter property, it is continuous from the left (see Definition A.3) on $[t_\mathrm{s}, t_\mathrm{f}]$. In order to find the appropriate duality pairing, we have to extend the linear functional $\mathfrak{L}$ defined by (4.1) from $C^0[t_\mathrm{s}, t_\mathrm{f}]$ to $\bigcup_{n=0}^{N-1} C_\mathrm{b}^0(I_n)$ by generalizing Definition 4.6 of the Riemann-Stieltjes integral to allow integrands that are continuous from the left.

5.3.1 Extension of the Riemann-Stieltjes integral

As mentioned in Section 4.2 and 4.3, in this thesis we follow the convention that the functions of bounded variation are continuous from the right. To allow Riemann-Stieltjes integration of integrands that have discontinuouties at the same points as the generating function and are continuous from the left we have to extend Definition 4.6 of the Riemann-Stieltjes integral appropriately.

Definition 5.4 *Let g, Φ be two functions on $[a, b]$ with Φ being of bounded variation and g being continuous from the left with a single discontinuity at $c \in (a, b)$. Assume that $\int_a^c f(t)\mathrm{d}\Phi(t)$ and $\int_{(c,b]} f(t)\mathrm{d}\Phi(t)$ exist where partitions $\mathcal{T}^k \in \mathcal{T}((c, b])$ take the*

form $c < \tau_0^k < \tau_1^k < \cdots < \tau_m^k = b$. *Then, the* extended Riemann-Stieltjes integral *is given by*

$$\oint_a^b g(t)\mathrm{d}\Phi(t) = \int_a^c g(t)\mathrm{d}\Phi(t) + \int_{(c,b]} g(t)\mathrm{d}\Phi(t).$$

In words, the extended Riemann-Stieltjes integral splits the standard Riemann-Stieltjes integral into a sum of those parts where the integrand is continuous. Hence, if $g \in C^0[a,b]$, then both Riemann-Stieltjes integrals coincide, i.e.

$$\oint_a^b g(t)\mathrm{d}\Phi(t) = \int_a^b g(t)\mathrm{d}\Phi(t).$$

If $g \in C^0[a,b]$, then it also holds that

$$\int_{(a,b]} g(t)\mathrm{d}\Phi(t) = \int_a^b g(t)\mathrm{d}\Phi(t)$$

since Φ is continuous from the right. Subsequently, we will always use the notion of the extended Riemann-Stieltjes integral in terms of the standard Riemann-Stieltjes integral on half open intervals.

5.3.2 Solution of the Constrained Variational Problem in $Y[t_\mathrm{s}, t_\mathrm{f}]^d$

The existence of an extension $\widehat{\mathfrak{L}}$ of the linear functional $\mathfrak{L}$ defined by (4.1) from $C^0[t_\mathrm{s}, t_\mathrm{f}]$ to $Y[t_\mathrm{s}, t_\mathrm{f}]$ is guaranteed by the Hahn-Banach Extension Theorem (Theorem 4.21). A suitable extension is provided by

$$\widehat{\mathfrak{L}}(g) = \sum_{n=0}^{N-1} \int_{I_n} g(t)\mathrm{d}\Psi(t)$$

using the extended Riemann-Stieltjes integral of Section 5.3.1. This extension $\widehat{\mathfrak{L}}$ restricted to the continuous functions $g \in C^0[t_\mathrm{s}, t_\mathrm{f}]$ coincides with $\mathfrak{L}$ defined by (4.1).

Now, solving the CVP (5.1) on $Y[t_\mathrm{s}, t_\mathrm{f}]^d$, the extended Lagrangian $\widehat{\mathcal{L}} : Y[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d \to \mathbb{R}$ of (5.1) on $Y[t_\mathrm{s}, t_\mathrm{f}]^d$ is given by

$$\widehat{\mathcal{L}}(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l}) := J(\boldsymbol{y}(t_\mathrm{f})) - \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t))\,\mathrm{d}\boldsymbol{\Lambda}(t) - \boldsymbol{l}^\mathsf{T}\left[\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right]. \quad (5.10)$$

The Lagrangian $\widehat{\mathcal{L}}$ is based on the extension $\widehat{\mathfrak{L}}$, and thus restricted to $\boldsymbol{y} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ it coincides with the Lagrangian $\mathcal{L}$ defined by (5.5).

With these definitions at hand, we first state the main result of the section.

Theorem 5.5 *The optimality conditions of the CVP* (5.1) *on* $Y[t_\mathrm{s},t_\mathrm{f}]^d$*, i.e.*

$$J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f}) - \sum_{n=0}^{N-1}\int_{I_n} \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t,\boldsymbol{y}(t))\boldsymbol{w}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) - \boldsymbol{l}^\intercal\boldsymbol{w}(t_\mathrm{s}) = 0, \tag{5.11a}$$

$$-\sum_{n=0}^{N-1}\int_{I_n} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t,\boldsymbol{y}(t))\,\mathrm{d}\boldsymbol{\Gamma}(t) = 0, \tag{5.11b}$$

$$-\boldsymbol{r}^\intercal\left[\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] = 0, \tag{5.11c}$$

$$\forall(\boldsymbol{w},\boldsymbol{\Gamma},\boldsymbol{r}) \in Y[t_\mathrm{s},t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d \times \mathbb{R}^d,$$

possess a unique solution $(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})$ *in* $Y[t_\mathrm{s},t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d \times \mathbb{R}^d$ *that coincides with the solution of* (5.6)*.*

We start with considering the nominal equations (5.11c)-(5.11b).

Lemma 5.6 *The solution* $\boldsymbol{y}(t)$ *of* (5.6c)-(5.6b) *solves the extended variational formulation* (5.11c)-(5.11b)*.*

Proof *Let* $\boldsymbol{y}(t)$ *be the solution of* (5.6c)-(5.6b)*. From* $C^1[t_\mathrm{s},t_\mathrm{f}]^d \subset Y[t_\mathrm{s},t_\mathrm{f}]^d$ *follows that* $\boldsymbol{y} \in Y[t_\mathrm{s},t_\mathrm{f}]^d$*. Since the integral* $\int_{t_\mathrm{s}}^{t_\mathrm{f}} g_i(t)\mathrm{d}\Gamma_i(t)$ *for the continuous integrand* $g_i(t) := \dot{y}_i(t) - f_i(t,y(t))$ *exists (Theorem 4.7), Lemma 4.8 states that also the integrals* $\int_{t_n}^{t_{n+1}} g_i(t)\mathrm{d}\Gamma_i(t)$ *over the subintervals exist and it holds*

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} g_i(t)\mathrm{d}\Gamma_i(t) = \sum_{n=0}^{N-1}\int_{t_n}^{t_{n+1}} g_i(t)\mathrm{d}\Gamma_i(t) = \sum_{n=0}^{N-1}\int_{I_n} g_i(t)\mathrm{d}\Gamma_i(t)$$

where the second equality is due to Section 5.3.1 on the extended Riemann-Stieltjes integral, $i = 1,\ldots,d$*. Thus, equation* (5.6b) *becomes* $\forall\boldsymbol{\Gamma} \in \mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d$

$$0 = \int_{t_\mathrm{s}}^{t_\mathrm{f}} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t,\boldsymbol{y}(t))\,\mathrm{d}\boldsymbol{\Gamma}(t) = \sum_{n=0}^{N-1}\int_{I_n} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t,\boldsymbol{y}(t))\,\mathrm{d}\boldsymbol{\Gamma}(t)$$

which coincides with (5.11b)*. Hence,* $\boldsymbol{y}(t)$ *also solves* (5.11b) *and trivially* (5.11c)*.*□

Lemma 5.7 *The extended variational formulation* (5.11b)-(5.11c) *possesses a unique solution* $\boldsymbol{y}(t)$*.*

Proof *Let* $\boldsymbol{y}(t)$ *be a solution of* (5.11b)-(5.11c)*. The space* $\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d$ *contains, in particular, the continuous functions of bounded variation that vanish everywhere except on* (t_n,t_{n+1})*. Thus, a necessary condition for* $\boldsymbol{y}(t)$ *being a solution of* (5.11b)-(5.11c) *is that each addend has to vanish, i.e.* $\int_{I_n} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t,\boldsymbol{y}(t))\mathrm{d}\boldsymbol{\Gamma}(t) = 0\ \forall\boldsymbol{\Gamma} \in \mathrm{NBV}(I_n)^d$ *with* $\boldsymbol{\Gamma}(t_{n+1}) = \boldsymbol{0}$*. The Fundamental Theorem of Variational Calculus yields* $\dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t,\boldsymbol{y}(t)) = \boldsymbol{0}$ *on* (t_n,t_{n+1}) *for all* $n = 0,\ldots,N-1$*. On the other hand,* $\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d$ *contains also the constant functions having a single jump in* t_n*. They give according to Section 4.1.2 and 5.3.1 the necessary conditions* $\dot{\boldsymbol{y}}(t_n) - \boldsymbol{f}(t_n,\boldsymbol{y}(t_n)) = \boldsymbol{0}$ *for* $n = 1,\ldots,N$*. Since* $\boldsymbol{f}(t,\boldsymbol{y})$ *is continuous in both variables and* $\boldsymbol{y} \in C^0[t_\mathrm{s},t_\mathrm{f}]^d$*,* $\boldsymbol{y}(t)$ *is necessarily continuously differentiable on* $[t_\mathrm{s},t_\mathrm{f}]$*. Thus, every solution of* (5.11b)-(5.11c) *satisfies* (5.6b)-(5.6c) *which possesses a unique solution.*□

As conclusion of this lemma, the extended variational formulation (5.11b)-(5.11c) is well-posed according to the well-posedness of (5.6b)-(5.6c).

Now, we focus on the adjoint problem in extended variational formulation which is for a given $\boldsymbol{y}(t)$: Find $(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ such that (5.11a) holds for all $\boldsymbol{w} \in Y[t_\mathrm{s}, t_\mathrm{f}]^d$.

Lemma 5.8 *For the solution $\boldsymbol{y}(t)$ of (5.11b)-(5.11c), the corresponding adjoint solution $(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ of (5.11a) is provided by (5.7).*

Proof *We proceed in the same way as in the proof of Lemma 5.2, but choose $\boldsymbol{w} \in Y[t_\mathrm{s}, t_\mathrm{f}]^d$ for the multiplication and split the integral in (5.8) using the subintervals I_n (same arguments as in the proof of Lemma 5.6). Integration by parts of all integrals yields the equivalent equation*

$$-\boldsymbol{\lambda}^\mathsf{T}(t_\mathrm{s})\boldsymbol{w}(t_\mathrm{s}) - \sum_{n=0}^{N-1} \int_{I_n} \boldsymbol{\lambda}^\mathsf{T}(t) \left[\dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)\right] \mathrm{d}t + J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f}) = 0.$$

Thus, the choice (5.7) provides a solution of (5.11a). □

Lemma 5.9 *For the solution $\boldsymbol{y}(t)$ of (5.11b)-(5.11c), the corresponding adjoint solution $(\boldsymbol{\Lambda}, \boldsymbol{l}) \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ of (5.11a) is unique.*

Proof *We follow mainly the proof of Lemma 5.3. Equation (5.11a) is equivalent to*

$$\underbrace{\sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)\, \mathrm{d}\boldsymbol{\Lambda}(t) + \boldsymbol{l}^\mathsf{T}\boldsymbol{w}(t_\mathrm{s})}_{=:\hat{\boldsymbol{A}}(\boldsymbol{\Lambda},\boldsymbol{l})(\boldsymbol{w})} = \underbrace{J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}(t_\mathrm{f})}_{=B(\boldsymbol{w})} \quad \forall \boldsymbol{w} \in Y[t_\mathrm{s}, t_\mathrm{f}]^d$$

where $\hat{\boldsymbol{A}}(\boldsymbol{\Lambda}, \boldsymbol{l})$ is also a linear functional on $Y[t_\mathrm{s}, t_\mathrm{f}]^d$ and $\hat{\boldsymbol{A}} : \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d \to \left(Y[t_\mathrm{s}, t_\mathrm{f}]^d\right)'$ is linear in $(\boldsymbol{\Lambda}, \boldsymbol{l})$. We show again that $\mathcal{N}(\hat{\boldsymbol{A}}) = \{(\boldsymbol{0}, \boldsymbol{0})\}$. Since $C^1[t_\mathrm{s}, t_\mathrm{f}]^d \subset Y[t_\mathrm{s}, t_\mathrm{f}]^d$, $\boldsymbol{l}$ has to vanish due to the same arguments as used in the proof of Lemma 5.3. Thus, the following equation

$$\hat{\boldsymbol{A}}(\boldsymbol{\Lambda}, \boldsymbol{0})(\boldsymbol{w}) = \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)\, \mathrm{d}\boldsymbol{\Lambda}(t) = 0 \quad \forall \boldsymbol{w} \in Y[t_\mathrm{s}, t_\mathrm{f}]^d$$

has to be satisfied also for $\boldsymbol{w} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d \subset Y[t_\mathrm{s}, t_\mathrm{f}]^d$, i.e. with $\boldsymbol{g}(t) := \dot{\boldsymbol{w}}(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))\boldsymbol{w}(t)$ it becomes

$$\sum_{n=0}^{N-1} \int_{I_n} \boldsymbol{g}(t)\, \mathrm{d}\boldsymbol{\Lambda}(t) = 0 \quad \forall \boldsymbol{g} \in C^0[t_\mathrm{s}, t_\mathrm{f}]^d.$$

Furthermore, as $\boldsymbol{g}(t)$ is continuous the integral $\int_{t_\mathrm{s}}^{t_\mathrm{f}} \boldsymbol{g}(t)\mathrm{d}\boldsymbol{\Lambda}(t)$ exists and coincides with the sum of the integrals over the subintervals (same arguments as in the proof of Lemma 5.6) and the proof can be finished in the same way as that of Lemma 5.3. □

With all this at hand we are able to prove Theorem 5.5.

Proof (of Theorem 5.5) *Lemma 5.6 and 5.7 prove the existence of a unique solution of* (5.11b)-(5.11c) *coinciding with the solution of* (5.6b)-(5.6c)*. For this solution,* (5.11a) *has a unique solution given by* (5.7) *due to Lemma 5.8 and 5.9.* □

The novel functional-analytic framework derived in this chapter holds generally for IVPs in ODEs. Hence, it is not limited to the analysis of BDF methods and their adjoint IND schemes but rather allows to analyze integration methods that provide at least a continuous and piecewise continuously differentiable approximation to the solution of the nominal IVP (1.1). In the following chapter we propose a FE Petrov-Galerkin discretization of the infinite dimensional optimality conditions (5.11) that is particularly suitable to relate the discrete adjoint IND values of BDF methods and the solution of the adjoint IVP by the help of the weak adjoint solution.

6 Petrov-Galerkin Finite Element discretization

In order to solve the infinite dimensional optimality conditions (5.11) derived in the last chapter numerically, the infinite dimensional function spaces $Y[t_\mathrm{s}, t_\mathrm{f}]^d$ and $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ have to be approximated by finite dimensional subspaces, the Finite Element (FE) spaces. This so-called *Petrov-Galerkin approximation* transfers the infinite dimensional conditions into a finite dimensional system of equations. This chapter follows again our own work Beigel et al. [21] with small modifications. In the first part we propose particular basis functions to span the finite dimensional subspaces, and the second part is devoted to the resulting system of equations and its equivalence to the Backward Differentiation Formula (BDF) method and its discrete adjoint Internal Numerical Differentiation (IND) scheme. Finally, we focus on the so obtained commutativity of discretization and differentiation in the case of multistep BDF methods.

6.1 Finite Element spaces

This section deals with the discretization of the infinite dimensional function spaces $Y[t_\mathrm{s}, t_\mathrm{f}]^d$ and $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ by choosing appropriate sets of basis functions. This is the general procedure of FE methods which are mostly used to solve Partial Differential Equations (PDEs) and only sometime to solve Ordinary Differential Equations (ODEs). For more details on FE methods for PDEs, we refer the reader to Ern and Guermond [57], Brenner and Scott [39] and Braess [37].

6.1.1 Trial space

To discretize the trial space $Y[t_\mathrm{s}, t_\mathrm{f}]^d$ we use piecewise polynomials of order k_n on the subinterval I_n

$$Y_{\mathcal{P}}[t_\mathrm{s}, t_\mathrm{f}]^d := \left\{ \boldsymbol{y} \in C^0[t_\mathrm{s}, t_\mathrm{f}]^d : \boldsymbol{y}\big|_{I_n} \in \mathcal{P}^{(k_n)}(I_n)^d \right\} \tag{6.1}$$

where $\mathcal{P}^{(k_n)}(I_n)$ denotes the space of all polynomials of degree k_n on I_n. We choose local basis functions φ_n that are composed of the fundamental Lagrangian polynomials (2.4) restricted to the particular subinterval. Figure 6.1 shows the basis function $\varphi_n \in Y_{\mathcal{P}}[t_\mathrm{s}, t_\mathrm{f}]^d$ for $n \geq 2$ with $k_0 = 1$, $k_m = 2$ for $m > 0$ and $h_m = h$ for all $m = 0, \ldots, N-1$. The support of a single basis function depends on the orders and

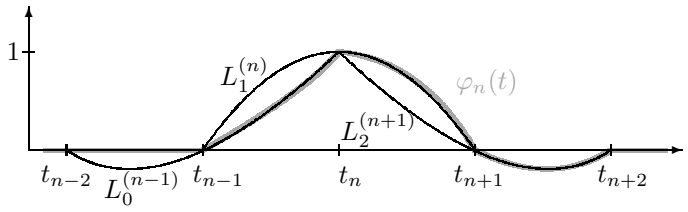


Figure 6.1: Basis function φ_n of $Y_{\mathcal{P}}[t_\mathrm{s}, t_\mathrm{f}]^d$ for $n \geq 2$ with $k_0 = 1$, $k_m = 2$ for $m > 0$ and constant stepsizes $h_m = h$ for all m.

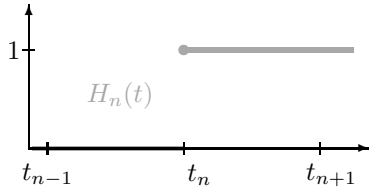


Figure 6.2: Basis function H_n of $Z_H[t_\mathrm{s}, t_\mathrm{f}]^d$.

contains at most seven adjacent subintervals as BDF methods are zero-stable up to order 6, see Theorem 2.19.

The function $\boldsymbol{y} \in Y[t_\mathrm{s}, t_\mathrm{f}]^d$ is then approximated by

$$\boldsymbol{y}(t) \approx \boldsymbol{y}^h(t) := \boldsymbol{y}_\mathrm{s}\varphi_0(t) + \sum_{n=1}^{N} \boldsymbol{y}_n \varphi_n(t)$$

with $(N+1)\cdot d$ degrees of freedom $\{\boldsymbol{y}_n \in \mathbb{R}^d\}_{n=0}^N$. To achieve locally the order $k_n > 1$, former values $\boldsymbol{y}_{n+1-k_n}, \ldots, \boldsymbol{y}_n$ are reused to set up the interpolation polynomial of order k_n which is afterwards restricted to I_n, cl. Section 2.1.

6.1.2 Test space

We approximate the test space $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ using Heaviside functions as basis functions. We choose them to be continuous from the right with discontinuity in t_n

$$H_n(t) := \begin{cases} 0 & t < t_n \\ 1 & t \geq t_n \end{cases}$$

as depicted in Figure 6.2. Thus, a function $\boldsymbol{\Lambda} \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ is approximated by the linear combination of these basis functions in the form

$$\boldsymbol{\Lambda}(t) \approx \boldsymbol{\Lambda}^h(t) := \sum_{n=1}^{N} h_{n-1} \boldsymbol{\lambda}_n H_n(t) \tag{6.2}$$

where the h_{n-1} appear for reasons which will become clear later. Note that $\boldsymbol{\Lambda}^h$ is a step function with initial value $\boldsymbol{\Lambda}^h(t_\mathrm{s}) = \mathbf{0}$ and jumps of magnitude $h_{n-1}\boldsymbol{\lambda}_n$ at t_n

for $n = 1, \dots, N$. Thus, it is $\mathbf{\Lambda}^h(t_n) = \mathbf{\Lambda}^h(t_{n-1}) + h_{n-1}\boldsymbol{\lambda}_n$ at the grid points and $\mathbf{\Lambda}^h(t) = \mathbf{\Lambda}^h(t_n)$ for inner points $t \in (t_n, t_{n+1})$. We denote this space by $Z_{\mathrm{H}}[t_\mathrm{s}, t_\mathrm{f}]^d$.

Regarding the relation (5.7) between the adjoint solutions $\boldsymbol{\lambda}$ and $\mathbf{\Lambda}$, the classical derivative of $\mathbf{\Lambda}^h$ fails to exist. But $\mathbf{\Lambda}^h$ is still differentiable in a weak form such that its weak derivative is given by the Dirac measures at $\{t_1, \dots, t_N\}$ with heights $\{h_0\boldsymbol{\lambda}_1, \dots, h_{N-1}\boldsymbol{\lambda}_N\}$, see e.g. Alt [7].

6.2 Finite dimensional optimality conditions

In this section, we approximate the infinite dimensional optimality conditions (5.11) by finite dimensional equations that result from approximating the function spaces $Y[t_\mathrm{s}, t_\mathrm{f}]^d$ and $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ by the FE spaces $Y_{\mathcal{P}}[t_\mathrm{s}, t_\mathrm{f}]^d$ and $Z_H[t_\mathrm{s}, t_\mathrm{f}]^d$ of Section 6.1. The resulting system of equations will be discussed in the following.

Theorem 6.1 *The discretized optimality conditions, i.e.*

$$
\begin{aligned}
&J'(\boldsymbol{y}^h(t_\mathrm{f}))\boldsymbol{w}^h(t_\mathrm{f}) \\
&\quad - \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{w}}^h(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}^h(t))\boldsymbol{w}^h(t)\,\mathrm{d}\mathbf{\Lambda}^h(t) - [\boldsymbol{l}^h]^\intercal \boldsymbol{w}^h(t_\mathrm{s}) = 0, && \text{(6.3a)} \\
&\qquad\qquad - \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}\mathbf{\Gamma}^h(t) = 0, && \text{(6.3b)} \\
&\qquad\qquad\qquad -[\boldsymbol{r}^h]^\intercal \left[\boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] = 0, && \text{(6.3c)} \\
&\qquad \forall (\boldsymbol{w}^h, \mathbf{\Gamma}^h, \boldsymbol{r}^h) \in Y_{\mathcal{P}}[t_\mathrm{s}, t_\mathrm{f}]^d \times Z_H[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d,
\end{aligned}
$$

with $\boldsymbol{l}^h = \boldsymbol{\lambda}_0$ are equivalent to the BDF scheme (2.2) *with prescribed stepsizes $\{h_n\}_{n=0}^{N-1}$ and orders $\{k_n\}_{n=0}^{N-1}$ together with its discrete adjoint IND scheme* (3.2)*.*

The above theorem is the main result of this section. The proof follows directly from the two lemmas given below.

Lemma 6.2 *The equations* (6.3b)-(6.3c) *are equivalent to the BDF scheme* (2.2) *with prescribed stepsizes $\{h_n\}_{n=0}^{N-1}$ and orders $\{k_n\}_{n=0}^{N-1}$.*

Proof *We first consider one addend of* (6.3b)

$$
\begin{aligned}
&\int_{I_n} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}\mathbf{\Gamma}^h(t) \\
&\quad = \left[\mathbf{\Gamma}^h(t_{n+1}) - \mathbf{\Gamma}^h(t_n)\right]^\intercal \left\{\dot{\boldsymbol{y}}^h(t_{n+1}) - \boldsymbol{f}(t_{n+1}, \boldsymbol{y}^h(t_{n+1}))\right\} \\
&\quad = \boldsymbol{\gamma}_{n+1}^\intercal \left\{ \sum_{i=0}^{k_n} \underbrace{h_n \dot{\varphi}_{n+1-i}(t_{n+1})}_{=\alpha_i^{(n)}} \boldsymbol{y}_{n+1-i} - h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) \right\}
\end{aligned}
$$

where the first equality holds due to the extended Riemann-Stieltjes integral of Section 5.3.1 in vector-valued version with coefficients $h_n\boldsymbol{\gamma}_{n+1}$ of $\boldsymbol{\Gamma}^h$ in (6.2)*. The second equality uses the properties of the basis functions φ_n. Here the appearance of the h_n in the coefficients of $\boldsymbol{\Lambda}^h$ given by* (6.2) *becomes clear. Thus,* (6.3b) *can be written as a system of equations that is nonlinear in $\{\boldsymbol{y}_n\}_{n=1}^N$ and linear in $\boldsymbol{\gamma}^\intercal := \begin{bmatrix} \boldsymbol{\gamma}_1^\intercal & \boldsymbol{\gamma}_2^\intercal & \cdots & \boldsymbol{\gamma}_N^\intercal \end{bmatrix} \in \left(\mathbb{R}^d\right)^N$*

$$\boldsymbol{\gamma}^\intercal \left[(\boldsymbol{A} \otimes \boldsymbol{I}) \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} + \begin{pmatrix} \alpha_1^{(0)} \boldsymbol{y}_\mathrm{s} \\ 0 \\ \vdots \\ 0 \end{pmatrix} - \begin{pmatrix} h_0 \boldsymbol{f}(t_1, \boldsymbol{y}_1) \\ h_1 \boldsymbol{f}(t_2, \boldsymbol{y}_2) \\ \vdots \\ h_{N-1} \boldsymbol{f}(t_N, \boldsymbol{y}_N) \end{pmatrix} \right] = 0, \ \forall \boldsymbol{\gamma} \tag{6.4}$$

where $\boldsymbol{A} \otimes \boldsymbol{I}$ denotes the Kronecker tensor product, i.e. the $(N \cdot d) \times (N \cdot d)$ matrix with $d \times d$ blocks $a_{ij}\boldsymbol{I}$, and the quadratic matrix $\boldsymbol{A}$ is lower triangular with band structure

$$\boldsymbol{A} = \begin{pmatrix} \alpha_0^{(0)} & 0 & 0 & 0 & \cdots \\ \alpha_1^{(1)} & \alpha_0^{(1)} & 0 & 0 & \cdots \\ \vdots & & & & \\ \cdots & 0 & \alpha_{k_{N-1}}^{(N-1)} & \cdots & \alpha_0^{(N-1)} \end{pmatrix}.$$

Equation (6.4) *holds if and only if the term in the squared brackets vanishes. Since $\boldsymbol{A}$ is lower triangular, each $\boldsymbol{y}_{n+1}$ is determined directly from $\boldsymbol{y}_\mathrm{s}, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$ by the nth equation of the squared brackets term in* (6.4) *which coincides with the nth step of* (2.2b)*. So, together with the equivalence between* (2.2a) *and* (6.3c) *the lemma is shown.* □

Lemma 6.3 *For the solution $\boldsymbol{y}^h(t)$ of* (6.3b)-(6.3c)*, the equation* (6.3a) *with $\boldsymbol{l}^h = \boldsymbol{\lambda}_0$ is equivalent to the discrete adjoint IND scheme* (3.2) *of the nominal BDF scheme* (2.2)*.*

Proof *Analogously to the beginning of the proof of Lemma 6.2, each integral in* (6.3a) *is given by*

$$\int_{I_n} \dot{\boldsymbol{w}}^h(t) - \boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}^h(t)) \boldsymbol{w}^h(t) \, \mathrm{d}\boldsymbol{\Lambda}^h(t) = \boldsymbol{\lambda}_{n+1}^\intercal \left\{ \sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{w}_{n+1-i} - h_n \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1}) \boldsymbol{w}_{n+1} \right\}.$$

Thus, equation (6.3a) *can be formulated equivalently in matrix form with $\boldsymbol{w}^\intercal := \begin{bmatrix} \boldsymbol{w}_1^\intercal & \boldsymbol{w}_2^\intercal & \cdots & \boldsymbol{w}_N^\intercal \end{bmatrix} \in \left(\mathbb{R}^d\right)^N$ and $\boldsymbol{\lambda}^\intercal := \begin{bmatrix} \boldsymbol{\lambda}_1^\intercal & \boldsymbol{\lambda}_2^\intercal & \cdots & \boldsymbol{\lambda}_N^\intercal \end{bmatrix}$*

$$\begin{bmatrix} \boldsymbol{0} & \cdots & \boldsymbol{0} & J'(\boldsymbol{y}_N) \end{bmatrix} \boldsymbol{w} - \left[\alpha_1^{(0)} \boldsymbol{\lambda}_1 + \boldsymbol{l}^h \right]^\intercal \boldsymbol{w}_0 - \boldsymbol{\lambda}^\intercal \left[\boldsymbol{A} \otimes \boldsymbol{I} - \begin{pmatrix} h_0 \boldsymbol{f}_{\boldsymbol{y}}(t_1, \boldsymbol{y}_1) & & \boldsymbol{0} \\ & \ddots & \\ \boldsymbol{0} & & h_{N-1} \boldsymbol{f}_{\boldsymbol{y}}(t_N, \boldsymbol{y}_N) \end{pmatrix} \right] \boldsymbol{w} = 0, \ \forall \boldsymbol{w}_0, \boldsymbol{w} \tag{6.5}$$

which is linear in both the variations $\boldsymbol{w}_0$, $\boldsymbol{w}$ *and the unknown* $\boldsymbol{\lambda}$. *The equivalent time-stepping scheme goes backwards in time starting with* $J'(\boldsymbol{y}_N) - \alpha_0^{(N-1)}\boldsymbol{\lambda}_N^\intercal + h_{N-1}\boldsymbol{\lambda}_N^\intercal \boldsymbol{f}_{\boldsymbol{y}}(t_N, \boldsymbol{y}_N) = 0$. *Thus,* (6.5) *with* $\boldsymbol{l}^h = \boldsymbol{\lambda}_0$ *is equivalent to* (3.2) *which finishes the proof.* □

It remains to spend some words on the well-posedness of the Petrov-Galerkin equations (6.3). The system (6.3b)-(6.3c) admits a unique solution $\boldsymbol{y}^h$ via $\{\boldsymbol{y}_n\}_{n=0}^N$ if $\left|h_n/\alpha_0^{(n)}\, L\right| < 1$ for $n = 0, \dots, N-1$ with Lipschitz constant L of $\boldsymbol{f}(t, \boldsymbol{y})$, or less restrictive, if $\left|h_n/\alpha_0^{(n)}\right| \|\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})\| < 1$ for all n, see Chapter 2. The solution depends continuously on the input data due to the zero- and $A(\alpha)$-stability of the integration scheme, cf. Section 2.3. Since $\|\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y})\|$ is bounded by L for all $(t, \boldsymbol{y})$ and h_n, k_n satisfy $\left|h_n/\alpha_0^{(n)}\, L\right| < 1$, the matrix in (6.5) is nonsingular and thus (6.3a) possesses a unique weak adjoint solution $\boldsymbol{\Lambda}^h$ via $\{\boldsymbol{\lambda}_n\}_{n=0}^N$. The solution depends continuously on the input data since the zero-stability of the nominal integration scheme (2.2) is carried over to the discrete adjoint IND scheme (3.2), see Section 3.6. The well-posedness of (6.3a) is also established by the derivation of the equivalent scheme (3.2) using Algorithmic Differentiation (AD) of (2.2), cf. Section 3.4.1.

Interpretation of the FE weak adjoint The FE weak adjoint $\boldsymbol{\Lambda}^h$ defined by (6.2) represents the numerical quadrature of the adjoint IND values $\{\boldsymbol{\lambda}_n\}_{n=0}^N$, i.e. the discrete counterpart of the integration in (5.7). The weighting with h_{n-1} guarantees that constant functions are integrated exactly which is the fundamental property of quadrature formulas. An example is provided by the discrete BDF adjoint λ_1^h in Figure 3.1(a) and the resulting FE weak adjoint Λ_1^h in Figure 6.4(a). Like in the infinite dimensional case described on page 52, the FE weak adjoint $\boldsymbol{\Lambda}^h$ represents the accumulation of all adjoint IND values from t_s to t for any time $t \in [t_\mathrm{s}, t_\mathrm{f}]$ due to (6.2). Furthermore, $\boldsymbol{\Lambda}^h$ describes the discrete stability of the system of equations given by (6.3c)-(6.3b).

6.3 Commutativity of differentiation and discretization

In Section 3.5 we highlighted the advantages and disadvantages of the discretize-then-differentiate and the differentiate-then-discretize approach to generate derivatives of solutions of Initial Value Problems (IVPs) with respect to initial values. Furthermore, we highlighted the desirable property that both approaches lead to the same discrete system, i.e. that discretization and differentiation commute. Unfortunately, in Section 3.6 we saw that the discrete adjoint IND schemes of BDF methods are not consistent discretizations of the adjoint IVP in the classical sense. However, with the novel functional-analytic framework of Chapter 5 and the Petrov-Galerkin FE discretization of this chapter we now obtain the commutativity of differentiation and discretization also for the multistep BDF methods. The following considerations can be seen as two separate argumentative ways to show the commutativity of

discretization and differentiation.

Discretize-then-differentiate approach The BDF method (2.2) with prescribed orders and stepsizes can be understood as discretization of the constraints of the infinite dimensional Constrained Variational Problem (CVP) (5.1) to end up with a finite dimensional Nonlinear Program (NLP)

$$\min_{\boldsymbol{y}_0,\dots,\boldsymbol{y}_N} \quad J(\boldsymbol{y}_N) \tag{6.6a}$$

$$\text{s. t.} \quad \sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} = h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}),\ n = 0, \dots, N-1 \tag{6.6b}$$

$$\boldsymbol{y}_0 = \boldsymbol{y}_\mathrm{s}. \tag{6.6c}$$

By the introduction of Lagrange multipliers $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ the Lagrangian of (6.6) reads

$$\mathcal{L}(\boldsymbol{y}_0, \dots, \boldsymbol{y}_N, \boldsymbol{\lambda}_0, \dots, \boldsymbol{\lambda}_N) := J(\boldsymbol{y}_N) - \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^\intercal \left(\sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} - h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) \right) - \boldsymbol{\lambda}_0^\intercal (\boldsymbol{y}_0 - \boldsymbol{y}_\mathrm{s}). \tag{6.7}$$

The necessary conditions, i.e. the Karush-Kuhn-Tucker conditions, for a stationary point $(\boldsymbol{y}_0, \dots, \boldsymbol{y}_N, \boldsymbol{\lambda}_0, \dots, \boldsymbol{\lambda}_N)$ of the Lagrangian are that the first order derivative of $\mathcal{L}$ vanishes, i.e. $\mathcal{L}'(\boldsymbol{y}_0, \dots, \boldsymbol{y}_N, \boldsymbol{\lambda}_0, \dots, \boldsymbol{\lambda}_N) = \boldsymbol{0}$ has to be satisfied, see e.g. Fletcher [59] and Nocedal and Wright [98]. The conditions on the derivative with respect to the Lagrange multipliers are the BDF method (2.2) itself with prescribed orders and stepsizes. The conditions on the derivative with respect to the nominal approximations are provided by the adjoint IND scheme (3.2) associated to (2.2). In Figure 6.3 this *discrete* approach is visualized by the upper arrow pointing to the right and the downward-pointing arrow at the right hand side.

From this point of view, the adjoint IND value $\boldsymbol{\lambda}_{n+1}$ describe the rate of change of the optimal value $J(\boldsymbol{y}_N)$ with respect to changes in the n-th constraint, i.e. with respect to a perturbation $\boldsymbol{\delta}_{n+1}$ of the constraint

$$\boldsymbol{c}_n(\boldsymbol{y}_0, \dots, \boldsymbol{y}_N) := \sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} - h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) = \boldsymbol{0}.$$

The perturbation $\boldsymbol{\delta}_{n+1}$ affects the future approximations $\boldsymbol{y}_{n+1}, \dots, \boldsymbol{y}_N$ and therefore also the objective value which becomes, in first order, $J(\boldsymbol{y}_N) - \boldsymbol{\lambda}_{n+1}^\intercal \boldsymbol{\delta}_{n+1}$, i.e. $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ describes the discrete stability of (2.2).

We obtain the same discrete Lagrangian as defined in (6.7) by inserting the FE approximations of Section 6.1 into the extended Lagrangian defined by (5.10)

$$\begin{aligned}
\widehat{\mathcal{L}}\left(\boldsymbol{y}^h, \boldsymbol{\Lambda}^h, \boldsymbol{l}^h\right) &= J(\boldsymbol{y}_N) - \sum_{n=0}^{N-1} h_n \boldsymbol{\lambda}_{n+1}^\intercal \left\{ \dot{\boldsymbol{y}}^h(t_{n+1}) - \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) \right\} - (\boldsymbol{l}^h)^\intercal [\boldsymbol{y}_0 - \boldsymbol{y}_\mathrm{s}] \\
&= J(\boldsymbol{y}_N) - \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^\intercal \left\{ \sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} - \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) \right\} - (\boldsymbol{l}^h)^\intercal [\boldsymbol{y}_0 - \boldsymbol{y}_\mathrm{s}]
\end{aligned}$$

see the beginning of the proof of Lemma 6.2. Hence, the function space discretization of $C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ and $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ by $Y_\mathcal{P}[t_\mathrm{s}, t_\mathrm{f}]^d$ and $Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d$ with a subsequent differentiation also describes the upper arrow pointing to the right and the downward-pointing arrow at the right hand side of Figure 6.3. The resulting finite dimensional optimality conditions of this discrete adjoint approach are the BDF method (2.2) together with its associated adjoint IND scheme (3.2).

Differentiate-then-discretize approach In this approach we first solve the CVP (5.1) in $C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ to obtain the infinite dimensional optimality conditions (5.6), as done in Section 5.2. Subsequently, the FE spaces of Section 6.1 are used for discretization to end up in the finite dimensional optimality conditions (6.3). Hence, we interpret the finite system (6.3) as a *non-conformal* discretization of (5.6) since $Y_\mathcal{P}[t_\mathrm{s}, t_\mathrm{f}]^d \not\subset C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ whereas $Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d \subset \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$. For a definition of non-conformity we refer e.g. to Ern and Guermond [57] and Großmann and Roos [66]. Nevertheless, the approximation setting comprising $\widehat{\mathcal{L}}(\cdot,\cdot,\cdot)$, $Y_\mathcal{P}[t_\mathrm{s}, t_\mathrm{f}]^d$ and $Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d$ is *consistent* since the exact solution $\boldsymbol{y}$ of (5.6c)-(5.6b) fulfills the discrete system (6.3c)-(6.3b), as we show in the following lemma.

Lemma 6.4 *The FE discretization of the nominal infinite dimensional system* (5.6c)-(5.6b) *using* $\widehat{\mathcal{L}}_{(\boldsymbol{\Lambda},\boldsymbol{l})}(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})(\cdot,\cdot)$ *and the FE spaces* $Y_\mathcal{P}[t_\mathrm{s}, t_\mathrm{f}]^d$ *and* $Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d$ *is consistent, i.e. the exact solution* $\boldsymbol{y} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ *of* (5.6c)-(5.6b) *also satisfies the nominal FE discretization* (6.3c)-(6.3b)

$$\widehat{\mathcal{L}}_{(\boldsymbol{\Lambda},\boldsymbol{l})}(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})(\boldsymbol{\Gamma}^h,\boldsymbol{r}^h) = 0 \quad \forall (\boldsymbol{\Gamma}^h,\boldsymbol{r}^h) \in Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$$

with arbitrary $\boldsymbol{\Lambda} \in \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ *and* $\boldsymbol{l} \in \mathbb{R}^d$.

Proof *It is* $\widehat{\mathcal{L}}_{\boldsymbol{l}}(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})(\boldsymbol{r}^h) = -[\boldsymbol{r}^h]^\intercal[\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}] = -[\boldsymbol{r}^h]^\intercal \boldsymbol{0} = 0$ *since* $\boldsymbol{y}$ *solves* (5.6c). *Furthermore, it holds*

$$\begin{aligned}\widehat{\mathcal{L}}_{\boldsymbol{\Lambda}}(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})(\boldsymbol{\Gamma}^h) &= -\sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t,\boldsymbol{y}(t))\,\mathrm{d}\boldsymbol{\Gamma}^h(t) \\ &= -\sum_{n=0}^{N-1} \left[\dot{\boldsymbol{y}}(t_{n+1}) - \boldsymbol{f}(t_{n+1},\boldsymbol{y}(t_{n+1}))\right] \boldsymbol{\gamma}_{n+1} h_n = 0\end{aligned}$$

since $\boldsymbol{y}$ *solves* (5.6b) *and hence* (1.1a) *due to the beginning of Section 5.2.* □

Remark 6.5 *The so-called* nonlinear Galerkin orthogonality*, cf. Bangerth and Rannacher [15], is satisfied, if*

$$\widehat{\mathcal{L}}_{(\boldsymbol{\Lambda},\boldsymbol{l})}(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})(\boldsymbol{\Gamma}^h,\boldsymbol{r}^h) - \widehat{\mathcal{L}}_{(\boldsymbol{\Lambda},\boldsymbol{l})}(\boldsymbol{y}^h,\boldsymbol{\Lambda},\boldsymbol{l})(\boldsymbol{\Gamma}^h,\boldsymbol{r}^h) = 0 \quad \forall (\boldsymbol{\Gamma}^h,\boldsymbol{r}^h) \in Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d,$$

where $\boldsymbol{y} \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d$ *solves* (5.6c)-(5.6b) *and* $\boldsymbol{y}^h \in Y_\mathcal{P}[t_\mathrm{s}, t_\mathrm{f}]^d$ *solves* (6.3c)-(6.3b). *The consistency obtained by Lemma 6.4 immediately implies that the nonlinear Galerkin orthogonality holds.*

The *adjoint consistency* as defined in Oliver and Darmofal [99] is provided by the following lemma. Adjoint consistency for discontinuous Galerkin methods has also been analyzed by Hartmann [70]. It is an important property concerning the commutativity of differentiation and discretization since it gurarantees that the exact solution of (5.6) also satisfies the discrete adjoint system (6.3a).

Lemma 6.6 *The FE discretization of* (5.6c)-(5.6b) *using* $\widehat{\mathcal{L}}_{(\boldsymbol{\Lambda},\boldsymbol{l})}(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})(\cdot,\cdot)$ *and the FE spaces* $Y_{\mathcal{P}}[t_\mathrm{s},t_\mathrm{f}]^d$ *and* $Z_\mathrm{H}[t_\mathrm{s},t_\mathrm{f}]^d$ *is adjoint consistent, i.e. the solution* $(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l}) \in C^1[t_\mathrm{s},t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d \times \mathbb{R}^d$ *of* (5.6) *satisfies the adjoint FE discretization* (6.3a)

$$\widehat{\mathcal{L}}_{\boldsymbol{y}}(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})(\boldsymbol{w}^h) = 0 \quad \forall \boldsymbol{w}^h \in Y_{\mathcal{P}}[t_\mathrm{s},t_\mathrm{f}]^d.$$

Proof *Since* $(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l})$ *solves* (5.6), *we have* $\boldsymbol{\Lambda}(t) = \int_{t_\mathrm{s}}^t \boldsymbol{\lambda}(\tau)\mathrm{d}\tau$, $\boldsymbol{l} = \boldsymbol{\lambda}(t_\mathrm{s})$ *where* $\boldsymbol{\lambda}(t)$ *solves* (1.8) *and is continuously differentiable, cf. Section 5.2 and 1.2.2. Hence, for the Fréchet derivative of* $\widehat{\mathcal{L}}$ *defined by* (5.10) *with respect to* $\boldsymbol{y}$ *in direction* $\boldsymbol{w}^h$ *evaluated at the nominal solution* $\boldsymbol{y}$ *we obtain due to Section 5.3.1 and 4.1.2 as well as integration by parts*

$$\begin{aligned}
&\widehat{\mathcal{L}}_{\boldsymbol{y}}\left(\boldsymbol{y},\boldsymbol{\Lambda},\boldsymbol{l}\right)\left(\boldsymbol{w}^h\right)\\
&= J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}^h(t_\mathrm{f}) - \sum_{n=0}^{N-1}\int_{I_n} \boldsymbol{\lambda}^\intercal(t)\left[\dot{\boldsymbol{w}}^h(t) - \boldsymbol{f}_{\boldsymbol{y}}(t,\boldsymbol{y}(t))\boldsymbol{w}^h(t)\right]\mathrm{d}t - \boldsymbol{\lambda}^\intercal(t_\mathrm{s})\boldsymbol{w}^h(t_\mathrm{s})\\
&= J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{w}^h(t_\mathrm{f}) - \boldsymbol{\lambda}^\intercal(t_\mathrm{f})\boldsymbol{w}^h(t_\mathrm{f}) + \boldsymbol{\lambda}^\intercal(t_\mathrm{s})\boldsymbol{w}^h(t_\mathrm{s})\\
&\quad + \sum_{n=0}^{N-1}\int_{I_n}\left[\dot{\boldsymbol{\lambda}}(t) + \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t,\boldsymbol{y}(t))\boldsymbol{\lambda}(t)\right]^\intercal \boldsymbol{w}^h(t)\mathrm{d}t - \boldsymbol{\lambda}^\intercal(t_\mathrm{s})\boldsymbol{w}^h(t_\mathrm{s})\\
&= \left[J'(\boldsymbol{y}(t_\mathrm{f})) - \boldsymbol{\lambda}^\intercal(t_\mathrm{f})\right]\boldsymbol{w}^h(t_\mathrm{f}) + \int_{t_\mathrm{s}}^{t_\mathrm{f}}\left[\dot{\boldsymbol{\lambda}}(t) + \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t,\boldsymbol{y}(t))\boldsymbol{\lambda}(t)\right]^\intercal \boldsymbol{w}^h(t)\mathrm{d}t = 0.
\end{aligned}$$

The penultimate equality holds due to Section 5.3.1 since the integrand of the extended Riemann-Stieltjes integral is continuous on the whole interval $[t_\mathrm{s},t_\mathrm{f}]$. *The last equality holds since* $\boldsymbol{\lambda}$ *solves* (1.8). □

In Figure 6.3 the differentiate-then-discretize approach is depicted by the downward-pointing arrow at the left hand side being followed by the lower arrow pointing to the right. The resulting finite dimensional optimality conditions of this continuous adjoint approach are provided by (6.3).

Commutativity Due to Theorem 6.1 the resulting discrete systems of equations of both adjoint approaches, the discrete and the continuous one, are equivalent. Thus, discretization and differentiation commute in the novel functional-analytic setting of Chapter 5 and, in Figure 6.3, both paths from the infinite dimensional CVP (upper left corner) lead to the same finite dimensional system of equations (lower right corner) also in the case of multistep BDF methods.

These properties can also be observed numerically. To this end, we recall the example of Section 3.6 with analytic solutions: The nonlinear Catenary problem solved

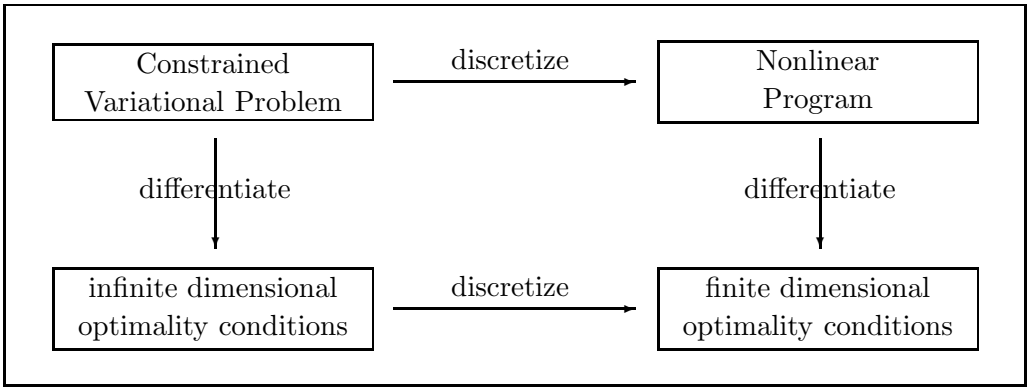
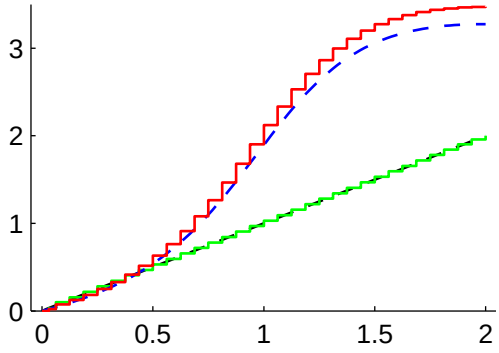


Figure 6.3: The two ways to transfer the infinite dimensional CVP (upper left corner) into a finite dimensional optimality system (lower right corner). The *discrete* approach first discretizes the CVP to give an NLP that is then differentiated. The *continuous* approach first differentiates the CVP to give infinite dimensional optimality conditions that are then discretized.

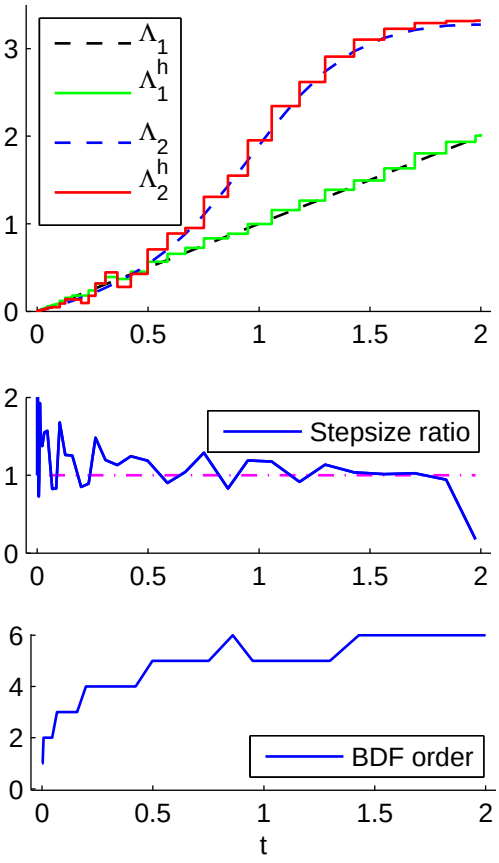
by a constant BDF method with order 2 and the variable BDF method `DAESOL-II`. We use the discrete adjoint IND values $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ generated by the corresponding adjoint IND schemes (3.2) and depicted in Figure 3.1 to compute by (6.2) the FE approximation $\boldsymbol{\Lambda}^h(t)$ to the weak adjoint $\boldsymbol{\Lambda}(t)$. The resulting FE weak adjoints are depicted in Figure 6.4.

Also in areas of varying stepsizes and changing orders, cf. penultimate and last row of Figure 6.4(b), the FE weak adjoint gives a good approximation to the exact weak adjoint. Remember that this was not fulfilled for the classical adjoint where huge oscillations of the adjoint IND values are present, cf. first row of Figure 3.1(b).

At the end of Chapter 5 we emphasized that the novel functional-analytic framework holds generally for IVPs with continuous or piecewise continuous defects. To utilize the framework for the analysis of different integration methods than BDF methods the choice of basis functions is cruicial. The basis functions chosen in Section 6.1 particularly fit to BDF methods and their discrete adjoint IND schemes.



(a) Constant BDF method



(b) Variable BDF method

Figure 6.4: Comparison of the FE weak adjoint $\boldsymbol{\Lambda}^h = [\Lambda_1^h, \Lambda_2^h]^\intercal$ and the analytic weak adjoint solution $\boldsymbol{\Lambda} = [\Lambda_1, \Lambda_2]^\intercal$ of the adjoint IVP on the Catenary test case. Stepsize ratio (penultimate row) and BDF order (bottom) of the variable BDF method.

7 Convergence analysis for discrete adjoints

To finish the investigation of the relation between the discrete adjoint Internal Numerical Differentiation (IND) values and the solution of the adjoint Initial Value Problem (IVP) via the weak adjoints defined in Chapter 5 and Chapter 6 we have to quantify the approximation quality of $\mathbf{\Lambda}^h \in Z_{\mathrm{H}}[t_{\mathrm{s}}, t_{\mathrm{f}}]^d$ to $\mathbf{\Lambda} \in \mathrm{NBV}[t_{\mathrm{s}}, t_{\mathrm{f}}]^d$. To this end, we demonstrate the convergence of $\mathbf{\Lambda}^h$ to $\mathbf{\Lambda}$ in the total variation norm of $\mathrm{NBV}[t_{\mathrm{s}}, t_{\mathrm{f}}]^d$ which directly implies the convergence of $\mathbf{\Lambda}^h(t)$ to $\mathbf{\Lambda}(t)$ at any time $t \in [t_{\mathrm{s}}, t_{\mathrm{f}}]$. This will be the subject of the second part of the chapter. As preparation for the convergence proof in $\mathrm{NBV}[t_{\mathrm{s}}, t_{\mathrm{f}}]^d$ we first show the convergence of the discrete IND adjoint values $\boldsymbol{\lambda}_n$ to $\boldsymbol{\lambda}(t_n)$ on the open interval $(t_{\mathrm{s}}, t_{\mathrm{f}})$. Therefore, we consider a constant Backward Differentiation Formula (BDF) method with order k and stepsize h using a self-starting procedure for $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ with $m \geq k-1$ fixed.

7.1 Convergence of discrete adjoints of BDF methods

The discrete adjoint IND scheme (3.2) of a constant BDF method reads (see Section 3.4.1)

$$\alpha_0 \boldsymbol{\lambda}_N - J'(\boldsymbol{y}_N)^\intercal = h \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t_N, \boldsymbol{y}_N) \boldsymbol{\lambda}_N \tag{7.1a}$$

$$\sum_{i=0}^{N-1-n} \alpha_i \boldsymbol{\lambda}_{n+1+i} = h \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t_{n+1}, \boldsymbol{y}_{n+1}) \boldsymbol{\lambda}_{n+1}, \; n = N-2, \ldots, N-k \tag{7.1b}$$

$$\sum_{i=0}^{k} \alpha_i \boldsymbol{\lambda}_{n+1+i} = h \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t_{n+1}, \boldsymbol{y}_{n+1}) \boldsymbol{\lambda}_{n+1}, \; n = N-k-1, \ldots, m \tag{7.1c}$$

$$\sum_{i=0}^{k} \alpha_i^{(n+i)} \boldsymbol{\lambda}_{n+1+i} = h \boldsymbol{f}_{\boldsymbol{y}}^\intercal(t_{n+1}, \boldsymbol{y}_{n+1}) \boldsymbol{\lambda}_{n+1}, \; n = m-1, \ldots, 0 \tag{7.1d}$$

$$\boldsymbol{\lambda}_0 + \alpha_1^{(0)} \boldsymbol{\lambda}_1 = \mathbf{0} \tag{7.1e}$$

where (7.1d) accounts for the nominal starting procedure. As shown in Lemma 3.4 the scheme (7.1) is a Linear Multistep Method (LMM) applied to the perturbed adjoint IVP (3.5). According to Lemma 3.5 the main steps (7.1c) are consistent with (3.5). The adjoint initialization steps (7.1a)-(7.1b) can be now interpreted as a starting procedure for (7.1c) giving inconsistent start values $\boldsymbol{\lambda}_N, \ldots, \boldsymbol{\lambda}_{N-k+1}$.

In the following, we study the asymptotic behavior for $h \to 0$ and $n \to \infty$ such

that $t_n = t_\mathrm{s} + nh$ remains fixed. The interval $[t_{m+1}, t_{N-k}]$ of the main part of (7.1) increases and approaches $(t_\mathrm{s}, t_\mathrm{f})$ for $h \to 0$.

Lemma 7.1 *Let $\boldsymbol{f_y}(t, \tilde{\boldsymbol{y}}(t))$ be continuously differentiable in $t \in [t_\mathrm{s}, t_\mathrm{f}]$ and $\tilde{\boldsymbol{y}}(t_n) = \boldsymbol{y}_n$ for $n = 0, \dots, N$ where $\{\boldsymbol{y}_n\}_{n=0}^N$ is generated by the constant BDF method with order k and stepsize h. Let $\tilde{\boldsymbol{\lambda}}(t)$ be the exact solution of the perturbed adjoint IVP (3.5) along $\tilde{\boldsymbol{y}}(t)$ and let $\{\boldsymbol{\lambda}_n\}_{n=1}^N$ be generated by (7.1). Then, for $t_n \in (t_\mathrm{s}, t_\mathrm{f})$ fixed there exists $H > 0$ such that*

$$\left\| \boldsymbol{\lambda}_n - \tilde{\boldsymbol{\lambda}}(t_n) \right\| = \mathcal{O}(h)$$

as the stepsize is reduced with $H > h \to 0$.

Proof *To ease the notion, we consider a scalar IVP. Nevertheless, the proof is also valid for systems of IVPs. Furthermore, we define some abbreviations $B(t) := f_y^\intercal(t, \tilde{y}(t))$ and $\eta := J'(\tilde{y}(t_\mathrm{f}))^\intercal$. Thus, the starting procedure (7.1a)-(7.1b) can be written equivalently using $\boldsymbol{\lambda}^\intercal := \begin{bmatrix} \lambda_N & \cdots & \lambda_{N-k+1} \end{bmatrix}$ and the $k \times 1$ unit vector $\boldsymbol{e}_1$*

$$\left[\tilde{\boldsymbol{A}} - h\boldsymbol{B}(t_N, h) \right] \boldsymbol{\lambda} = \boldsymbol{e}_1 \eta$$

where $\tilde{\boldsymbol{A}} = \bar{\boldsymbol{I}} \left[\boldsymbol{A}_{N-k+1:N, N-k+1:N} \right]^\intercal \bar{\boldsymbol{I}}$ for the reverse identity matrix $\bar{\boldsymbol{I}}$ and the matrix $\boldsymbol{A}$ from page 62, and

$$\boldsymbol{B}(t_N, h) := \begin{pmatrix} B(t_N) & & 0 \\ & \ddots & \\ 0 & & B(t_N - (k-1)h) \end{pmatrix} = B(t_N)\boldsymbol{I} + \mathcal{O}(h) \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 1 & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & 1 \end{pmatrix}$$

using the Taylor series expansion of the entries $B(t_N - ih)$ around t_N. The matrix $\tilde{\boldsymbol{A}}$ is nonsingular since $\alpha_0 \neq 0$. Furthermore, for h small enough to satisfy $\left\| h\tilde{\boldsymbol{A}}^{-1}\boldsymbol{B}(t_N, h) \right\| < 1$ we can express the inverse of $\boldsymbol{I} - h\tilde{\boldsymbol{A}}^{-1}\boldsymbol{B}(t_N, h)$ by its Neumann series (see Theorem A.4) to obtain

$$\begin{aligned} \boldsymbol{\lambda} &= \left[\tilde{\boldsymbol{A}} \left(\boldsymbol{I} - h\tilde{\boldsymbol{A}}^{-1}\boldsymbol{B}(t_N, h) \right) \right]^{-1} \boldsymbol{e}_1 \eta = \left\{ \sum_{j=0}^{\infty} \left(h\tilde{\boldsymbol{A}}^{-1}\boldsymbol{B}(t_N, h) \right)^j \right\} \tilde{\boldsymbol{A}}^{-1}\boldsymbol{e}_1 \eta \\ &= \left\{ \boldsymbol{I} + h\tilde{\boldsymbol{A}}^{-1}\boldsymbol{B}(t_N, h) + \mathcal{O}(h^2) \right\} \tilde{\boldsymbol{A}}^{-1}\boldsymbol{e}_1 \eta \\ &= \tilde{\boldsymbol{A}}^{-1}\boldsymbol{e}_1 \eta + h\tilde{\boldsymbol{A}}^{-1} B(t_N) \tilde{\boldsymbol{A}}^{-1}\boldsymbol{e}_1 \eta + \mathcal{O}(h^2). \end{aligned} \tag{7.2}$$

Due to (7.2) the starting procedure satisfies the assumption (2.14) of Theorem 2.24 applied to the linear IVP (3.5) with continuously differentible coefficient $B(t) = f_y(t, \tilde{y}(t))$. Thus, Theorem 2.24 yields for certain constants K_1 and K_2

$$\lambda_n - \tilde{\lambda}(t_n) = \exp\left(\int_{t_\mathrm{f}}^{t_n} -B(\tau)\mathrm{d}\tau \right) \zeta + \theta \left(K_1 + \frac{K_2}{t_n - h - t_\mathrm{f}} \right) h$$

where $|\theta| < 1$ *and*

$$\zeta := \frac{1}{\rho'(1)} \sum_{i=0}^{k-1} \gamma_i (\lambda_{N-i} - \eta).$$

The coefficients γ_i *sum up to 1, i.e.* $\sum_{i=0}^{k-1} \gamma_i = 1$*, such that together with* (7.2) *we obtain for* $\boldsymbol{\gamma}^\intercal := \begin{bmatrix} \gamma_0 & \cdots & \gamma_{k-1} \end{bmatrix}$

$$\begin{aligned} \zeta = \boldsymbol{\gamma}^\intercal \boldsymbol{\lambda} - \eta &= \boldsymbol{\gamma}^\intercal \left[\tilde{\boldsymbol{A}}^{-1} \boldsymbol{e}_1 \eta + h \tilde{\boldsymbol{A}}^{-1} B(t_N) \tilde{\boldsymbol{A}}^{-1} \boldsymbol{e}_1 \eta + \mathcal{O}(h^2) \right] - \eta \\ &= \left[\boldsymbol{\gamma}^\intercal \tilde{\boldsymbol{A}}^{-1} \boldsymbol{e}_1 - 1 \right] \eta + h \boldsymbol{\gamma}^\intercal \tilde{\boldsymbol{A}}^{-1} B(t_N) \tilde{\boldsymbol{A}}^{-1} \boldsymbol{e}_1 \eta + \mathcal{O}(h^2). \end{aligned}$$

The coefficient $\boldsymbol{\gamma}^\intercal \tilde{\boldsymbol{A}}^{-1} \boldsymbol{e}_1 - 1$ *of the first addend vanishes which can be verified easily for all zero-stable BDF methods (i.e.* $k \leq 6$ *according to Theorem 2.19). Thus, we obtain*

$$\begin{aligned} \lambda_n - \tilde{\lambda}(t_n) = h \exp\left(\int_{t_n}^{t_\mathrm{f}} B(\tau) \mathrm{d}\tau \right) \boldsymbol{\gamma}^\intercal \tilde{\boldsymbol{A}}^{-1} B(t_N) \tilde{\boldsymbol{A}}^{-1} \boldsymbol{e}_1 \eta \\ + h\, \theta \left(K_1 + \frac{K_2}{t_n - h - t_\mathrm{f}} \right) + \mathcal{O}(h^2) \end{aligned} \tag{7.3}$$

where both coefficients are bounded. □

The main result of this section is the following.

Theorem 7.2 *Let* $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y})$ *be continuously differentiable with respect to* $(t, \boldsymbol{y})$*. Let* $\boldsymbol{\lambda}(t)$ *be the exact solution of the adjoint IVP* (1.8) *and let* $\{\boldsymbol{\lambda}_n\}_{n=1}^N$ *be generated by* (7.1)*. Then, for* $t_n \in (t_\mathrm{s}, t_\mathrm{f})$ *fixed there exists* $H > 0$ *such that*

$$\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}(t_n)\| = \mathcal{O}(h) \tag{7.4}$$

as the stepsize is reduced with $H > h \to 0$*.*

Proof *Let the continuously differentiable spline* $\tilde{\boldsymbol{y}}(t)$ *be composed of quadratic polynomials on* I_n *such that* $\tilde{\boldsymbol{y}}(t_n) = \boldsymbol{y}_n$*,* $\tilde{\boldsymbol{y}}(t_{n+1}) = \boldsymbol{y}_{n+1}$ *and* $\dot{\tilde{\boldsymbol{y}}}(t_{n+1}) = \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})$ *for* $n = 0, \ldots, N-1$*. Furthermore, we define the interpolation operator* $\boldsymbol{\mathcal{I}}$ *that maps a continuously differentiable function* $\boldsymbol{g}(t)$ *to a continuously differentiable spline* $\boldsymbol{\mathcal{I}}\boldsymbol{g}(t)$ *that is composed of quadratic polynomials on* I_n *with* $\boldsymbol{\mathcal{I}}\boldsymbol{g}(t_n) = \boldsymbol{g}(t_n)$*,* $\boldsymbol{\mathcal{I}}\boldsymbol{g}(t_{n+1}) = \boldsymbol{g}(t_{n+1})$ *and* $\dot{\boldsymbol{\mathcal{I}}}\boldsymbol{g}(t_{n+1}) = \dot{\boldsymbol{g}}(t_{n+1})$ *for* $n = 0, \ldots, N-1$*. Then, the difference of* $\tilde{\boldsymbol{y}}(t)$ *and* $\boldsymbol{\mathcal{I}}\boldsymbol{y}(t)$ *in the* C^0*-norm is*

$$\|\tilde{\boldsymbol{y}}(t) - \boldsymbol{\mathcal{I}}\boldsymbol{y}(t)\|_{C^0[t_\mathrm{s}, t_\mathrm{f}]^d} = \mathcal{O}(h)$$

using Taylor series expansions and the at least linear convergence of the nominal constant BDF method with self-starter (cf. Theorem 2.21 or Shampine and Zhang

[112]). Due to the assumption on $\boldsymbol{f}(t, \boldsymbol{y})$*, the exact nominal solution* $\boldsymbol{y}(t)$ *of* (1.1) *is twice continuously differentiable such that*

$$\|\boldsymbol{y}(t) - \boldsymbol{\mathcal{I}}\boldsymbol{y}(t)\|_{C^0[t_\mathrm{s},t_\mathrm{f}]^d} = \mathcal{O}(h^2)$$

due to the approximation property of quadratic splines. Thus, it is

$$\|\tilde{\boldsymbol{y}}(t) - \boldsymbol{y}(t)\|_{C^0[t_\mathrm{s},t_\mathrm{f}]^d} \leq \|\tilde{\boldsymbol{y}}(t) - \boldsymbol{\mathcal{I}}\boldsymbol{y}(t)\|_{C^0} + \|\boldsymbol{\mathcal{I}}\boldsymbol{y}(t) - \boldsymbol{y}(t)\|_{C^0} = \mathcal{O}(h). \tag{7.5}$$

Hence, due to (3.6) *the perturbed adjoint solution* $\tilde{\boldsymbol{\lambda}}(t)$ *converges to* $\boldsymbol{\lambda}(t)$ *in the same manner*

$$\left\|\tilde{\boldsymbol{\lambda}}(t) - \boldsymbol{\lambda}(t)\right\|_{C^0[t_\mathrm{s},t_\mathrm{f}]^d} = \mathcal{O}(h) \tag{7.6}$$

which implies directly the pointwise convergence for every $t \in [t_\mathrm{s}, t_\mathrm{f}]$*. Since* $\boldsymbol{f}_{\boldsymbol{y}}(t, \tilde{\boldsymbol{y}}(t))$ *is continuously differentiable in* t*, Lemma 7.1 yields*

$$\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}(t_n)\| \leq \left\|\boldsymbol{\lambda}_n - \tilde{\boldsymbol{\lambda}}(t_n)\right\| + \left\|\tilde{\boldsymbol{\lambda}}(t_n) - \boldsymbol{\lambda}(t_n)\right\| = \mathcal{O}(h)$$

for $t_n \in (t_\mathrm{s}, t_\mathrm{f})$*.* □

Remark 7.3 *If* $\boldsymbol{f}(t, \boldsymbol{y})$ *is* k*-times continuously differentiable in* $(t, \boldsymbol{y})$*, the start errors of the nominal BDF method with order* k *are small enough (i.e. such that the convergence of order* k *is guaranteed, see Theorem 2.21 or Shampine and Zhang [112]), and the spline* $\tilde{\boldsymbol{y}}$ *is of corresponding order, then* (7.6) *holds with order* k *in* h*. However, this is not necessary for Theorem 7.2 since the limiting factor in the convergence order is Lemma 7.1.*

So far, we have considered the approximation properties of $\boldsymbol{\lambda}_n$ generated by (7.1) for $n = m+1, \cdots, N-k$, i.e. of those values defined on $[t_{m+1}, t_{N-k}] \subset [t_\mathrm{s}, t_\mathrm{f}]$. In the following we investigate the discrete adjoint IND values $\boldsymbol{\lambda}_n$ for $n = N, \ldots, N-k+1$ and $n = m, \ldots, 0$ resulting from the adjoint initialization and termination steps of (7.1).

Lemma 7.4 *If the same assumptions like in Theorem 7.2 hold, then, for* $n = N, \ldots, N-k+1$ *fixed, there exists a positive, bounded constant* c_n *such that*

$$\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}(t_n)\| \leq c_n \left\|J'(\boldsymbol{y}(t_\mathrm{f}))\right\| + \mathcal{O}(h)$$

with $\boldsymbol{\lambda}(t_\mathrm{f}) = J'(\boldsymbol{y}(t_\mathrm{f}))^\intercal$*.*

Proof *Due to the Taylor series expansion of* $\tilde{\boldsymbol{\lambda}}(t_n)$ *around* $t_N = t_\mathrm{f}$ *it is*

$$\tilde{\boldsymbol{\lambda}}(t_n) = \tilde{\boldsymbol{\lambda}}(t_\mathrm{f}) - (N-n)h \cdot \dot{\tilde{\boldsymbol{\lambda}}}(t_\mathrm{f}) + \mathcal{O}(h^2) = J'(\tilde{\boldsymbol{y}}(t_\mathrm{f}))^\intercal + \mathcal{O}(h) = J'(\boldsymbol{y}_N)^\intercal + \mathcal{O}(h).$$

Solving (7.1a) *and* (7.1c) *for* $\boldsymbol{\lambda}_n$ *with* $n = N, \dots, N-k+1$ *and using the Neumann series of* $(\boldsymbol{I} - h/\alpha_0 \boldsymbol{f}_{\boldsymbol{y}}(t_n, \boldsymbol{y}_n))^{-1}$ *we obtain*

$$\boldsymbol{\lambda}_N = (\alpha_0 \boldsymbol{I} - h\boldsymbol{f}_{\boldsymbol{y}}(t_N, \boldsymbol{y}_N))^{-\mathsf{T}} J'(\boldsymbol{y}_N)^{\mathsf{T}} = \frac{1}{\alpha_0} J'(\boldsymbol{y}_N)^{\mathsf{T}} + \mathcal{O}(h)$$

$$\boldsymbol{\lambda}_n = -(\alpha_0 \boldsymbol{I} - h\boldsymbol{f}_{\boldsymbol{y}}(t_n, \boldsymbol{y}_n))^{-\mathsf{T}} \sum_{i=1}^{N-n} \alpha_i \boldsymbol{\lambda}_{n+i} = -\frac{1}{\alpha_0} \sum_{i=1}^{N-n} \alpha_i \boldsymbol{\lambda}_{n+i} + \mathcal{O}(h)$$

such that $\boldsymbol{\lambda}_n$ *can be successively expressed in terms of* $J'(\boldsymbol{y}_N)^{\mathsf{T}}$. *Since* $J'(\boldsymbol{y}_N) = J'(\boldsymbol{y}(t_\mathrm{f})) + \mathcal{O}(\|\boldsymbol{y}_N - \boldsymbol{y}(t_\mathrm{f})\|) = J'(\boldsymbol{y}(t_\mathrm{f})) + \mathcal{O}(h)$ *due to the convergence of the nominal BDF method we exemplarily obtain for* $n = N$ *and* $n = N-1$ *that*

$$\left\| \tilde{\boldsymbol{\lambda}}(t_N) - \boldsymbol{\lambda}_N \right\| = \left| 1 - \frac{1}{\alpha_0} \right| \cdot \left\| J'(\boldsymbol{y}_N) \right\| + \mathcal{O}(h) = \left| 1 - \frac{1}{\alpha_0} \right| \cdot \left\| J'(\boldsymbol{y}(t_\mathrm{f})) \right\| + \mathcal{O}(h)$$

$$\left\| \tilde{\boldsymbol{\lambda}}(t_{N-1}) - \boldsymbol{\lambda}_{N-1} \right\| = J'(\boldsymbol{y}_N) + \frac{\alpha_1}{\alpha_0} \boldsymbol{\lambda}_N + \mathcal{O}(h) = \left| 1 + \frac{\alpha_1}{\alpha_0^2} \right| \cdot \left\| J'(\boldsymbol{y}(t_\mathrm{f})) \right\| + \mathcal{O}(h).$$

To complete the proof we again use that

$$\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}(t_n)\| \leq \left\| \boldsymbol{\lambda}_n - \tilde{\boldsymbol{\lambda}}(t_n) \right\| + \left\| \tilde{\boldsymbol{\lambda}}(t_n) - \boldsymbol{\lambda}(t_n) \right\| = \left\| \boldsymbol{\lambda}_n - \tilde{\boldsymbol{\lambda}}(t_n) \right\| + \mathcal{O}(h)$$

due to the triangle inequality and (7.6). □

For the discrete adjoint IND values from the adjoint termination steps (7.1d), the procedure is nearly the same.

Lemma 7.5 *If the same assumptions like in Theorem 7.2 hold, then, for* $n = m, \dots, 0$ *fixed, there exists a positive, bounded constant* c_n *such that*

$$\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}(t_n)\| \leq c_n \|\boldsymbol{\lambda}(t_\mathrm{s})\| + \mathcal{O}(h).$$

Proof *For* $n = m, \dots, 1$ *the new approximation* $\boldsymbol{\lambda}_m$ *determined by* (7.1d) *depends on the past values coming from the main part. Due to* (7.3) *in the proof of Lemma 7.1 with* $n \geq m+1$ *and the Taylor series expansion of* $\tilde{\boldsymbol{\lambda}}(t_n)$ *around* $t_0 = t_\mathrm{s}$ *we obtain*

$$\boldsymbol{\lambda}_n = \tilde{\boldsymbol{\lambda}}(t_n) + C_n h = \tilde{\boldsymbol{\lambda}}(t_\mathrm{s}) + \mathcal{O}(h).$$

Solving (7.1d) *for* $\boldsymbol{\lambda}_n$ *(*$n = m, \dots, 1$*) and using the above relation,* $\boldsymbol{\lambda}_n$ *can be expressed in terms of* $\tilde{\boldsymbol{\lambda}}(t_\mathrm{s})$. *Furthermore, we use again the Taylor series expansion of* $\tilde{\boldsymbol{\lambda}}(t_n)$ *around* $t_0 = t_\mathrm{s}$ *to obtain*

$$\left\| \tilde{\boldsymbol{\lambda}}(t_n) - \boldsymbol{\lambda}_n \right\| = \left\| \tilde{\boldsymbol{\lambda}}(t_\mathrm{s}) - \boldsymbol{\lambda}_n \right\| = c_n \left\| \tilde{\boldsymbol{\lambda}}(t_\mathrm{s}) \right\|$$

and finish in the same way like in the proof of Lemma 7.4. □

Since $\boldsymbol{\lambda}(t_\mathrm{s})$ and $J'(\boldsymbol{y}(t_\mathrm{f}))$ keep bounded by the assumptions on $\boldsymbol{f}(t, \boldsymbol{y})$ and J (cf. Chapter 1), the difference $\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}(t_n)\|$ remains bounded for $h \to 0$.

Remark 7.6 *If the stepsizes in the nominal self-starting procedure vary and are less than the constant setpsize, then the assertions of this section remain true but the adjoint main steps are those on* $[t_{m+k}, t_{N-k}]$ *and for the adjoint termination holds* $n \leq m + k - 1$*, cf. Remark 3.6.*

Without modifications of the adjoint initialization steps (7.1a)-(7.1b), we have demonstrated that the discrete adjoint IND values of the main part converge linearly to the exact adjoint solution $\boldsymbol{\lambda}(t)$ of (1.8). Nevertheless, we still have to consider the oscillations of the discrete adjoint IND values at the interval ends of $[t_\mathrm{s}, t_\mathrm{f}]$ which are due to the inconsistency of the adjoint initialization and termination steps, cf. Section 3.6. We will concentrate on this in the next section.

7.2 Convergence of FE weak adjoints

In this section we focus on the Finite Element (FE) approximation $\boldsymbol{\Lambda}^h$ of the weak adjoint and its convergence to the exact weak adjoint $\boldsymbol{\Lambda}$ of (1.8) resulting from (5.6a). We show the strong convergence, i.e. convergence in the total variation norm of $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ (see Definition 4.17). Nevertheless, the proof is based on the distance of $\boldsymbol{\Lambda}^h$ and $\boldsymbol{\Lambda}$ measured in the dual norm of $C^0[t_\mathrm{s}, t_\mathrm{f}]^d$ (see Section 4.3) which yields together with the Riesz Representation Theorem (Theorem 4.23) the strong convergence.

Theorem 7.7 *The FE approximation* $\boldsymbol{\Lambda}^h(t) = \sum_{n=1}^N h_{n-1} \boldsymbol{\lambda}_n H_n(t)$ *given by the discrete adjoint IND scheme* (7.1) *of a constant BDF method with order* k *and stepsize* h *converges to the exact weak adjoint solution* $\boldsymbol{\Lambda}(t) = \int_{t_\mathrm{s}}^t \boldsymbol{\lambda}(\tau)\mathrm{d}\tau$ *where* $\boldsymbol{\lambda}(\tau)$ *solves* (1.8)*. The convergence is with respect to the total variation norm on* $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$*.*

Proof *Let* $h := \frac{t_\mathrm{f} - t_\mathrm{s}}{N}$ *be the stepsize of the equidistant grid. Thus, the nodes are* $t_n = t_\mathrm{s} + nh$ *for* $n = 0, \ldots, N$*. We consider firstly the* i*-th component,* $1 \leq i \leq d$*. To ease the notion, we set* $\Lambda := \boldsymbol{\Lambda}_i$*,* $\Lambda^h := \boldsymbol{\Lambda}_i^h$*,* $g := \boldsymbol{g}_i$ *such that the dual norm (see Section 4.3) reads*

$$\left\| \Lambda - \Lambda^h \right\|_{\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]} = \sup_{\|g\|_{C^0[t_\mathrm{s}, t_\mathrm{f}]} = 1} \left| \int_{t_\mathrm{s}}^{t_\mathrm{f}} g(t) \mathrm{d} \left(\Lambda - \Lambda^h \right)(t) \right|.$$

As Λ *is given by* $\Lambda(t) = \int_{t_\mathrm{s}}^t \lambda(\tau)\mathrm{d}\tau$ *and* Λ^h *is a jump function it holds, cf. Section 4.1.2,*

$$\int_{t_\mathrm{s}}^{t_\mathrm{f}} g(t) \mathrm{d} \left(\Lambda - \Lambda^h \right)(t) = \int_{t_\mathrm{s}}^{t_\mathrm{f}} \lambda(t) g(t) \mathrm{d}t - \sum_{n=1}^N h \lambda_n g(t_n).$$

Approximating the integral by the composite trapezoidal rule for equidistant grids yields

$$h\left\{\frac{1}{2}\lambda(t_0)g(t_0)+\sum_{n=1}^{N-1}\lambda(t_n)g(t_n)+\frac{1}{2}\lambda(t_N)g(t_N)\right\}+\mathcal{O}(h^2)-\sum_{n=1}^{N}h\lambda_n g(t_n)$$
$$=h\left\{\frac{1}{2}\lambda(t_0)g(t_0)+\sum_{n=1}^{N}\left[\lambda(t_n)-\lambda_n\right]g(t_n)-\frac{1}{2}\lambda(t_N)g(t_N)\right\}+\mathcal{O}(h^2).$$

We obtain a bound for the $\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d$*-dual norm of* $\mathbf{\Lambda}-\mathbf{\Lambda}^h$ *by taking the absolute value, using the triangle inequality and the fact that* $\|g\|_{C^0[t_\mathrm{s},t_\mathrm{f}]}=1$*, i.e.*

$$\left\|\Lambda-\Lambda^h\right\|_{\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]}\leq h\left\{|\lambda(t_0)|+\sum_{n=1}^{N}|\lambda(t_n)-\lambda_n|+|\lambda(t_N)|\right\}+\mathcal{O}(h^2).$$

With Theorem 7.2 the sum over the main part becomes

$$\sum_{n=m+1}^{N-k}|\lambda(t_n)-\lambda_n|=\sum_{n=m+1}^{N-k}\mathcal{O}(h)=\mathcal{O}(1)$$

such that the norm is bounded by

$$\left\|\Lambda-\Lambda^h\right\|_{\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]}$$
$$\leq h\left\{|\lambda(t_0)|+\sum_{n=1}^{m}|\lambda(t_n)-\lambda_n|+\mathcal{O}(1)+\sum_{n=1}^{k-1}|\lambda(t_{N-n})-\lambda_{N-n}|+|\lambda(t_N)|\right\}+\mathcal{O}(h^2).$$

Since the magnitude of all remaining addends is bounded according to Lemma 7.4 and 7.5 and their number is independent of the step number N*, it is* $\left\|\Lambda-\Lambda^h\right\|_{\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]}=\mathcal{O}(h)$*. Since this holds for all* $i=1,\ldots,d$ *and the value of the dual norm coincides with that of the total variation norm due to the Riesz Representation Theorem (Theorem 4.23), the assertion is shown.* □

Remark 7.8 *By small modifications in the proof of Theorem 7.7, the assertion can be widened to variable stepsizes in the self-starting procedure provided that the variable stepsizes are less than the constant stepsize, cf. Remark 7.6.*

The uniform convergence of $\mathbf{\Lambda}^h$ to $\mathbf{\Lambda}$ in the total variation norm of $\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d$, as demonstrated in the above theorem, implies the pointwise convergence on the entire time interval $[t_\mathrm{s},t_\mathrm{f}]$. In general, for $\mathbf{\Phi}\in\mathrm{NBV}[t_\mathrm{s},t_\mathrm{f}]^d$ and the particular partition $\{t_\mathrm{s},\theta,t_\mathrm{f}\}$ of $[t_\mathrm{s},t_\mathrm{f}]$ with arbitrary time point $\theta\in[t_\mathrm{s},t_\mathrm{f}]$ holds

$$|\mathbf{\Phi}(\theta)|\leq|\mathbf{\Phi}(\theta)|+|\mathbf{\Phi}(t_\mathrm{f})-\mathbf{\Phi}(\theta)|=|\mathbf{\Phi}(\theta)-\mathbf{\Phi}(t_\mathrm{s})|+|\mathbf{\Phi}(t_\mathrm{f})-\mathbf{\Phi}(\theta)|\leq V_{t_\mathrm{s}}^{t_\mathrm{f}}(\mathbf{\Phi})$$

due to Definition 4.3. Thus, Theorem 7.7 implies the pointwise convergence of $\mathbf{\Lambda}^h(t)$ to $\mathbf{\Lambda}(t)$ on the entire time interval $t\in[t_\mathrm{s},t_\mathrm{f}]$ at least with the same linear convergence rate.

Part III

Novel goal-oriented global error estimation for BDF methods

8 Goal-oriented global error estimation

With the novel interpretation for the discrete adjoints of multistep Backward Differentiation Formula (BDF) methods presented in the previous Part II we now derive novel goal-oriented global error estimators for BDF methods. The derivation is based on concepts developed for a posteriori error estimation in Galerkin-type Finite Element (FE) methods.

In this chapter we derive for the first time a posteriori global error estimators that use information computed by adjoint Internal Numerical Differentiation (IND) of multistep BDF method. For a criterion of interest J they estimate the difference

$$J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) \tag{8.1}$$

where $\boldsymbol{y}$ is the unknown exact Initial Value Problem (IVP) solution and $\boldsymbol{y}^h$ the computed BDF approximation. We call this difference in J the *goal-oriented global error* of $\boldsymbol{y}^h$. Throughout this chapter we suppose that J is sufficiently often differentiable.

Generally, we distinguish between *error approximations* and *error estimators*. The former still include unknown exact quantities whereas the latter only depend on computed approximations that are available in practical implementations. Thus, the error approximations are useful for theoretical investigations, for example, of their asymptotic behavior for constant BDF methods and decreasing stepsizes. However, for practical use error estimators are required that perform good for variable BDF methods which choose the stepsizes as large as possible.

The novel goal-oriented error estimators are inspired by the well-established counterpart in Galerkin-type FE methods for Partial Differential Equations (PDEs). After a literature review we start with the derivation of an error representation for (8.1) that includes the unknown exact adjoint solution. Then we derive two approximations for the goal-oriented error representation and propose another goal-oriented error approximation motivated additionally by the classical theory of BDF methods described in Section 2.3. Subsequently, we examine the asymptotic behavior of all three novel goal-oriented error approximations. Finally, we develop efficient goal-oriented error estimators that we have incorporated as well into the variable BDF method `DAESOL-II`. The main computational cost for each estimator is that of a single adjoint IND sweep.

8.1 Literature review of global error estimation in ODEs

In the 1960s and 1970s the field of global error estimation in numerical integration of IVPs in Ordinary Differential Equations (ODEs) was a center of researcher's in-

terest. Zadunaisky [127] proposed to use a continuous approximation obtained by integration of the IVP solution to set up a neighboring IVP with known solution and to solve the neighboring problem with the same integration scheme to obtain an estimate for the global error by subtracting the solutions. Some years before, Henrici [72] proposed to solve another related IVP, which unfortunately involved the unknown local truncation error, cf. Zadunaisky [128]. Stetter [115] used Zadunaisky's technique for iterative improvement of the nominal approximation by addition of the estimated global error. An overview of global error estimation in that period is given by Skeel [113]. Later on, these approaches have been investigated also for BDF methods, see Skeel [114]. However, these approaches suffer from several aspects, amongst others they are costly, not stringent and assume small constant stepsizes.

In the subsequent years, local techniques for error control in numerical integration of ODEs were in the focus of research. Stepsize and order selection strategies based on estimates of local error quantities were developed, see e.g. Hairer et al. [67, 68] or Shampine [109] for a comprehensive presentation of adaptive integration methods. A summary of common local error estimates and stepsize selection techniques for error control can also be found in Shampine [110]. Although, these approaches work satisfactorily and allow an efficient integration promising benefits can be expected from incorporation of the IVP's conditioning by adjoint information.

In the 2000s, a posteriori global error estimation for ODE integration became an active research field again. To estimate the global error in a criterion of interest the solution of the adjoint variational IVP is used as weight for local error quantities, see Moon et al. [96], Cao and Petzold [43], Lang and Verwer [84] and Tran and Berzins [118]. For these global error estimates the adjoint IVP along an approximation of the nominal solution is solved by an additional adaptive integration. This includes the difficult choice of integrator options by the user and the expensive choice of adaptive components by the integrator also for the numerical solution of the adjoint IVP, cf. Section 3.2 and 3.5.

Residual-based a posteriori error estimation in FE methods for PDEs goes back to Babuška and Rheinboldt [13, 12] at the end of the 1970s. The term 'residual' refers to the error given by inserting the approximate solution into the ODE, which in our notion is the defect given in Definition 2.11. In the 1980s, Babuška and Miller [9, 10, 11] introduced the idea to use adjoint information within a posteriori error estimation. Residual-based a posteriori estimates have been investigated also by Estep, Johnson and co-workers, see e.g. Eriksson et al. [55, 56]. The error estimators of FE methods for PDEs have been considered for ODEs as well, see Johnson [77], Estep [58] and Eriksson et al. [55]. These authors summarized the stability of the nominal problem, described by the solution of the adjoint (also called dual) problem according to Section 1.3, in a single (global) stability constant. In the 1990s, Becker and Rannacher [19, 18] refined the latter approach by using distributed stability factors provided by the adjoint solution. This gave rise to the Dual Weighted Residual (DWR) a posteriori error estimates. The approach was

also used in discontinuous Galerkin methods for ODEs, see Böttcher and Rannacher [36]. Moreover, it has been generalized to estimate the error in a given functional, the so-called DWR method for goal-oriented error estimation. More on the wide field of a posteriori error estimation in PDE numerics can be found, e.g. in Verfürth [120] and Babuška and Strouboulis [14]. For details on the DWR method we refer to Becker and Rannacher [20] as well as to the book of Bangerth and Rannacher [15].

We choose the DWR approach as starting point to derive novel global error estimators for BDF-type methods. This approach promises to give efficient and accurate estimators that are also suitable for global error control.

8.2 Goal-oriented error representation

In this section we derive an error representation for the particular Petrov-Galerkin FE discretization developed in Chapter 5 and 6. We carry over some concepts described in Bangerth and Rannacher [15] and Meidner [91] to the particular setting of IVPs in ODEs and BDF methods. Throughout this section we suppose that all systems of equations are solved exactly, notably the BDF equations (2.2b) and hence (6.3b).

Theorem 8.1 *Let $(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l}) \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ be the solution of (5.6) and $\boldsymbol{y}^h \in Y_{\mathcal{P}}[t_\mathrm{s}, t_\mathrm{f}]^d$ the solution of the nominal Petrov-Galerkin FE discretization (6.3c)-(6.3b). Then, the global error in the criterion of interest $J : \mathbb{R}^d \to \mathbb{R}$ takes the following form*

$$\begin{aligned} J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) = -\sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t)) \,\mathrm{d}\left[\boldsymbol{\Lambda} - \boldsymbol{i}_h\boldsymbol{\Lambda}\right](t) \\ - \left[\boldsymbol{l} - \boldsymbol{i}_h\boldsymbol{l}\right]^\intercal \left[\boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] + \mathcal{R}_h \end{aligned} \tag{8.2}$$

for an interpolation operator $\boldsymbol{i}_h : \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d \to Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$. The remainder $\mathcal{R}_h$ is quadratic in the global error function $\boldsymbol{e}(t) := \boldsymbol{y}(t) - \boldsymbol{y}^h(t)$ (cf. Definition 2.10)

$$\begin{aligned} \mathcal{R}_h := &- \int_0^1 \boldsymbol{e}^\intercal(t_\mathrm{f}) J''\left(\boldsymbol{y}^h(t_\mathrm{f}) + s\boldsymbol{e}(t_\mathrm{f})\right) \boldsymbol{e}(t_\mathrm{f}) \cdot s \,\mathrm{d}s \\ &- \int_0^1 \left\{ \sum_{n=0}^{N-1} \int_{I_n} \frac{\mathrm{d}}{\mathrm{d}s} \left\{ \boldsymbol{f}_{\boldsymbol{y}}\left(t, \boldsymbol{y}^h(t) + s\boldsymbol{e}(t)\right) \boldsymbol{e}(t) \right\} \mathrm{d}\boldsymbol{\Lambda}(t) \right\} \cdot s \,\mathrm{d}s. \end{aligned} \tag{8.3}$$

Proof *With integration by parts and the Fundamental Theorem of Calculus the first*

term in $\mathcal{R}_h$ becomes

$$\begin{aligned}
&\int_0^1 \boldsymbol{e}^{\intercal}(t_\mathrm{f}) J''\left(\boldsymbol{y}^h(t_\mathrm{f}) + s\boldsymbol{e}(t_\mathrm{f})\right)\boldsymbol{e}(t_\mathrm{f})\cdot s\,\mathrm{d}s \\
&\qquad = J'\left(\boldsymbol{y}^h(t_\mathrm{f}) + s\boldsymbol{e}(t_\mathrm{f})\right)\boldsymbol{e}(t_\mathrm{f})\cdot s\Big|_0^1 - \int_0^1 J'\left(\boldsymbol{y}^h(t_\mathrm{f}) + s\boldsymbol{e}(t_\mathrm{f})\right)\boldsymbol{e}(t_\mathrm{f})\,\mathrm{d}s \\
&\qquad = J'\left(\boldsymbol{y}(t_\mathrm{f})\right)\boldsymbol{e}(t_\mathrm{f}) - \Big[J\left(\boldsymbol{y}(t_\mathrm{f})\right) - J(\boldsymbol{y}^h(t_\mathrm{f}))\Big].
\end{aligned}$$

In the same way the second term in $\mathcal{R}_h$ becomes

$$\begin{aligned}
&\int_0^1 \left\{\sum_{n=0}^{N-1}\int_{I_n}\frac{\mathrm{d}}{\mathrm{d}s}\left\{\boldsymbol{f}_{\boldsymbol{y}}\left(t,\boldsymbol{y}^h(t)+s\boldsymbol{e}(t)\right)\boldsymbol{e}(t)\right\}\mathrm{d}\boldsymbol{\Lambda}(t)\right\}\cdot s\,\mathrm{d}s \\
&= \left\{\sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}_{\boldsymbol{y}}\left(t,\boldsymbol{y}^h(t)+s\boldsymbol{e}(t)\right)\boldsymbol{e}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t)\right\}\cdot s\Big|_0^1 \\
&\quad - \int_0^1 \left\{\sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}_{\boldsymbol{y}}\left(t,\boldsymbol{y}^h(t)+s\boldsymbol{e}(t)\right)\boldsymbol{e}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t)\right\}\mathrm{d}s \\
&= \sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}_{\boldsymbol{y}}\left(t,\boldsymbol{y}(t)\right)\boldsymbol{e}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) - \sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}\left(t,\boldsymbol{y}(t)\right) - \boldsymbol{f}(t,\boldsymbol{y}^h(t))\,\mathrm{d}\boldsymbol{\Lambda}(t)
\end{aligned}$$

Thus, the remainder becomes

$$\begin{aligned}
\mathcal{R}_h = &- J'\left(\boldsymbol{y}(t_\mathrm{f})\right)\boldsymbol{e}(t_\mathrm{f}) + J\left(\boldsymbol{y}(t_\mathrm{f})\right) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}_{\boldsymbol{y}}\left(t,\boldsymbol{y}(t)\right)\boldsymbol{e}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) \\
&+ \sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}\left(t,\boldsymbol{y}(t)\right) - \boldsymbol{f}(t,\boldsymbol{y}^h(t))\,\mathrm{d}\boldsymbol{\Lambda}(t).
\end{aligned}$$

We focus on the terms containing $\boldsymbol{e}(t)$ and replace $\boldsymbol{e}(t)$ by its expression and start with those terms containing $\boldsymbol{y}$. Due to the extended Riemann-Stieltjes integral (Section 5.3.1) and due to (5.6a) *we obtain*

$$\begin{aligned}
&-J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{y}(t_\mathrm{f}) - \sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}_{\boldsymbol{y}}(t,\boldsymbol{y}(t))\boldsymbol{y}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) \\
&\quad = -J'(\boldsymbol{y}(t_\mathrm{f}))\boldsymbol{y}(t_\mathrm{f}) - \int_{t_\mathrm{s}}^{t_\mathrm{f}}\boldsymbol{f}_{\boldsymbol{y}}(t,\boldsymbol{y}(t))\boldsymbol{y}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) = -\int_{t_\mathrm{s}}^{t_\mathrm{f}}\dot{\boldsymbol{y}}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) - \boldsymbol{l}^{\intercal}\boldsymbol{y}(t_\mathrm{s}) \\
&\quad = -\sum_{n=0}^{N-1}\int_{I_n}\dot{\boldsymbol{y}}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) - \boldsymbol{l}^{\intercal}\boldsymbol{y}(t_\mathrm{s}).
\end{aligned}$$

With Lemma 6.6 the terms containing $\boldsymbol{y}^h$ become

$$J'\left(\boldsymbol{y}(t_\mathrm{f})\right)\boldsymbol{y}^h(t_\mathrm{f}) + \sum_{n=0}^{N-1}\int_{I_n}\boldsymbol{f}_{\boldsymbol{y}}\left(t,\boldsymbol{y}(t)\right)\boldsymbol{y}^h(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) = \sum_{n=0}^{N-1}\int_{I_n}\dot{\boldsymbol{y}}^h(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) + \boldsymbol{l}^{\intercal}\boldsymbol{y}^h(t_\mathrm{s}).$$

With these two expressions, the remainder further transfers to

$$\begin{aligned}
\mathcal{R}_h =& J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) - \boldsymbol{l}^\intercal \boldsymbol{y}(t_\mathrm{s}) + \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}^h(t)\,\mathrm{d}\boldsymbol{\Lambda}(t) \\
&+ \boldsymbol{l}^\intercal \boldsymbol{y}^h(t_\mathrm{s}) + \sum_{n=0}^{N-1} \int_{I_n} \boldsymbol{f}(t, \boldsymbol{y}(t)) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}\boldsymbol{\Lambda}(t) \\
=& J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \boldsymbol{l}^\intercal \left[\boldsymbol{y}(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] + \boldsymbol{l}^\intercal \left[\boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] \\
&- \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}(t) - \boldsymbol{f}(t, \boldsymbol{y}(t))\,\mathrm{d}\boldsymbol{\Lambda}(t) + \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}\boldsymbol{\Lambda}(t) \\
=& J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) + \boldsymbol{l}^\intercal \left[\boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] + \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}\boldsymbol{\Lambda}(t)
\end{aligned}$$

where the last equality holds since $\boldsymbol{y}$ *solves* (5.6b)-(5.6c)*. Hence, we now have*

$$\begin{aligned}
&J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) \\
&\qquad = -\boldsymbol{l}^\intercal \left[\boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] - \sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}\boldsymbol{\Lambda}(t) + \mathcal{R}_h \qquad (8.4)
\end{aligned}$$

and since $\boldsymbol{y}^h$ *solves* (6.3b)-(6.3c) *and* $\boldsymbol{i}_h\boldsymbol{\Lambda} \in Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d$ *the assertion is shown.* □

The error representation (8.2) contains the stability of the continuous IVP via the weak adjoint solution $\boldsymbol{\Lambda}$ and not that of the BDF discretization (2.2) given by the FE weak adjoint $\boldsymbol{\Lambda}^h$. Moreover, the weights $\boldsymbol{\Lambda} - \boldsymbol{i}_h\boldsymbol{\Lambda}$ in (8.2) include the local interpolation error of the exact weak adjoint in $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d$ by its interpolant in $Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d$. For the evaluation of the error representation (8.2), guesses for the unknown exact solutions $\boldsymbol{\Lambda}$ and $\boldsymbol{l}$ are required. We will address this issue in Section 8.3.1.

Most other a posteriori error estimates for ODE approximations like those of Cao and Petzold [43], Lang and Verwer [84] and Tran and Berzins [118] are based on a similar error representation as (8.4) in the classical sense. Such an error representation using the defect and the classical adjoint can also be derived in another way, see e.g. Cao and Petzold [43], and is generally valid for any integration method. All these authors approximate the exact adjoint via the expensive numerical integration of the adjoint IVP (1.8) along the nominal approximation. We might think of using the same representation in conjunction with our weak adjoint approximation $\boldsymbol{\Lambda}^h$: Nevertheless, if we would approximate $\boldsymbol{\Lambda}$ by the Petrov-Galerkin FE weak adjoint $\boldsymbol{\Lambda}^h$ within the error representation (8.4) this would result in a useless estimate being zero since $\boldsymbol{y}^h$ solves (6.3c)-(6.3b). Another a posteriori error estimate related to (8.4) is based on local errors and has been derived by Moon et al. [96].

So far, the error representation (8.2) is not computable since it involves the unknown exact weak adjoint solution $\boldsymbol{\Lambda}$ and the unknown exact solution $\boldsymbol{l}$ of (5.6a).

8.3 Approximation of the error representation

In this section we derive approximations for the error representation (8.2). We do this within three steps. Firstly, the remainder $\mathcal{R}_h$ of (8.2) is neglected since it is quadratic in the global error function $\boldsymbol{e}(t) = \boldsymbol{y}(t) - \boldsymbol{y}^h(t)$. We get

$$J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) \approx E(\boldsymbol{y}^h) := -\sum_{n=0}^{N-1} \int_{I_n} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t)) \,\mathrm{d}\left[\boldsymbol{\Lambda} - \boldsymbol{i}_h\boldsymbol{\Lambda}\right](t) - \left[\boldsymbol{l} - \boldsymbol{i}_h\boldsymbol{l}\right]^\intercal \left[\boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right]. \tag{8.5}$$

Secondly, we will approximate the local interpolation errors $\boldsymbol{\Lambda} - \boldsymbol{i}_h\boldsymbol{\Lambda}$ and $\boldsymbol{l} - \boldsymbol{i}_h\boldsymbol{l}$ using the computed adjoint solutions $\boldsymbol{\Lambda}^h$ and $\boldsymbol{l}^h$ of (6.3a), respectively. This gives us the error approximation based on defect integrals. Thirdly, we will approximate the defect in (8.5). In accordance with Definition 2.11, we use the abbreviation $\boldsymbol{r}_n(t) = \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))$ on I_n. This leads us to the error approximation using local errors. Finally, we will combine these new concepts with the classical ODE theory of BDF methods to propose a third error approximation that uses the local truncation errors.

8.3.1 Approximation of the weights

To approximate the weights $\boldsymbol{\Lambda} - \boldsymbol{i}_h\boldsymbol{\Lambda}$ in (8.5) we use higher order interpolation of the computed weak adjoint solution $\boldsymbol{\Lambda}^h$. There exist also several other approaches to estimate the weights. However, the trade-off between accuracy and effort of higher order interpolation is well-balanced, cf. Becker and Rannacher [20] as well as Bangerth and Rannacher [15]. Since $\boldsymbol{\Lambda}^h$ is piecewise constant, cf. Section 6.1.2, we use piecewise linear interpolation of $\boldsymbol{\Lambda}^h$.

Let $\boldsymbol{\mathcal{I}^{(1)}}$ be the piecewise linear interpolation operator of the form

$$\boldsymbol{\mathcal{I}^{(1)}}\boldsymbol{g}(t) = \boldsymbol{g}(t_n) + \frac{\boldsymbol{g}(t_{n+1}) - \boldsymbol{g}(t_n)}{h_n}(t - t_n), \quad t \in [t_n, t_{n+1}]$$

for $n = 0, \ldots, N-1$. Then, the local interpolation error of the exact weak adjoint $\boldsymbol{\Lambda}$ is approximated by the local interpolation error of the computed FE weak adjoint $\boldsymbol{\Lambda}^h$, i.e.

$$\boldsymbol{\Lambda} - \boldsymbol{i}_h\boldsymbol{\Lambda} \approx \boldsymbol{\mathcal{I}^{(1)}}\boldsymbol{\Lambda}^h - \boldsymbol{\Lambda}^h.$$

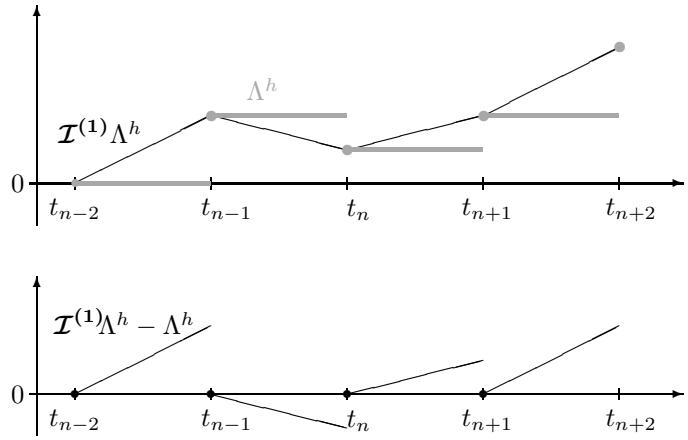


Figure 8.1: FE weak adjoint $\boldsymbol{\Lambda}^h$, its linear interpolation $\boldsymbol{\mathcal{I}}^{(\mathbf{1})}\boldsymbol{\Lambda}^h$ and the resulting weights $\boldsymbol{\mathcal{I}}^{(\mathbf{1})}\boldsymbol{\Lambda}^h - \boldsymbol{\Lambda}^h$ used for the approximations of the error representation (8.2).

On the closed subinterval $\bar{I}_n = \{t_n\} \cup I_n$ of $[t_\mathrm{s}, t_\mathrm{f}]$ the function $\boldsymbol{\mathcal{I}}^{(\mathbf{1})}\boldsymbol{\Lambda}^h - \boldsymbol{\Lambda}^h$ reads

$$\begin{aligned}\left(\boldsymbol{\mathcal{I}}^{(\mathbf{1})}\boldsymbol{\Lambda}^h - \boldsymbol{\Lambda}^h\right)(t) =& \boldsymbol{\Lambda}^h(t_n) + \frac{\boldsymbol{\Lambda}^h(t_{n+1}) - \boldsymbol{\Lambda}^h(t_n)}{h_n}(t - t_n) - \boldsymbol{\Lambda}^h(t) \\ =& \boldsymbol{\Lambda}^h(t_n) + \frac{h_n \boldsymbol{\lambda}_{n+1}}{h_n}(t - t_n) - \boldsymbol{\Lambda}^h(t) \\ =& \boldsymbol{\Lambda}^h(t_n) + \boldsymbol{\lambda}_{n+1}(t - t_n) - \boldsymbol{\Lambda}^h(t) \\ =& \begin{cases} \mathbf{0} & t = t_n \\ \boldsymbol{\lambda}_{n+1}(t - t_n) & t \in (t_n, t_{n+1}) \\ \mathbf{0} & t = t_{n+1} \end{cases} \end{aligned} \tag{8.6}$$

where the definition (6.2) of $\boldsymbol{\Lambda}^h$ is used. Hence, $\boldsymbol{\mathcal{I}}^{(\mathbf{1})}\boldsymbol{\Lambda}^h - \boldsymbol{\Lambda}^h$ as generating function in (8.5) has a jump at the right endpoint t_{n+1} of $\bar{I}_n$ and is continuous from the right as visualized in the lower part of Figure 8.1.

Since the local defect $\boldsymbol{r}_n(t)$ on I_n can be extended continuously to the left endpoint t_n we obtain due to Section 5.3.1 that each integral term in $E(\boldsymbol{y}^h)$ is approximated by

$$\begin{aligned}\int_{I_n} \boldsymbol{r}_n(t)\,\mathrm{d}\left(\boldsymbol{\mathcal{I}}^{(\mathbf{1})}\boldsymbol{\Lambda}^h - \boldsymbol{\Lambda}^h\right)(t) =& \int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}\left(\boldsymbol{\mathcal{I}}^{(\mathbf{1})}\boldsymbol{\Lambda}^h - \boldsymbol{\Lambda}^h\right)(t) \\ =& \boldsymbol{\lambda}_{n+1}^\intercal \int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t - h_n \boldsymbol{\lambda}_{n+1}^\intercal \boldsymbol{r}_n(t_{n+1}) \\ =& \boldsymbol{\lambda}_{n+1}^\intercal \int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t - \boldsymbol{\lambda}_{n+1}^\intercal \boldsymbol{\delta}_{n+1}\end{aligned}$$

where $h_n \boldsymbol{r}_n(t_{n+1}) = h_n \dot{\boldsymbol{y}}^h(t_{n+1}) - h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}^h(t_{n+1})) = \boldsymbol{\delta}_{n+1}$ is the residual of the nonlinear BDF equation (2.2b), cf. Definition 2.9. Although we assumed in Section

8.2 that the BDF equations are solved exactly, we include here the residuals $\boldsymbol{\delta}_{n+1}$ explicitly.

We do not need to approximate the point weight $\boldsymbol{l} - \boldsymbol{i}_h\boldsymbol{l}$ since the residual $\boldsymbol{\delta}_0 = \boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s} = \boldsymbol{y}_0 - \boldsymbol{y}_\mathrm{s} = \mathbf{0}$ always vanishes.

8.3.2 Error approximation with defect integrals

Using the above approximation of the weights the error approximation $E(\boldsymbol{y}^h)$ defined by (8.5) is further approximated by the novel goal-oriented error approximation

$$\bar{E}(\boldsymbol{y}^h) := -\sum_{n=0}^{N-1} \left\{ \boldsymbol{\lambda}_{n+1}^\intercal \int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t - \boldsymbol{\lambda}_{n+1}^\intercal \boldsymbol{\delta}_{n+1} \right\}. \tag{8.7}$$

This error approximation uses only quantities that are available in practical implementations. These are the nominal approximation $\boldsymbol{y}^h$, which is given by the discrete approximations $\{\boldsymbol{y}_n\}_{n=0}^N$ computed by the BDF method, and the discrete adjoints $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ computed by the adjoint IND scheme. This error approximation weights the sum of a nominal local error quantity, the defect integral, and the residual of the nonlinear BDF equation with the discrete stability of the BDF scheme provided by the adjoint IND values. For each integration step, the defect integrals can be evaluated as exactly as desired by numerical quadrature.

8.3.3 Approximation of the defect integrals

In this section we relate the integrals of the defects used in (8.7)

$$\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t = \int_{t_n}^{t_{n+1}} \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}t$$

to the local errors given by Definition 2.4.

Lemma 8.2 *The local error* $\mathbf{LE}(t_{n+1})$ *and the local defect* $\boldsymbol{r}_n(t)$ *of the approximation* $\boldsymbol{y}^h$ *given by* (6.3c)-(6.3b) *are related by*

$$\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t = -\mathbf{LE}(t_{n+1}) + \boldsymbol{\mathcal{R}}_n$$

where the remainder is $\|\boldsymbol{\mathcal{R}}_n\| = \mathcal{O}\left(h_n \cdot \|\mathbf{LE}(t_{n+1})\|\right)$.

Proof *We substract a zero from the defect* $\boldsymbol{r}_n(t)$ *using the ODE of the local IVP given in Definition 2.4 to obtain*

$$\boldsymbol{r}_n(t) = \dot{\boldsymbol{y}}^h(t) - \dot{\boldsymbol{u}}_n(t) + \boldsymbol{f}(t, \boldsymbol{u}_n(t)) - \boldsymbol{f}(t, \boldsymbol{y}^h(t)).$$

Integration over $[t_n, t_{n+1}]$ *yields*

$$\begin{aligned}
\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t &= \int_{t_n}^{t_{n+1}} \dot{\boldsymbol{y}}^h(t) - \dot{\boldsymbol{u}}_n(t)\,\mathrm{d}t + \int_{t_n}^{t_{n+1}} \boldsymbol{f}(t, \boldsymbol{u}_n(t)) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}t \\
&= \boldsymbol{y}^h(t_{n+1}) - \boldsymbol{u}_n(t_{n+1}) - \boldsymbol{y}^h(t_n) + \boldsymbol{u}_n(t_n) + \boldsymbol{\mathcal{R}}_n \\
&= \boldsymbol{y}_{n+1} - \boldsymbol{u}_n(t_{n+1}) - \boldsymbol{y}_n + \boldsymbol{y}_n + \boldsymbol{\mathcal{R}}_n = -\mathbf{LE}(t_{n+1}) + \boldsymbol{\mathcal{R}}_n
\end{aligned}$$

with $\boldsymbol{\mathcal{R}}_n := \int_{t_n}^{t_{n+1}} \boldsymbol{f}(t, \boldsymbol{u}_n(t)) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))\,\mathrm{d}t$. *This remainder is bounded by*

$$\|\boldsymbol{\mathcal{R}}_n\| \leq \int_{t_n}^{t_{n+1}} L \left\| \boldsymbol{u}_n(t) - \boldsymbol{y}^h(t) \right\| \mathrm{d}t \leq h_n L \max_{t\in[t_n,t_{n+1}]} \left\| \boldsymbol{u}_n(t) - \boldsymbol{y}^h(t) \right\|$$

where L *is the Lipschitz constant of* $\boldsymbol{f}(t,\boldsymbol{y})$. *Since* $\boldsymbol{y}^h$ *on* $[t_n, t_{n+1}]$ *is the continuous representation provided by the* n*-th BDF integration step we obtain by Lemma 2.28 that*

$$\max_{t\in[t_n,t_{n+1}]} \left\| \boldsymbol{u}_n(t) - \boldsymbol{y}^h(t) \right\| \leq \left(\frac{\alpha_0^{(n)}}{4} + 1 \right) \|\mathbf{LE}(t_{n+1})\| .$$

Due to the boundedness assumption on $\alpha_0^{(n)}$ *of Section 2.1 the assertion is shown.*□

8.3.4 Error approximation with local errors

With the above result we obtain from (8.7) the following goal-oriented global error approximation

$$\hat{E}(\boldsymbol{y}^h) := \sum_{n=0}^{N-1} \left\{ \boldsymbol{\lambda}_{n+1}^{\mathsf{T}} \mathbf{LE}(t_{n+1}) + \boldsymbol{\lambda}_{n+1}^{\mathsf{T}} \boldsymbol{\delta}_{\boldsymbol{n+1}} \right\} \tag{8.8}$$

for the global error (8.2) in the criterion of interest. This error approximation uses the local error as nominal local error quantity. By this procedure we have introduced the theoretical local solutions $\boldsymbol{u}_n(t_{n+1})$ given in Definition 2.4. Generally, they are not given in practical implementations. Nevertheless, this error approximation gives us theoretical insights within subsequent sections.

8.3.5 Error approximation with local truncation errors

The classical theory of BDF methods, described in Chapter 2, provides an a priori bound on the global error $\mathbf{GE}(t_\mathrm{f}) = \boldsymbol{y}(t_\mathrm{f}) - \boldsymbol{y}^h(t_\mathrm{f})$ by

$$\|\mathbf{GE}(t_N)\| \leq K \left\{ \max_{0\leq n\leq m} \|\mathbf{GE}(t_n)\| + \frac{1}{h} \left(\max_{0\leq n\leq N} \|\boldsymbol{\delta}_n\| + \max_{0\leq n\leq N} \|\mathbf{LTE}(t_n)\| \right) \right\},$$

see (2.13). In this formula the constant K describes, in a worst-case scenario, the stability of the IVP. Hence, to obtain a more rigorous approximation we might replace K by the local stability factors $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ computed by the adjoint IND scheme. In the goal-oriented error approximation (8.7) the defect integrals are used as nominal local error quantities. This indicates that the nominal error quantities are at the $\boldsymbol{y}$-level instead of the $\dot{\boldsymbol{y}}$-level. Accordingly, we might integrate $1/h\,\|\mathbf{LTE}(t_n)\|$ on $[t_n, t_{n+1}]$, i.e. multiply by $h = t_{n+1} - t_n$, to end up at the $\boldsymbol{y}$-level. Altogether we propose a third goal-oriented global error approximation by

$$\tilde{E}(\boldsymbol{y}^h) := \sum_{n=0}^{N-1} \left\{ \boldsymbol{\lambda}_{n+1}^{\mathsf{T}} \mathbf{LTE}(t_{n+1}) + \boldsymbol{\lambda}_{n+1}^{\mathsf{T}} \boldsymbol{\delta}_{\boldsymbol{n+1}} \right\} \tag{8.9}$$

which is motivated by the classical theory of BDF methods. Unfortunately, the local truncation errors $\mathbf{LTE}(t_{n+1})$ given by (2.15) use time-derivatives of the unknown nominal solution $\boldsymbol{y}(t)$ which is not available in practical implementations. Nevertheless, they can be estimated as described in Section 2.4.1.

8.4 Asymptotic behavior of the error approximations

In this section we address the asymptotic correctness of the three goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$ given by (8.7), $\hat{E}(\boldsymbol{y}^h)$ given by (8.8) and $\tilde{E}(\boldsymbol{y}^h)$ given by (8.9) to measure the true goal-oriented global error $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))$. We consider again a constant BDF method with appropriate self-starter and sufficiently accurate solved nonlinear BDF equations such that the global error satisfies $\|\boldsymbol{e}(t_\mathrm{f})\| = \left\|\boldsymbol{y}(t_\mathrm{f}) - \boldsymbol{y}^h(t_\mathrm{f})\right\| = \|\boldsymbol{y}(t_\mathrm{f}) - \boldsymbol{y}_N\| = \mathcal{O}(h^k)$, cf. Theorem 2.21 or Shampine and Zhang [112].

8.4.1 Notation

We start with some definitions that are important to measure the quality of an error approximation. However, they can be directly transfered to error estimators.

Definition 8.3 *Let $\boldsymbol{y}^h(t_\mathrm{f})$ be a numerical approximation of the exact solution $\boldsymbol{y}(t_\mathrm{f})$ of IVP* (1.1) *that converges at order k and let the criterion of interest J be continuously differentiable. Then, an a posteriori error approximation $\check{E}(\boldsymbol{y}^h)$ is called* asympotically correct *for $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))$ if*

$$J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \check{E}(\boldsymbol{y}^h) = \mathcal{O}(h^{k+1}) \tag{8.10}$$

holds.

Thus, asymptotical correctness means that the error in the error approximation is of higher order in h than the true error.

Definition 8.4 *The* signed effectivity index $I_\mathrm{eff}^\mathrm{s}$ *of an a posteriori error approximation $\check{E}(\boldsymbol{y}^h)$ for $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))$ is given by*

$$I_\mathrm{eff}^\mathrm{s} =: \frac{\check{E}(\boldsymbol{y}^h)}{J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))}. \tag{8.11}$$

In FE methods for PDEs the notion of asymptotic correctness of a posteriori error approximations is slightly different. In the PDE community the quality of an a posteriori error approximation $\check{E}(\boldsymbol{y}^h)$ is usually measured by the ratio of its absolute value and the absolute value of the true error, cf. Babuška and Strouboulis [14] and Verfürth [120]. This so-called *effectivity index* of $\check{E}(\boldsymbol{y}^h)$ given by

$$I_\mathrm{eff} =: \frac{\left|\check{E}(\boldsymbol{y}^h)\right|}{\left|J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))\right|}.$$

should be near one to have an accurate error approximation. In this context, an approximation is called *asymptotically correct* if the effectivity index tends to one if the stepsizes converge to zero. If $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) \neq 0$ for all discretization grids, we may divide (8.10) by $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))$ to obtain

$$1 - I_\mathrm{eff}^\mathrm{s} = \mathcal{O}(h) \tag{8.12}$$

due to $\left|J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))\right| = \mathcal{O}(h^k)$. Thus, both concepts for the definition of asymptotic correctness coincide.

We base the theoretical investigations on the asymptotic correctness of Definition 8.3 and use the signed effectivity index for the numerical investigations. In practice, we are already satisfied if the absolute value of the effectivity index and its reciprocal remain reasonable bounded for all grids, e.g. if $|I_\mathrm{eff}^\mathrm{s}| \in [0.5, 2]$ holds. The sign of $I_\mathrm{eff}^\mathrm{s}$ describes whether the error approximation is able to give information about sign and magnitude of the true error or only about magnitude.

8.4.2 Theoretical investigations of the asymptotic behavior

In this section we investigate theoretically the asymptotic behavior of the goal-oriented global error approximations $\bar{E}(\boldsymbol{y}^h)$ given by (8.7) and $\hat{E}(\boldsymbol{y}^h)$ given by (8.8).

Theorem 8.5 *Let $(\boldsymbol{y}, \boldsymbol{\Lambda}, \boldsymbol{l}) \in C^1[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ be the solution of* (5.6) *and $\left(\boldsymbol{y}^h, \boldsymbol{\Lambda}^h, \boldsymbol{l}^h\right) \in Y_\mathcal{P}[t_\mathrm{s}, t_\mathrm{f}]^d \times Z_\mathrm{H}[t_\mathrm{s}, t_\mathrm{f}]^d \times \mathbb{R}^d$ the solution of the Petrov-Galerkin FE discretization* (6.3) *where the nominal BDF method* (6.3c)-(6.3b) *converges at order k. Then, the error approximation $\bar{E}(\boldsymbol{y}^h)$ given by* (8.7) *for $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))$ is asymptotically correct, i.e. it holds*

$$\left|J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \bar{E}(\boldsymbol{y}^h)\right| = \mathcal{O}(h^{k+1}). \tag{8.13}$$

Proof *With the approximation defined in* (8.5) *we obtain by the triangle inequality*

$$\left|J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \bar{E}(\boldsymbol{y}^h)\right| \leq \left|J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - E(\boldsymbol{y}^h)\right| + \left|E(\boldsymbol{y}^h) - \bar{E}(\boldsymbol{y}^h)\right|.$$

According to Theorem 8.1 the approximation $E(\boldsymbol{y}^h)$ differs from $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))$ by the remainder term $\mathcal{R}_h$ which is quadratic in $\boldsymbol{e}$ and hence $|\mathcal{R}_h| = \mathcal{O}(h^{2k})$ due to Theorem 2.21. For the second summand we substract (8.7) *from* (8.5) *to obtain*

$$\begin{aligned}\left|E(\boldsymbol{y}^h) - \bar{E}(\boldsymbol{y}^h)\right| = \Bigg|&- \sum_{n=0}^{N-1} \int_{I_n} \boldsymbol{r}_n(t) \,\mathrm{d}\left[\boldsymbol{\Lambda} - \boldsymbol{i}_h\boldsymbol{\Lambda}\right](t) - \left[\boldsymbol{l} - \boldsymbol{i}_h\boldsymbol{l}\right]^\intercal \left[\boldsymbol{y}^h(t_\mathrm{s}) - \boldsymbol{y}_\mathrm{s}\right] \\ &+ \sum_{n=0}^{N-1} \left\{\boldsymbol{\lambda}_{n+1}^\intercal \int_{I_n} \boldsymbol{r}_n(t) \,\mathrm{d}t - \boldsymbol{\lambda}_{n+1}^\intercal \boldsymbol{\delta}_{n+1}\right\}\Bigg|.\end{aligned}$$

With the initial condition (6.3c), *the weak adjoint $\boldsymbol{\Lambda}$ fulfilling* (5.7), *the interpolation $\boldsymbol{i}_h\boldsymbol{\Lambda}(t) = \sum_{n=0}^{N-1} h_n \boldsymbol{\lambda}(t_{n+1}) H_{n+1}(t)$ and the extended Riemann-Stieltjes integral of*

Section 5.3.1 the above difference becomes

$$\left|\sum_{n=0}^{N-1}\left\{\int_{I_n}\boldsymbol{\lambda}^{\mathsf{T}}(t)\boldsymbol{r}_n(t)\,\mathrm{d}t - h_n\boldsymbol{\lambda}^{\mathsf{T}}(t_{n+1})\boldsymbol{r}_n(t_{n+1}) - \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\int_{I_n}\boldsymbol{r}_n(t)\,\mathrm{d}t + \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\boldsymbol{\delta}_{n+1}\right\}\right|. \tag{8.14}$$

Now we make use of the triangle inequality to find an upper bound. We start with the non-integral terms of (8.14) *and use the fact that* $h_n\boldsymbol{r}_n(t_{n+1}) = \boldsymbol{\delta}_{n+1}$

$$\left|\sum_{n=0}^{N-1}[\boldsymbol{\lambda}(t_{n+1}) - \boldsymbol{\lambda}_{n+1}]^{\mathsf{T}}\boldsymbol{\delta}_{n+1}\right| \leq \sum_{n=0}^{N-1}\|\boldsymbol{\lambda}(t_{n+1}) - \boldsymbol{\lambda}_{n+1}\|\cdot\|\boldsymbol{\delta}_{n+1}\|\,.$$

For $n = m,\dots,N-k$ *it is* $\|\boldsymbol{\lambda}(t_{n+1}) - \boldsymbol{\lambda}_{n+1}\| = \mathcal{O}(h)$ *due to Theorem 7.2 whereas all others are* $\|\boldsymbol{\lambda}(t_{n+1}) - \boldsymbol{\lambda}_{n+1}\| = \mathcal{O}(1)$ *due to Lemma 7.4 and 7.5. And hence, by the assumption that* $\|\boldsymbol{\delta}_{n+1}\| = \mathcal{O}(h^{k+1})$, $\mathcal{O}(N)$ *summands are* $\mathcal{O}(h^{k+2})$ *and only* $\mathcal{O}(1)$ *are* $\mathcal{O}(h^{k+1})$. *Since* $h = (t_{\mathrm{f}} - t_{\mathrm{s}})/N$ *the sum becomes* $\mathcal{O}(h^{k+1})$. *Secondly, using the Taylor series* $\boldsymbol{\lambda}(t) = \boldsymbol{\lambda}(t_{n+1}) + \mathcal{O}(h)$ *on* $[t_n, t_{n+1}]$ *the sum of the integral terms in* (8.14) *becomes*

$$\begin{aligned}
&\left|\sum_{n=0}^{N-1}\left\{\boldsymbol{\lambda}^{\mathsf{T}}(t_{n+1})\int_{I_n}\boldsymbol{r}_n(t)\mathrm{d}t + \mathcal{O}(h)\int_{I_n}\boldsymbol{r}_n(t)\mathrm{d}t - \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\int_{I_n}\boldsymbol{r}_n(t)\mathrm{d}t\right\}\right| \\
&= \left|\sum_{n=0}^{N-1}\left\{\boldsymbol{\lambda}^{\mathsf{T}}(t_{n+1}) - \boldsymbol{\lambda}_{n+1}^{\mathsf{T}} + \mathcal{O}(h)\right\}\int_{I_n}\boldsymbol{r}_n(t)\mathrm{d}t\right| \\
&\leq \sum_{n=0}^{N-1}\|\boldsymbol{\lambda}(t_{n+1}) - \boldsymbol{\lambda}_{n+1} + \mathcal{O}(h)\|\cdot\left\|\int_{I_n}\boldsymbol{r}_n(t)\mathrm{d}t\right\|.
\end{aligned}$$

The norm of the defect integral is $\left\|\int_{I_n}\boldsymbol{r}_n(t)\mathrm{d}t\right\| = \mathcal{O}(h^{k+1})$ *due to Lemma 8.2 and 2.8 as well as the consistency order* k *of the constant BDF method, cf. Section 2.3.1. Hence, again* $\mathcal{O}(N)$ *summands are* $\mathcal{O}(h^{k+2})$ *and only* $\mathcal{O}(1)$ *are* $\mathcal{O}(h^{k+1})$ *such that the assertion is shown.* □

The convergence of the defect integrals at order $k+1$ to zero is confirmed numerically in Section 10.2.1, more precisely in the top row of Figure 10.4. The result of Theorem 8.5 can be further used to describe the asymptotic behavior of the error approximation $\hat{E}(\boldsymbol{y}^h)$ obtained by approximating the defect integral by the local error as described in Section 8.3.3.

Corollary 8.6 *Let the assumptions of Theorem 8.5 hold. Then, for the error approximation* $\hat{E}(\boldsymbol{y}^h)$ *given by* (8.8) *also holds*

$$\left|J(\boldsymbol{y}(t_{\mathrm{f}})) - J(\boldsymbol{y}^h(t_{\mathrm{f}})) - \hat{E}(\boldsymbol{y}^h)\right| = \mathcal{O}(h^{k+1}).$$

Proof *The first summand of*

$$\left|J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \hat{E}(\boldsymbol{y}^h)\right| \leq \left|J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) - \bar{E}(\boldsymbol{y}^h)\right| + \left|\bar{E}(\boldsymbol{y}^h) - \hat{E}(\boldsymbol{y}^h)\right|$$

behaves like $\mathcal{O}(h^{k+1})$ *according to Theorem 8.5. Substracting* (8.8) *from* (8.7) *yields*

$$\begin{aligned}\left|\bar{E}(\boldsymbol{y}^h) - \hat{E}(\boldsymbol{y}^h)\right| &= \left|-\sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^\intercal \left\{\int_{t_n}^{t_{n+1}} \boldsymbol{r}^h(t)\,\mathrm{d}t + \mathbf{LE}(t_{n+1})\right\}\right| \\ &= \left|\sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^\intercal \mathcal{R}_n\right| \leq \sum_{n=0}^{N-1} \left|\boldsymbol{\lambda}_{n+1}^\intercal \mathcal{R}_n\right| \leq \sum_{n=0}^{N-1} \|\boldsymbol{\lambda}_{n+1}\|\,\|\mathcal{R}_n\|\end{aligned}$$

due to Lemma 8.2. For all $n = 0, \ldots, N-1$ *the value* $\|\boldsymbol{\lambda}_n\|$ *remains bounded since* $\|\boldsymbol{\lambda}_n - \boldsymbol{\lambda}(t_n)\| = \mathcal{O}(1)$ *due to Theorem 7.2, Lemma 7.4 and 7.5 and* $\boldsymbol{\lambda}(t)$ *is bounded on* $[t_\mathrm{s}, t_\mathrm{f}]$ *due to its continuity, cf. Section 5.1. Due to Lemma 8.2 it is* $\|\mathcal{R}_n\| = \mathcal{O}(h\,\|\mathbf{LE}(t_{n+1})\|) = \mathcal{O}(h^{k+2})$ *where the last equality is due to Lemma 2.8 and the BDF consistency order* k. □

8.4.3 First numerical experiments

We investigate the asymptotic behavior of the three goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ numerically using constant BDF methods and IVP examples that provide all analytic solutions, i.e. exact global and local solutions as well as exact time-derivatives of the solutions. In Section 10.2.2 we present the numerical results in all details.

For a constant BDF method of order $k = 1$ the signed effectivity indices defined in (8.11) of the error approximations $\bar{E}(\boldsymbol{y}^h)$ given by (8.7) and $\hat{E}(\boldsymbol{y}^h)$ given by (8.8) converge linearly to the desired value one for decreasing stepsizes. Thus, the results of Theorem 8.5 and Corollary 8.6 are confirmed numerically for the one-step BDF method.

Surprisingly, for multistep BDF methods the linear convergence of the effectivity indices to one is not affirmed numerically. The signed effectivity indices of $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ for a constant BDF method of order $k = 2$ with two first-order steps of size $h/2$ as self-starter show a problem-dependent offset from the desired value one. These numerical observations raise the question which assumptions used by Theorem 8.5 are not fulfilled in practice. We illuminate this in the next section.

However, the signed effectivity indices of the error approximation $\tilde{E}(\boldsymbol{y}^h)$ given by (8.9) show a linear convergence to one for both BDF methods, the one-step method with $k = 1$ and the multistep method with order $k = 2$.

In summary, from the numerical point of view the goal-oriented error approximation $\tilde{E}(\boldsymbol{y}^h)$ seems to be the approximation of choice. However, from the theoretical derivation in function spaces the error approximations $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ seem to be the appropriate ones.

8.4.4 Further investigations concerning the asymptotic behavior

In this section we focus on the relation of the error approximations $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ compared to the error approximation $\tilde{E}(\boldsymbol{y}^h)$. We have seen that their asymptotic behaviors are different and in particular for multistep methods other than described by Theorem 8.5 and Corollary 8.6.

With the linear Dahlquist equation and the linear criterion of interest $J(y(t_\mathrm{f})) = y(t_\mathrm{f})$ of Section 10.2.2 we can eliminate the term $\mathcal{R}_h$ defined in (8.3) as reason for the offset since $\mathcal{R}_h$ is zero in this case while the offset is observed, cf. first row of Figure 10.7(a). Furthermore, the convergence of the defect integrals $\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\mathrm{d}t$ at order $k+1$ can also be confirmed numerically, see upper left corner of Figure 10.4 of Section 10.2.1. Thus, we have to search for other reasons.

If we suppose that the Localizing Assumption of Definition 2.7 holds in every integration step, i.e. that $\boldsymbol{y}_{n+1-i} = \boldsymbol{y}(t_{n+1-i})$ holds for all past values used to compute $\boldsymbol{y}_{n+1}$, then the local truncation error $\mathbf{LTE}(t_{n+1})$ in $\tilde{E}(\boldsymbol{y}^h)$ can be written as

$$\mathbf{LTE}(t_{n+1}) = \alpha_0^{(n)}\mathbf{LE}(t_{n+1}) - \mathcal{O}(h_n)\mathbf{LE}(t_{n+1}) \tag{8.15}$$

which follows directly from Lemma 2.8. Thus, under the artificial Localizing Assumption the relation between $\tilde{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ is the following.

Lemma 8.7 *For a constant BDF method of order k with m variable starting steps and supposing that the Localizing Assumption holds in every integration step the goal-oriented error approximations $\tilde{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ are related by*

$$\begin{aligned}\tilde{E}(\boldsymbol{y}^h) = \alpha_0^{(l)}\hat{E}(\boldsymbol{y}^h) + \sum_{n=0}^{l-1}(\alpha_0^{(n)} - \alpha_0^{(l)})\boldsymbol{\lambda}_{n+1}^\intercal\mathbf{LE}(t_{n+1}) + (1-\alpha_0^{(l)})\sum_{n=0}^{N-1}\boldsymbol{\lambda}_{n+1}^\intercal\boldsymbol{\delta_{n+1}} \\ + \mathcal{O}(h^{k+1}).\end{aligned} \tag{8.16}$$

Proof *See Section A.2.2.* □

Moving all terms except $\mathcal{O}(h^{k+1})$ to one side of the equality sign in (8.16) and dividing by the true error $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f})) \neq 0$ we would expect linear convergence in this relation value to zero also numerically. However, this is not the case. Thus, we drop the artificial Localizing Assumption and express the local truncation error in terms of the local error as follows.

Lemma 8.8 *Without the Localizing Assumption the local truncation error $\mathbf{LTE}(t_{n+1})$ can be written as*

$$\begin{aligned}\mathbf{LTE}(t_{n+1}) =& \alpha_0^{(n)}\mathbf{LE}(t_{n+1}) - \mathcal{O}(h_n)\mathbf{LE}(t_{n+1}) \\ &+ \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] + \sum_{i=1}^{k_n}\alpha_i^{(n)}\mathbf{GE}(t_{n+1-i})\end{aligned} \tag{8.17}$$

where $\boldsymbol{u}_n(t_{n+1})$ is the local exact solution of $\dot{\boldsymbol{u}}_n(t) = \boldsymbol{f}(t, \boldsymbol{u}_n(t))$, $\boldsymbol{u}_n(t_n) = \boldsymbol{y}_n$, cf. Definition 2.4.

Proof *See Section A.2.2.* □

With this expression for the local truncation errors, the goal-oriented error approximations $\tilde{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ are related as follows.

Lemma 8.9 *For a constant BDF method of order k with m variable starting steps the goal-oriented error approximations $\tilde{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ are related by*

$$\begin{aligned}\tilde{E}(\boldsymbol{y}^h) = \alpha_0^{(l)}\hat{E}(\boldsymbol{y}^h) + \sum_{n=0}^{l-1}(\alpha_0^{(n)} - \alpha_0^{(l)})\boldsymbol{\lambda}_{n+1}^{\intercal}\mathbf{LE}(t_{n+1}) + (1-\alpha_0^{(l)})\sum_{n=0}^{N-1}\boldsymbol{\lambda}_{n+1}^{\intercal}\boldsymbol{\delta_{n+1}} \\ + \sum_{n=0}^{N-1}\boldsymbol{\lambda}_{n+1}^{\intercal}\left[\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] + \sum_{i=1}^{k_n}\alpha_i^{(n)}\mathbf{GE}(t_{n+1-i})\right] \\ + \mathcal{O}(h^{k+1}).\end{aligned} \tag{8.18}$$

Proof *The proof follows that of Lemma 8.7 given in Section A.2.2 but uses the expression* (8.17) *for the local truncation errors instead of* (8.15)*.* □

Numerical experiments indicate that the relation between $\tilde{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ is actually described by Lemma 8.9. Using this relation we are able to derive the following *implicit correction term*

$$\begin{aligned}\Delta\bar{E}(\boldsymbol{y}^h) :=& (\alpha_0^{(l)} - 1)\bar{E}(\boldsymbol{y}^h) + \sum_{n=0}^{l-1}(\alpha_0^{(n)} - \alpha_0^{(l)})\boldsymbol{\lambda}_{n+1}^{\intercal}\mathbf{LE}(t_{n+1}) + (1-\alpha_0^{(l)})\sum_{n=0}^{N-1}\boldsymbol{\lambda}_{n+1}^{\intercal}\boldsymbol{\delta}_{n+1} \\ &+ \sum_{n=0}^{N-1}\boldsymbol{\lambda}_{n+1}^{\intercal}\left[\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] + \sum_{i=1}^{k_n}\alpha_i^{(n)}\mathbf{GE}(t_{n+1-i})\right]\end{aligned} \tag{8.19}$$

for the goal-oriented error approximation $\bar{E}(\boldsymbol{y}^h)$ and similarly for $\hat{E}(\boldsymbol{y}^h)$. In fact, the signed effectivity indices of the corrected error approximations $\bar{E}(\boldsymbol{y}^h) + \Delta\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h) + \Delta\hat{E}(\boldsymbol{y}^h)$ approach the desired value one like depicted, see third row of Figure 10.7 in Section 10.2.2.

Furthermore, for a constant BDF method of order $k = 1$ we are able to show that the correction term coverges quadratically to zero which means that the leading terms of $\tilde{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ coincide as observed numerically.

Lemma 8.10 *For a constant BDF method of order $k = 1$ the correction terms converge quadratically for decreasing stepsize h, i.e.*

$$\Delta\bar{E}(\boldsymbol{y}^h) = \Delta\hat{E}(\boldsymbol{y}^h) = \mathcal{O}(h^2).$$

Proof *For a BDF method of order $k = 1$ it is $\alpha_0^{(n)} = \alpha_0^{(l)} = 1$ such that the correction terms* (8.19) *of $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ become*

$$\Delta\bar{E}(\boldsymbol{y}^h) = \Delta\hat{E}(\boldsymbol{y}^h) = \sum_{n=0}^{N-1}\boldsymbol{\lambda}_{n+1}^{\intercal}\left[\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] + \alpha_1^{(n)}\mathbf{GE}(t_n)\right].$$

Each term in brackets becomes

$$\begin{aligned}\boldsymbol{\mathcal{J}}^{(n)}_{\mathrm{BDF}}(\boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] + \alpha_1^{(n)}\mathbf{GE}(t_n)\\ &= \boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1}) - \boldsymbol{y}(t_n) + \boldsymbol{y}_n - h\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})]\\ &= h\dot{\boldsymbol{y}}(t_{n+1}) - \frac{h^2}{2}\ddot{\boldsymbol{y}}(t_{n+1}) - h\dot{\boldsymbol{u}}_n(t_{n+1}) + \frac{h^2}{2}\ddot{\boldsymbol{u}}_n(t_{n+1}) + \mathcal{O}(h^3)\\ &\quad - h\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] \qquad (8.20)\end{aligned}$$

using the Taylor series expansions of $\boldsymbol{y}(t_n)$ *and* $\boldsymbol{y}_n = \boldsymbol{u}_n(t_n)$ *around* t_{n+1}*. A central point of the proof is that due to Theorem 1.7 and the power series of the exponential function it holds*

$$\begin{aligned}\|\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})\| &\leq \|\mathbf{GE}(t_n)\| \exp(Lh)\\ &= \|\mathbf{GE}(t_n)\| \sum_{i=0}^{\infty} \frac{(Lh)^i}{i!} = \mathcal{O}(\|\mathbf{GE}(t_n)\|) = \mathcal{O}(h). \qquad (8.21)\end{aligned}$$

The first order derivatives of (8.20) *sum up in the following way using the Taylor series expansions of* $\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}(t_{n+1}))$ *around* $\boldsymbol{u}_n(t_{n+1})$ *and of* $\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{u}_n(t_{n+1}))$ *around* $\boldsymbol{y}_{n+1}$*, respectively,*

$$\begin{aligned}&h\left\{\dot{\boldsymbol{y}}(t_{n+1}) - \dot{\boldsymbol{u}}_n(t_{n+1}) - \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})]\right\}\\ &= h\left\{\boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1})) - \boldsymbol{f}(t_{n+1}, \boldsymbol{u}_n(t_{n+1})) - \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})]\right\}\\ &= h\left\{\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{u}_n(t_{n+1})) - \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})\right\}[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})]\\ &\quad + \mathcal{O}(h\,\|\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})\|^2)\\ &= h\mathcal{O}(\|\mathbf{LE}(t_{n+1})\|)[\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] + \mathcal{O}(h\,\|\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})\|^2)\end{aligned}$$

with $\mathbf{LE}(t_{n+1}) = \boldsymbol{u}_n(t_{n+1}) - \boldsymbol{y}_{n+1}$*. Using 8.21 the above sum behaves like* $\mathcal{O}(h^3)$*. Furthermore, the sum of the second order derivatives in* (8.20) *becomes*

$$\begin{aligned}&\frac{h^2}{2}\left\{\ddot{\boldsymbol{u}}_n(t_{n+1}) - \ddot{\boldsymbol{y}}(t_{n+1})\right\}\\ &= \frac{h^2}{2}\left\{\boldsymbol{f}_t(t_{n+1}, \boldsymbol{u}_n(t_{n+1})) + \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{u}_n(t_{n+1}))\boldsymbol{f}(t_{n+1}, \boldsymbol{u}_n(t_{n+1}))\right.\\ &\quad \left. - \boldsymbol{f}_t(t_{n+1}, \boldsymbol{y}(t_{n+1})) - \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}(t_{n+1}))\boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1}))\right\}\\ &= \frac{h^2}{2}\mathcal{O}(\|\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})\|)\end{aligned}$$

using the Taylor series expansions around $\boldsymbol{u}_n(t_{n+1})$ *of* $\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}(t_{n+1}))\boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1}))$ *and* $\boldsymbol{f}_t(t_{n+1}, \boldsymbol{y}(t_{n+1}))$ *and hence also behaves like* $\mathcal{O}(h^3)$*. With the boundedness of all* $\boldsymbol{\lambda}_{n+1}$*, as shown in the proof of Corollary 8.6, and* $\sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}\mathcal{O}(h^3) = \mathcal{O}(h^2)$ *since* $h = (t_{\mathrm{f}} - t_{\mathrm{s}})/N$ *the proof is finished.* □

In the proof of Lemma 8.10 it is shown that for BDF methods of order one those terms of the correction term caused by the removal of the Localizing Assumption

are negligible. Hence, for the goal-oriented error approximation of one-step BDF methods the Localizing Assumption does not cause any negative effect. In fact, the whole correction term is negligible due to Lemma 8.10. However, for BDF methods of higher order than one, i.e. for true multistep methods, the goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ which are based on function space arguments have to be corrected by (8.19) to give effectivity indices that approach one for decreasing stepsizes, see Section 10.2.2. The interpretation of these additional correction terms in the function space derivation is still an open issue.

8.5 Goal-oriented global error estimators

However, for practical usage in variable order variable stepsize BDF-type methods like the realization `DAESOL-II` the asymptotic behavior of an error approximation is not as important as its efficient and accurate evaluation for possibly large stepsizes h_n. For an efficient realization of the goal-oriented global error approximations $\bar{E}(\boldsymbol{y}^h)$ given by (8.7), $\hat{E}(\boldsymbol{y}^h)$ given by (8.8) and $\tilde{E}(\boldsymbol{y}^h)$ given by (8.9) we have to regard further aspects. These include for all three error approximations the efficient computation of the discrete adjoints $\boldsymbol{\lambda}_{n+1}$ and for each one either the quadrature of the defects, the estimation of the exact local errors $\mathbf{LE}(t_{n+1})$ or of the exact local truncation errors $\mathbf{LTE}(t_{n+1})$. Furthermore, the residuals $\boldsymbol{\delta}_{n+1}$ of the nonlinear BDF equations are needed for all three goal-oriented error approximations.

8.5.1 Discrete adjoints

The adjoint IND values $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ used in all three error approximations of Section 8.3 are given by the adjoint IND scheme (3.2) of the nominal BDF method. Nevertheless, it is more efficient to compute the adjoint values $\{\bar{\boldsymbol{y}}_n\}_{n=0}^N$ as solution of the adjoint IND scheme (3.4) corresponding to the domain space formulation (3.3) of a BDF step. In fact, the computation of $\{\bar{\boldsymbol{y}}_n\}_{n=0}^N$ saves N builts and decompositions of the BDF Jacobians and transposed solutions. Furthermore, in `DAESOL-II` this adjoint IND is realized in the more efficient iterative version. Since $\boldsymbol{\lambda}_{n+1}$ and $\bar{\boldsymbol{y}}_{n+1}$ are related by the inverse $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})^{-\intercal}$ of the transposed Jacobians according to Lemma 3.3 we approximate $\boldsymbol{\lambda}_{n+1}$ by

$$\widehat{\boldsymbol{\lambda}}_{n+1} := \frac{1}{\alpha_0^{(n)}}\, \bar{\boldsymbol{y}}_{n+1}$$

for $n = 0, \ldots, N-1$. These approximations are asymptotically correct due to the following reasons: The inverse $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})^{-\intercal}$ can be expressed by its Neumann series (see Theorem A.4) assuming that $h_n/\alpha_0^{(n)}\, \|\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1})\| < 1$ holds and hence approximated up to first order by the first summand of the series. The saving of computational costs by using $\{\widehat{\boldsymbol{\lambda}}_n\}_{n=1}^N$ instead of $\{\boldsymbol{\lambda}_n\}_{n=1}^N$ is significant, particularly if the Jacobian $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y})$ of the ODE right hand side is expensive to evaluate.

8.5.2 Nominal local error quantities

Now we consider the nominal local error quantities used in $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$, respectively. We start with the local truncation errors $\mathbf{LTE}(t_{n+1})$ required for $\tilde{E}(\boldsymbol{y}^h)$ given by (8.9). They are also the fundament for the classical strategies of variable BDF methods to control the integration accuracy locally by stepsize and order adaption in each integration step. For variable BDF methods the local truncation error is given by (2.15) and includes the derivative $\boldsymbol{y}^{(k_n+1)}(t_{n+1})$. As described in Section 2.4.1 practical implementations use finite differences of the past approximations $\boldsymbol{y}_{n-k_n}, \ldots, \boldsymbol{y}_n$ and $\boldsymbol{y}_{n+1}$ to estimate the aforementioned derivative and hence to give the estimated local truncation error $\widehat{\mathbf{LTE}}(t_{n+1})$. The asymptotic correctness of this estimator was shown by Gear [62]. Hence, $\tilde{E}(\boldsymbol{y}^h)$ is estimated efficiently by the following goal-oriented global error estimator

$$\tilde{\eta} := \sum_{n=0}^{N-1} \widehat{\boldsymbol{\lambda}}_{n+1}^{\intercal} \widehat{\mathbf{LTE}}(t_{n+1}) + \eta_{\boldsymbol{\delta}} \tag{8.22}$$

where the weighted sum of the residuals of the nonlinear BDF equations is summarized by

$$\eta_{\boldsymbol{\delta}} := \sum_{n=0}^{N-1} \widehat{\boldsymbol{\lambda}}_{n+1}^{\intercal} \boldsymbol{\delta}_{\boldsymbol{n+1}}. \tag{8.23}$$

This residual term appears in all goal-oriented error approximations of Section 8.3 and hence in all goal-oriented estimators developed here. It will be treated in more detail in Section 8.5.3 below.

The goal-oriented error approximation $\hat{E}(\boldsymbol{y}^h)$ defined by (8.8) makes use of the local errors $\mathbf{LE}(t_{n+1})$. They might be estimated by solving the local IVP of Definition 2.4 using a higher order integration method. Nevertheless, this is computationally very expensive and hence not recommendable for practical use. Therefore, we rather suppose that the Localizing Assumption (Definition 2.7) holds such that using Lemma 2.8 and the above approximation of $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})^{-1}$ the local error can be approximated by

$$\mathbf{LE}(t_{n+1}) \doteq \frac{1}{\alpha_0^{(n)}} \mathbf{LTE}(t_{n+1}). \tag{8.24}$$

Thus, we use the estimator $\widehat{\mathbf{LTE}}(t_{n+1})$ of the local truncation error to obtain an estimator $\widehat{\mathbf{LE}}(t_{n+1}) := \widehat{\mathbf{LTE}}(t_{n+1})/\alpha_0^{(n)}$ for the local error. This estimation is not very accurate since by using the Localizing Assumption the term $\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1}) + \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})^{-1} \sum_{i=1}^{k_n} \alpha_i^{(n)} \mathbf{GE}(t_{n+1-i})$ is completely neglected in (8.24). Nevertheless, with the estimator $\widehat{\mathbf{LE}}(t_{n+1})$ we have, at least, a reasonable and efficiently computed value at hand. Hence, we estimate $\hat{E}(\boldsymbol{y}^h)$ by the following goal-oriented

global error estimator

$$\hat{\eta} := \sum_{n=0}^{N-1} \widehat{\boldsymbol{\lambda}}_{n+1}^{\intercal} \widehat{\mathbf{LE}}(t_{n+1}) + \eta_{\boldsymbol{\delta}}. \tag{8.25}$$

Finally, to realize the goal-oriented error approximation $\bar{E}(\boldsymbol{y}^h)$ defined by (8.7) we use numerical quadrature of the exact defects $\boldsymbol{r}_n(t) = \dot{\boldsymbol{y}}^h(t) - \boldsymbol{f}(t, \boldsymbol{y}^h(t))$ on $[t_n, t_{n+1}]$ for $n = 0, \ldots, N-1$. The computational cost depends on the quadrature formula and the required tolerance. The resulting goal-oriented global error estimator to estimate $\bar{E}(\boldsymbol{y}^h)$ reads

$$\bar{\eta} := -\sum_{n=0}^{N-1} \widehat{\boldsymbol{\lambda}}_{n+1}^{\intercal} \int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t + \eta_{\boldsymbol{\delta}}. \tag{8.26}$$

Note that the derived error estimators $\tilde{\eta}$ and $\bar{\eta}$ are asymptotically correct to estimate the error approximations $\tilde{E}(\boldsymbol{y}^h)$ and $\bar{E}(\boldsymbol{y}^h)$, respectively. The estimator $\hat{\eta}$ is not asymptotically correct for $\hat{E}(\boldsymbol{y}^h)$ since the estimation $\widehat{\mathbf{LE}}(t_{n+1})$ is not asymptotically correct for the exact local error $\mathbf{LE}(t_{n+1})$.

8.5.3 Residuals of the nonlinear BDF equations

In all three goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ of Section 8.3, and hence in the estimators $\bar{\eta}$, $\hat{\eta}$ and $\tilde{\eta}$ of Section 8.5.2, the residuals $\{\boldsymbol{\delta}_n\}_{n=1}^N$ of the nonlinear BDF equations (2.2b) build a weighted sum. The residuals $\boldsymbol{\delta}_{n+1}$ themselves can be computed exactly by inserting the nominal approximation $\boldsymbol{y}_{n+1}$, i.e. the last iterate of the iterative procedure used to solve (2.2b), into the corresponding BDF equation. Nevertheless, if the ODE right hand side $\boldsymbol{f}(t, \boldsymbol{y})$ is expensive to evaluate, also the evaluation of the residuals is computationally expensive. Generally, in the implementation of implicit integration methods the accuracy achieved by the time discretization has to be the dominant one, particularly the implicit, nonlinear equations have to be solved to a higher accuracy. If so, the residuals are comparably small in contrast to the local error quantities utilized in $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$, respectively, and might be neglected therefore.

The stepsize and order selection rule together with the monitor strategy of Section 2.4 guarantee that the residuals $\boldsymbol{\delta}_{n+1}$ are heuristically smaller than the local truncation errors according to the following lemma and the appropriate choice of the Newton tolerance. We use the notion of Section 2.4.3.

Lemma 8.11 *Assume* $h_n/\alpha_0^{(n)} \left\| \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1}^*) \right\| < 1$*, such that the nonlinear BDF equation* (2.2b) *has a unique solution* $\boldsymbol{y}_{n+1}^*$*. Suppose that* $\boldsymbol{M}_n$ *and* $\boldsymbol{y}_{n+1}^{(0)} = \boldsymbol{y}_{n+1}^{\mathrm{P}}$ *satisfy the requirements of the Local Contraction Theorem (Theorem 2.29) and* $\left\| \Delta \boldsymbol{y}_{n+1}^{(s_n-1)} \right\| <$ `NTol` *is the termination criterion of the Newton-type method. If the method terminates with*

1. $s_n = 3$ *and* $0.25 \leq \delta_0 < 0.3$

2. $s_n = 2$ *and* $\delta_0 < 0.25$
3. $s_n = 1$ *and* $\left\| \boldsymbol{\mathcal{F}}_{\mathrm{BDF}}^{(n)} \left(\boldsymbol{y}_{n+1}^{(s_n)} \right) \right\| \leq \alpha_0^{(n)} \mathtt{NTol}$

then the approximation $\boldsymbol{y}_{n+1}^{(s_n)}$ *to* $\boldsymbol{y}_{n+1}^*$ *leads to a residual bounded by*

$$\left\| \boldsymbol{\delta}_{n+1}^{(s_n)} \right\| = \left\| \boldsymbol{\mathcal{F}}_{\mathrm{BDF}}^{(n)} \left(\boldsymbol{y}_{n+1}^{(s_n)} \right) \right\| \leq \alpha_0^{(n)} \mathtt{NTol}.$$

Proof *The residual* $\boldsymbol{\delta}_{n+1}^{(s_n)}$ *of the last iterate* $\boldsymbol{y}_{n+1}^{(s_n)}$ *is given by*

$$\begin{aligned} \boldsymbol{\delta}_{n+1}^{(s_n)} = \boldsymbol{\mathcal{F}}_{\mathrm{BDF}}^{(n)} \left(\boldsymbol{y}_{n+1}^{(s_n)} \right) &= \boldsymbol{\mathcal{F}}_{\mathrm{BDF}}^{(n)} \left(\boldsymbol{y}_{n+1}^* + \left(\boldsymbol{y}_{n+1}^{(s_n)} - \boldsymbol{y}_{n+1}^* \right) \right) \\ &= \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}^*) \left(\boldsymbol{y}_{n+1}^{(s_n)} - \boldsymbol{y}_{n+1}^* \right) + \mathcal{O}\left(\left\| \boldsymbol{y}_{n+1}^{(s_n)} - \boldsymbol{y}_{n+1}^* \right\|^2 \right) \end{aligned}$$

and hence it is bounded by

$$\left\| \boldsymbol{\delta}_{n+1}^{(s_n)} \right\| \leq \left\| \alpha_0^{(n)} \boldsymbol{I} - h_n \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1}^*) \right\| \cdot \left\| \boldsymbol{y}_{n+1}^{(s_n)} - \boldsymbol{y}_{n+1}^* \right\| < 2\alpha_0^{(n)} \left\| \boldsymbol{y}_{n+1}^{(s_n)} - \boldsymbol{y}_{n+1}^* \right\|.$$

For $s_n \in \{2, 3\}$*, the a priori estimate of the Local Contraction Theorem and the decrease of the sequence* (δ_i) *yield that*

$$\left\| \boldsymbol{y}_{n+1}^{(s_n)} - \boldsymbol{y}_{n+1}^* \right\| \leq \frac{\delta_{s_n - 1}}{1 - \delta_{s_n - 1}} \left\| \Delta \boldsymbol{y}_{n+1}^{(s_n - 1)} \right\| \leq \frac{\delta_0}{1 - \delta_0} \left\| \Delta \boldsymbol{y}_{n+1}^{(s_n - 1)} \right\| < \frac{1}{2} \mathtt{NTol}$$

since $\delta_0 < 1/3$ *by assumption. Hence, the assertion is shown.* □

At least for constant stepsizes we have $1 \leq \alpha_0^{(n)} \leq 2.45 < 5/2$, see Section 2.4.1, and together with the choice $\mathtt{NTol} = 0.08 \cdot \mathtt{RelTol}$ of Section 2.4.3 we obtain

$$\left\| \boldsymbol{\delta}_{n+1}^{(s_n)} \right\| < 5/2 \cdot 0.08 \cdot \mathtt{RelTol} = 0.2 \cdot \mathtt{RelTol}.$$

On the other hand, the estimated local truncation error is bounded by RelTol due to Section 2.4.1. In this way, the residuals $\boldsymbol{\delta}_{n+1}$ are negligible compared to $\widehat{\mathbf{LTE}}(t_{n+1})$ in the goal-oriented error estimator $\tilde{\eta}$ given by (8.22). This also holds for $\hat{\eta}$ given by (8.25) since at least for constant stepsize we have $||\widehat{\mathbf{LE}}(t_{n+1})|| \leq ||\widehat{\mathbf{LTE}}(t_{n+1})||$ according to (8.24) and $\alpha_0^{(n)} \geq 1$.

For fully variable BDF methods we will examine numerically the impact of the residual term $\eta_{\boldsymbol{\delta}}$ on the accuracy of the goal-oriented estimators in Section 10.3.2.

8.5.4 Computational complexity and vector-valued criterions of interest

The novel goal-oriented global error estimators $\bar{\eta}$, $\hat{\eta}$ and $\tilde{\eta}$ are available at different computational costs. All of them need the approximations $\widehat{\boldsymbol{\lambda}}_{n+1} = \bar{\boldsymbol{y}}_{n+1}/\alpha_0^{(n)}$ of the discrete adjoints $\boldsymbol{\lambda}_{n+1}$. Hence, one adjoint IND sweep of the domain space formulation of the BDF method is necessary to compute $\{\bar{\boldsymbol{y}}_n\}_{n=0}^N$. For this, the more efficient iterative version should be preferred to the direct version, cf. Section

3.4.2 or Albersmeyer and Bock [5]. Using $\widehat{\boldsymbol{\lambda}}_{n+1}$ instead of $\boldsymbol{\lambda}_{n+1}$ for $n = 0, \ldots, N$ requires only N scalar-vector multiplications with vector length d instead of N builts and decompositions of the $d \times d$ BDF Jacobian and transposed solutions.

For $\tilde{\eta}$ the local truncation errors have to be estimated. With the BDF method based on Newton interpolation polynomials the estimation of $\mathbf{LTE}(t_{n+1})$ is computationally very efficient, cf. Bleser [25], and also used for the local stepsize and order selection described in Section 2.4. Apart from the cost for the estimation of the local truncation errors, the goal-oriented error estimator $\hat{\eta}$ needs N additional scalar-vector multiplications. For $\bar{\eta}$, the computational effort depends directly on the quadrature formula used to obtain the defect integrals and the cost for the evaluation of the ODE right hand side $\boldsymbol{f}(t, \boldsymbol{y})$.

If the error estimator also includes the residual term $\eta_{\boldsymbol{\delta}}$, N residuals have to be evaluated at a cost depending on the evaluation cost for the ODE right hand side $\boldsymbol{f}(t, \boldsymbol{y})$.

Apart from the memory requirements of the adjoint IND sweep (cf. Albersmeyer and Bock [5]), additional memory is required during the nominal integration to store the vector-valued estimates of the nominal local error quantities and possibly the vector-valued defects in each integration step.

If the global error in a vector-valued criterion of interest $\boldsymbol{J} = [J_1, \cdots, J_M]^\intercal$ is required, this can be computed as well with the goal-oriented error estimators derived above. To this end, the error in each component J_i is estimated like in the case of a scalar criterion of interest. Altogether, this gives a vector-valued estimator $\check{\boldsymbol{\eta}}$ for the vector-valued error $\boldsymbol{J}(\boldsymbol{y}(t_\mathrm{f})) - \boldsymbol{J}(\boldsymbol{y}^h(t_\mathrm{f}))$. It is available at the cost of M adjoint IND sweeps, each in direction $J_i'(\boldsymbol{y}^h(t_\mathrm{f}))$ for $i = 1, \ldots, M$. The nominal error quantities have to be computed only once.

9 Application of the novel estimators for goal-oriented error control

So far, most adaptive integrators determine the integration accurcacy by means of a given relative tolerance. But, the relative tolerance applies to the local accuracy only and does not guarantee any global error bound for the approximation of the Initial Value Problem (IVP) solution. Furthermore, the appropriate choice of the relative tolerance in terms of accuracy and efficiency is still a challenge. If the IVP at hand is asymptotically stable, local inaccuracies are damped out and a loose tolerance already yields a good approximation. On the other hand, if the IVP to be solved is highly unstable, already small errors amplify in a disastrous manner and the approximation might become useless at all.

With the novel goal-oriented global error estimators for Backward Differentiation Formula (BDF) methods derived in Chapter 8 we hold a suitable tool to resolve such ambivalent situations. We now examine how the estimated information can be used to control the nominal integration such that the goal-oriented global error of the nominal approximation is influenced appropriately. 'Appropriately' in this context has two tendencies. It may mean to reduce the error of the nominal integration or to loose it since the nominal integration needs not to be that accurate and computational effort can be saved.

In this chapter we will not treat the choice of the required tolerance `GTol` for the goal-oriented global error that should be met by the nominal approximation. We rather assume to be given a global tolerance `GTol`. In the simulation context the choice of `GTol` is due to the user and his/her particular aims. In the optimal control context the choice of `GTol` should be done by the optimization procedure itself based on its progress towards the optimum and its convergence behavior, see Bock [31]. Nevertheless, this topic itself is a field of research and hence is beyond the scope of this thesis.

However, in this chapter we will use our novel goal-oriented error estimators within algorithmic frameworks to drive the approximation in such a way that its goal-oriented global error is below a given global tolerance `GTol`. In the whole chapter we assume that the nonlinear BDF equations are solved until the residuals are negligible compared to the nominal local error quantities, cf. Section 8.5.3, and hence that the residual term $\eta_{\boldsymbol{\delta}}$ in the goal-oriented estimators of Section 8.5.2 is negligible. Exemplarily, we focus here on the goal-oriented estimator $\tilde{\eta}$ defined by (8.22) using estimated local truncation errors. Nevertheless, the estimators $\bar{\eta}$ and $\hat{\eta}$ could be used as well. We start in Section 9.1 with an approach that uses the error estimator $\tilde{\eta}$ to

adjust the relative tolerances for subsequent integrations with the standard stepsize and order selection, cf. Section 2.4. In Section 9.2, we continue with a strategy that adapts the discretization scheme directly using each addend of the sum in (8.22) separately. The termination criterion of both strategies is satisfied if the estimator η fulfilles

$$|\eta| \leq c \cdot \texttt{GTol} \tag{9.1}$$

where c is a positive constant that accounts for the over- or underestimation tendency of the estimator η.

9.1 Goal-oriented local tolerance adaption

Here we use the information obtained, for example, by the novel goal-oriented global error estimator $\tilde{\eta}$ to influence the local integration tolerance in the so-called goal-oriented local tolerance adaption. Based on the error estimate $\tilde{\eta}$ the relative tolerance `RelTol` is reduced such that the local truncation errors in the subsequent integration are decreased in the hope for a corresponding reduction in the goal-oriented global error. After the first nominal integration with $\texttt{RelTol}^0 = \texttt{RelTol}$ for a user given relative tolerance `RelTol`, the goal-oriented error is estimated by $\tilde{\eta}^0$. As long as $\left|\tilde{\eta}^j\right| \leq c \cdot \texttt{GTol}$ is not fulfilled, the nominal integration is repeated with

$$\texttt{RelTol}^{j+1} = \texttt{RelTol}^j \cdot \min\left\{c_{\text{red}}, \frac{c \cdot \texttt{GTol}}{|\tilde{\eta}^j|}\right\} \tag{9.2}$$

where $c_{\text{red}} < 1$ is a positive factor assuring reduction. In this approach all integrations are performed with the stepsize and order selection as well as the monitor strategy described in Section 2.4.

The choice $c \cdot \texttt{GTol}/\left|\tilde{\eta}^j\right|$ in (9.2) is based on the following assumption: If the integration with $\texttt{RelTol}^j$ as upper bound on the local truncation errors yields an approximation with estimated global error $\left|\tilde{\eta}^j\right| > c \cdot \texttt{GTol}$, then a subsequent integration with a relative tolerance reduced by the factor $c \cdot \texttt{GTol}/\left|\tilde{\eta}^j\right| < 1$ is supposed to reduce the global error by the same factor such that the global error of the new approximation approaches the termination criterion (9.1).

The algorithmic procedure is summarized in Algorithm 1. It extends an idea described by Lang and Verwer [84]. The number of iterations (integrations) in Algorithm 1 is limited by J. If after J integrations the tolerance `GTol` is still not met, the algorithm terminates with a failure.

However, the goal-oriented local tolerance adaption has its limitations. It relies on the assumption that a reduction of the relative tolerance results in a reduction of the goal-oriented global error. But, the choice of the adaptive components does not guarantee that the estimated local truncation errors meet the upper bound of $||\widehat{\mathbf{LTE}}(t_{n+1})|| \leq \texttt{RelTol}$ in (2.17), cf. Section 2.4.1, and hence it is not guaranteed that the estimated local truncation errors for the more restrictive tolerance are always smaller than that of the less restrictive one. Furthermore, the conditioning

Algorithm 1: Goal-oriented local tolerance adaption

Input : Desired `GTol`, loose `RelTol`.

Output: Approximate solution $\boldsymbol{y}_N$ with estimated goal-oriented error $\tilde{\eta}$.

```
RelTol^0 = RelTol, j = 0;
Integration with RelTol^0;
Estimation of goal-oriented error η̃^0;
while |η̃^j| > c · GTol and RelTol^j > 10^{-14} and j < J do
    RelTol^{j+1} = RelTol^j · min{c_red, c · GTol / |η̃^j|};
    Integration with RelTol^{j+1};
    Estimation of goal-oriented error η̃^{j+1};
    j = j + 1;
N = N^j, η̃ = η̃^j;
```

of the IVP to be solved has a crucial impact on the propagation of local inaccuracies, see Section 1.3. Local inaccuracies in different areas of the time interval may be propagated differently. Hence, the relative tolerance might be unnecessarily restrictive for the local truncation errors on parts of asymptotic stability whereas on parts of IVP instability they should be smaller than the relative tolerance. These limitations of the goal-oriented local tolerance adaption call for a more flexible strategy that accounts not only for nominal local error quantities but also for the local stability.

However, in situations where an IVP has to be solved many times and the desired global accuracies are known, the goal-oriented error estimate $\tilde{\eta}$ can be used analogously to (9.2) to obtain an *educated guess* for a suitable relative tolerance. This also includes an increase of the relative tolerance if the previous one has been unnecessarily restrictive.

9.2 Goal-oriented scheme adaption

In this section we examine how to include not only the estimated nominal local error quantities but also the estimated stability of the IVP, which determines the propagation of local errors, into a goal-oriented global error control mechanism. We do this again exemplarily with the help of the goal-oriented estimator $\tilde{\eta}$ based on estimated local truncation errors. The estimator $\tilde{\eta}$ given by (8.22) with negligible residual term $\eta_{\boldsymbol{\delta}}$ is the sum of so-called *local error indicators* $\tilde{\eta}_n$

$$\tilde{\eta} = \sum_{n=0}^{N-1} \tilde{\eta}_n \text{ with } \tilde{\eta}_n = \widehat{\boldsymbol{\lambda}}_{n+1}^{\intercal} \widehat{\mathbf{LTE}}(t_{n+1}). \tag{9.3}$$

Based on these indicators we can adapt the integration scheme of the BDF method for a subsequent integration with prescribed scheme. This goal-oriented scheme adaption replaces the standard stepsize and order selection mechanism described in Section 2.4 completely and is summarized in Algorithm 2.

Algorithm 2: Goal-oriented scheme adaption

Input : Desired `GTol`, loose `RelTol`.

Output: Approximate solution $\boldsymbol{y}_N$ with estimated goal-oriented error $\tilde{\eta}$.

```
1 j = 0;
2 Integration with RelTol;
3 Estimation of local error indicators {η̃_n^0} and goal-oriented error η̃^0;
4 while |η̃^j| > c · GTol and j < J do
5 |   Indicator-based scheme adaption (Algorithm 3);
6 |   Integration with prescribed integration scheme;
7 |   Estimation of local error indicators {η̃_n^{j+1}} and goal-oriented error η̃^{j+1} ;
8 |   j = j + 1;
9 N = N^j, η̃ = η̃^j;
```

The total number of iterations in Algorithm 2 is again limited by J. The indicator-based scheme adaption for BDF methods and the integration with prescribed schemes, i.e. line 2 and 1 of Algorithm 1, are addressed in Section 9.2.1 and 9.2.2, respectively.

Moon et al. [95] used a similar approach for one-step methods to reduce the global error by dividing those integration steps with the largest error contributions into uniform substeps. Due to the simultaneous stepsize adaptation for all integration steps Bangerth and Rannacher [15] suggested to call this approach *implicit* stepsize control. This separate adjustment step has its origin in adaptive Finite Element (FE) methods for Partial Differential Equations (PDEs) where the space discretization is chosen adaptively. Similar adaptation procedures are applied to IVPs in Ordinary Differential Equations (ODEs) by Böttcher and Rannacher [36], Eriksson et al. [55] and Logg [87] using continuous and discontinuous Galerkin methods.

9.2.1 Indicator-based scheme adaption for BDF methods

We now focus on the indicator-based adaption of the integration scheme of BDF-type methods using the local error indicators of the novel goal-oriented error estimator $\tilde{\eta}$. There exist several implicit adaption strategies. For example, Logg [87] and Moon et al. [95] used error balancing over all integration steps. Beside this, Becker and Rannacher [20] also used strategies that refine a particular number of integration steps. One is to reduce the local error indicators of a fixed percentage of integration steps. Another one is to reduce those indicators that yield a fixed percentage of the estimated error. For the moment we focus on the reduction of the local error indicators of $\mathrm{p} \cdot 100$ percent of the integration steps and develop an approach to achieve this in the case of BDF methods.

Originally, BDF-type methods as realized in `DAESOL-II` are based on the relative tolerance `RelTol` provided by the user and strategies to control the local accuracy, cf. Section 2.4. However, the goal-oriented scheme adaption of Algorithm 2 totally replaces the standard selection rules for stepsize and order. Only a first integration

with a loose relative tolerance `RelTol` and the standard stepsize and order selection mechanism is performed. This yields the first integration scheme $\{h_n\}_{n=0}^{N-1}$, $\{k_n\}_{n=0}^{N-1}$ and $\{\texttt{NTol}_n\}_{n=0}^{N-1}$ with $\texttt{NTol}_n = 0.08 \cdot \texttt{RelTol}$ (cf. Section 2.4.3). After estimating the goal-oriented global error, the local error indicators of $\mathrm{p} \cdot 100$ percent of the integration steps are reduced.

Generally, the estimated local truncation error $\widehat{\mathbf{LTE}}(t_{n+1})$ of the n-th integration step, included in $\tilde{\eta}_n$ according to (9.3), can be influenced directly by the choice of stepsize h_n and order k_n. As a first attempt we try to reduce the error indicator of the n-th integration step by bisecting the subinterval I_n and performing two integration steps with stepsize $h_n/2$ and order k_n. To maintain the assumption that the residuals of the nonlinear BDF equations and hence the residual term $\eta_{\boldsymbol{\delta}}$ are negligible we have to adjust also the Newton tolerance $\texttt{NTol}_n$, cf. Section 8.5.3.

In Section 2.4.3 the Newton tolerance was given via the relative tolerance as $\texttt{NTol}_n = 0.08 \cdot \texttt{RelTol}$. For the particular Newton-type method fulfilling the assumptions of Lemma 8.11 this Newton tolerance yields a residual $\boldsymbol{\delta}_{n+1}$ with norm bounded by $\alpha_0^{(n)} \cdot 0.08 \cdot \texttt{RelTol}$. On the other hand, for the estimated local truncation error holds $||\widehat{\mathbf{LTE}}(t_{n+1})|| \leq \texttt{RelTol}$ which is guaranteed by the standard stepsize and order selection. If we assume constant stepsize and take (2.12), i.e. $\mathbf{LTE}(t_{n+1}) \doteq C_{k_n+1} h^{k_n+1} \boldsymbol{y}^{(k_n+1)}(t_{n+1})$, into account, then a bisection of the stepsize reduces the local truncation error by a factor of $1/2^{(k_n+1)}$. Hence, we have to reduce the Newton tolerance by this factor as well, such that the residual $\boldsymbol{\delta}_{n+1}$ of the nonlinear BDF equation remain negligible compared to $\widehat{\mathbf{LTE}}(t_{n+1})$.

The whole indicator-based scheme adaption is summarized in Algorithm 3.

Algorithm 3: Indicator-based scheme adaption

Input : $\mathrm{p}, \{\eta_n\}, \{h_n\}, \{k_n\}, \{\texttt{NTol}_n\}$ for $n = 0, \ldots, N-1$.
Output: Adjusted $\{h'_n\}, \{k'_n\}, \{\texttt{NTol}'_n\}$ for $n = 0, \ldots, N'-1$.
1 Sort $|\eta_{n_1}| \geq \cdots \geq |\eta_{n_N}|$;
2 **for** $i = 1 : \mathrm{p}N$ **do**
3 Bisect subinterval I_{n_i};
4 Use order k_{n_i} for both steps;
5 Use Newton tolerance $\texttt{NTol}_{n_i}/2^{(k_{n_i}+1)}$ for both steps;

Note that the termination tolerances for the numerical solution of the nonlinear BDF equations by Newton-type methods are not fixed anymore over all integration steps, but rather they are adjusted according to the local conditions, cf. line 3 in Algorithm 3.

9.2.2 Integration with prescribed integration scheme

After the indicator-based adaption of the integration scheme the next step of the goal-oriented scheme adaption is to integrate the IVP with prescribed adaptive components, cf. line 2 in Algorithm 2. For each integration step, stepsize h_n and order

k_n are already given and the nonlinear BDF equation has to be solved with given tolerance $\texttt{NTol}_{n+1}$. To retain the efficiency of the overall integration the hierarchical update procedure for the iteration matrices in the Newton-type method of Section 2.4.3 is not changed. Particularly, the monitor strategy is still used and limits the computational effort by keeping the iteration matrix fixed as long as convergence is observed. Furthermore, this integration is much cheaper compared to the standard integration (as e.g. used in line 2 of Algorithm 1) since the selection of stepsize and order is omitted.

The goal-oriented scheme adaption of Algorithm 2 and 3 is already good as the numerical results in Section 11.2 will demonstrate. Nevertheless, the indicator-based scheme adaption of Algorithm 3 provides potential for improvements and should be seen as a starting point. For example, it is an open question how to adapt the orders of the integration scheme and the described adaption of the Newton tolerance seems to be quite restrictive. Additionally, Algorithm 3 only allows a refinement but no coarsing of the integration grid.

In addition to the above goal-oriented global error control strategies a third strategy is of great interest: The stability described by the adjoints should be incorporated into the standard selection of stepsize, order and Newton tolerance and hence utilized in a subsequent integration with time stepping goal-oriented adaption. Cao and Petzold [43] investigated such an approach for their a posteriori error estimator using a costly integration of the adjoint IVP, a constant order BDF method and a stepsize adaption formula based on equidistant grids. They indicated a gain in accuracy and efficiency compared to their standard stepsize adaption for unstable and stiff IVPs, respectively. An interesting issue for future research would be the development of such a global error control strategy based on our novel goal-oriented error estimators, which are superior compared to that of Cao and Petzold due to Section 10.3, and our standard stepsize and order selection, which uses the local truncation error formula for variable grids, cf. Section 2.4.1 or Bleser [25] and Eich [52, 53].

Part IV

Numerical results

10 Numerical validation

In this chapter we give numerical evidence to the theoretical results derived in Part II and Chapter 8. To this end, we use Initial Value Problems (IVPs) with known analytic solutions and investigate the results computed with a constant Backward Differentiation Formula (BDF) method and the variable BDF-type method `DAESOL-II`. In the first part of this chapter we focus on the Finite Element (FE) approximations of the weak adjoints computed by adjoint Internal Numerical Differentiation (IND). We will confirm numerically the convergence results of Chapter 7 and demonstrate numerically the smooth behavior of the FE weak adjoints also for variable BDF-type methods. Secondly, we examine the goal-oriented global error approximations of Section 8.3 and their asymptotic behavior examined in Section 8.4. Finally, we investigate the accuracy of the novel goal-oriented global error estimators derived in Chapter 8, and in particular in Section 8.5, and compare our estimators to a corresponding one proposed by Cao and Petzold [43].

10.1 Weak adjoint solutions

We illustrate the theoretical results of Part II with the help of a nonlinear test case with known analytic nominal and adjoint solutions. The Catenary, see e.g. Hairer et al. [67], is given by a second order Ordinary Differential Equation (ODE)

$$\ddot{y}(t) = p\sqrt{1+\dot{y}(t)^2}, \quad p > 0.$$

We reformulate the ODE as system of first order equations

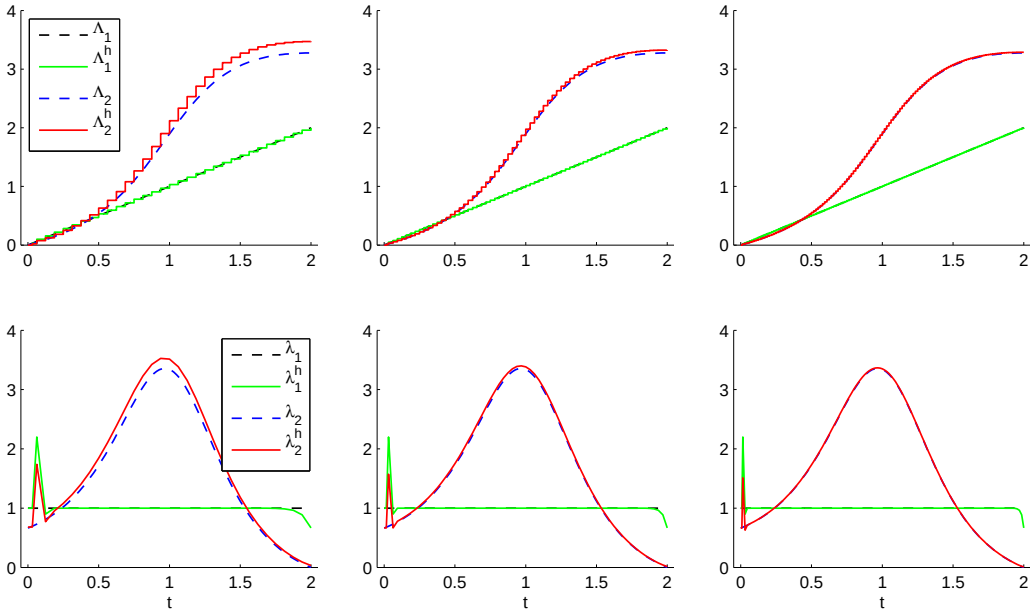
$$\begin{aligned} \dot{y}_1(t) &= y_2(t) \\ \dot{y}_2(t) &= p\sqrt{1+y_2(t)^2} \end{aligned}$$

and solve it on the interval $[t_\mathrm{s}, t_\mathrm{f}] = [0, 2]$ for the parameter choice $p = 3$ and the initial conditions $\boldsymbol{y}(0) = \boldsymbol{y}_\mathrm{s} = [1/3\cosh(-3),\ \sinh(-3)]^\intercal$. As criterion of interest we choose $J(\boldsymbol{y}(2)) = y_1(2)$. The analytic nominal solution and the analytic classical adjoint solution are

$$\boldsymbol{y}(t) = \begin{pmatrix} B + \frac{1}{p}\cosh(pt+A) \\ \sinh(pt+A) \end{pmatrix}, \quad \boldsymbol{\lambda}(t) = \begin{pmatrix} 1 \\ -\frac{\sinh(pt+A)-\sinh(pt_\mathrm{f}+A)}{p\cosh(pt+A)} \end{pmatrix} \tag{10.1}$$

and the analytic weak adjoint solution in the space $\mathrm{NBV}[t_\mathrm{s}, t_\mathrm{f}]^2$ is

$$\boldsymbol{\Lambda}(t) = \begin{pmatrix} t \\ -\frac{1}{p^2}\ln(\cosh(pt+A)) + \frac{2}{p^2}\sinh(pt_\mathrm{f}+A)\arctan\left(\mathrm{e}^{pt+A}\right) \end{pmatrix} \tag{10.2}$$



(a) $h = 2^{-4}$ (b) $h = 2^{-5}$ (c) $h = 2^{-6}$

Figure 10.1: Results for the constant BDF method. Comparison of the FE weak adjoints $\mathbf{\Lambda}^h$ to the analytic weak adjoints $\mathbf{\Lambda}$ (top) as well as the discrete adjoint IND values $\boldsymbol{\lambda}^h$ compared to analytic classical adjoints $\boldsymbol{\lambda}$ (bottom) for three different stepsizes.

where $A = -p$ and $B = 0$. The results presented in this section are also contained in Beigel et al. [21].

10.1.1 Constant BDF method

We use a constant BDF method with order 2 and stepsize h implemented in Matlab®. The self-starting procedure consists of two first-order BDF steps with stepsize $h/2$. The nonlinear BDF equations are solved up to a given small accuracy. Furthermore, the direct adjoint IND scheme (3.2) is realized.

The lower row of Figure 10.1 compares the discrete adjoint IND values $\boldsymbol{\lambda}^h = \{\boldsymbol{\lambda}_n\}_{n=0}^N$ for the three different stepsizes $h = 2^{-4}$, 2^{-5} and 2^{-6} to the analytic solution $\boldsymbol{\lambda}$ given by (10.1) of the adjoint IVP along the analytic nominal solution. The peaks of the discrete adjoints at the interval ends are due to the inconsistency of the adjoint initialization and termination steps of the discrete adjoint IND scheme with the adjoint IVP, cf. Section 3.6 and 7.1. Nevertheless, the discrete adjoints converge on the open interval $(0, 2)$ towards the analytic adjoint solution as we have demonstrated in Theorem 7.2. In the upper row of Figure 10.1 the FE approximation $\mathbf{\Lambda}^h$ is compared to the analytic weak adjoint $\mathbf{\Lambda}$ given by (10.2). It converges on the whole time interval as we have shown in Theorem 7.7.

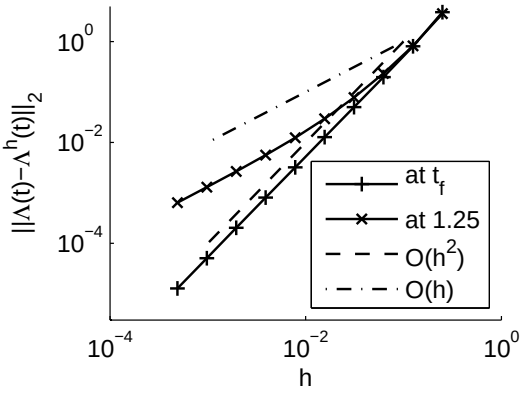


Figure 10.2: Convergence of the FE weak adjoints to the analytic weak adjoints. Error evaluated at the final time $t_\mathrm{f} = 2$ and at the interior time point $t = 1.25$.

Figure 10.2 shows the Euclidean norm of the difference between the analytic weak adjoint (10.2) and the FE approximation, i.e.

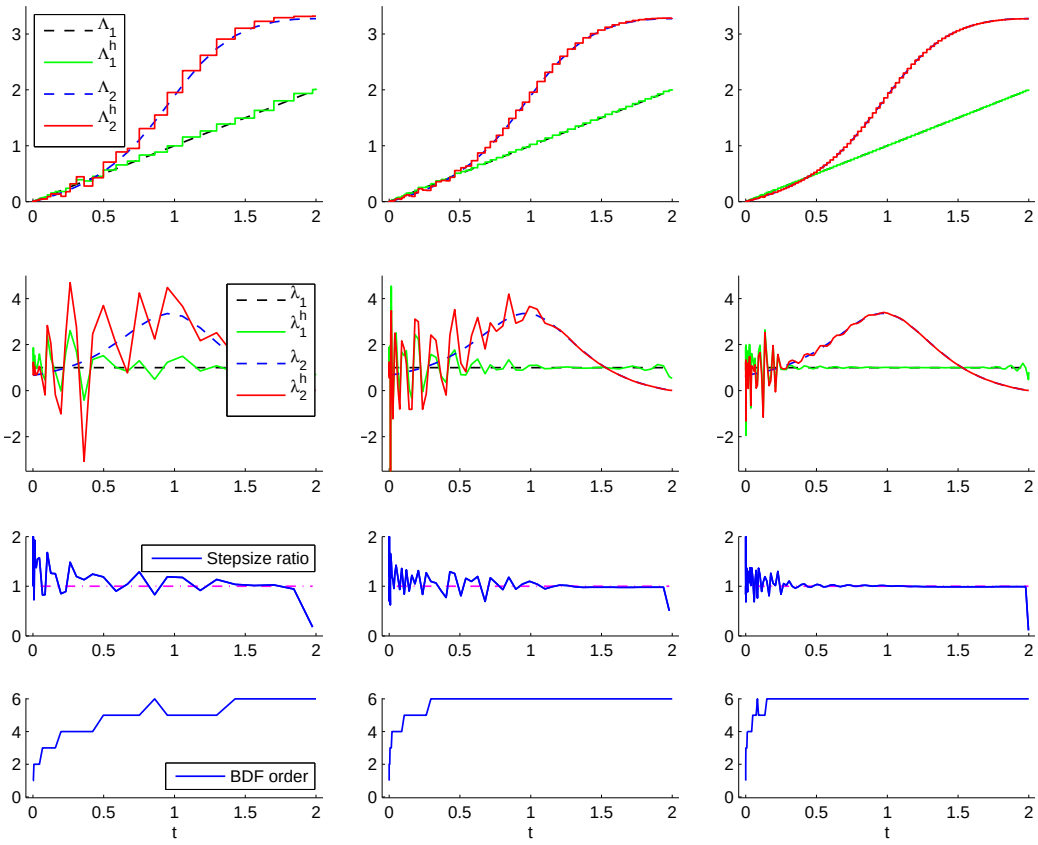
$$\text{Error} = \left\| \boldsymbol{\Lambda}(t) - \boldsymbol{\Lambda}^h(t) \right\|_2,$$

evaluated at the final time $t = t_\mathrm{f} = 2$ and at some interior time point $t = 1.25$, respectively, for decreasing stepsizes. The error evaluated at the final time decreases at second order rate, a somewhat better behavior than predicted by the convergence theory of Section 7.2 (Theorem 7.7 and the subsequent comment). This might be due to the second order convergence of the discrete adjoint $\boldsymbol{\lambda}_0$ at the initial time together with a possible cancellation of discrepancies of the discrete adjoints at the interval ends during the scaled summation of all $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ to give $\boldsymbol{\Lambda}^h$, see Chapter 6. Overall, this observation calls for a closer theoretical investigation. The error at the interior time point $t = 1.25$ shows the expected linear convergence, cf. Theorem 7.7 and the subsequent comment on the pointwise convergence.

10.1.2 Variable BDF method

We use the variable BDF-type method `DAESOL-II`, see Section 2.4 or Albersmeyer [3], to solve the Catenary for three different relative tolerances `RelTol` $= 10^{-4}$, 10^{-7} and 10^{-9}. We allow the method to use all strategies for an efficient integration and derivative generation, i.e. the stepsize and order selection rule, the monitor strategy (see Section 2.4) and iterative adjoint IND (see Section 3.4.2). Afterwards, we multiply the computed discrete adjoint IND values $\{\bar{\boldsymbol{y}}_n\}_{n=0}^N$ by the inverse of the Jacobians $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(t_{n+1}, \boldsymbol{y}_{n+1})$ to obtain $\{\boldsymbol{\lambda}_n\}_{n=0}^N$, see Lemma 3.3. The results are depicted in Figure 10.3.

In areas of constant BDF order (fourth row of Figure 10.3) and constant stepsizes (third row; depicting the stepsize ratio defined on page 11), the discrete adjoints



(a) `RelTol` $= 10^{-4}$ (b) `RelTol` $= 10^{-7}$ (c) `RelTol` $= 10^{-9}$

Figure 10.3: Results for the variable BDF-type method `DAESOL-II`. Comparison of the FE weak adjoints $\mathbf{\Lambda}^h$ to the analytic weak adjoints $\mathbf{\Lambda}$ (top) as well as the discrete adjoints $\boldsymbol{\lambda}^h$ compared to the analytic classical adjoints $\boldsymbol{\lambda}$ (second row). Stepsize ratio (third row) and BDF order (bottom) of the integration scheme are also depicted.

$\boldsymbol{\lambda}^h$ converge to the analytic adjoint solution $\boldsymbol{\lambda}$ (second row) as seen in the right column on the interval $(1, 1.7)$ approximately. In the other areas, i.e. where the order is varying and stepsize is changing, the discrete adjoints $\boldsymbol{\lambda}^h$ are highly oscillating (second row). Nevertheless, also in these cases, the FE approximations $\boldsymbol{\Lambda}^h$ converge to the analytic weak adjoint solution (10.2) on the entire time interval (first row of Figure 10.3).

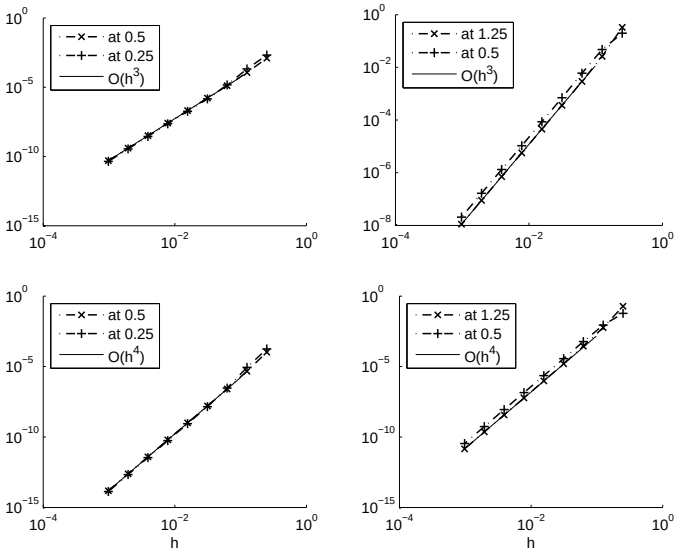
These examples give numerical evidence that the FE approximation serves as proper quantity to approximate the weak adjoint solution also for variable BDF-type methods with iterative adjoint IND as described in Section 2.4 and 3.4.2, i.e. also in areas of variable order and variable stepsize as well as in conjunction with efficient Newton-type methods and iterative adjoint IND.

10.2 Goal-oriented global error approximation for constant BDF methods

We use again our Matlab® implementation of constant BDF methods. To investigate the goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ derived in Section 8.3 as well as their behavior and relations described in Section 8.4 we augmented the program to evaluate the local errors $\mathbf{LE}(t_{n+1})$, the local truncation errors $\mathbf{LTE}(t_{n+1})$ and the defect integrals $\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t$. For $\mathbf{LE}(t_{n+1})$ and $\mathbf{LTE}(t_{n+1})$ we used (2.7) and (2.15) with analytic expressions, respectively, and for $\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t$ we used numerical quadrature. Although, the nonlinear BDF equations are solved up to a given small accuracy we keep the residual terms $\sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\intercal}\boldsymbol{\delta}_{n+1}$ in the goal-oriented error approximations.

10.2.1 Behavior of the defect integrals

Exemplarily, we use a constant BDF method of order $k = 2$ and stepsize h with two first-order BDF steps of size $h/2$ as self-starter to solve the Dahlquist equation (Example 1 in Section A.3.1) with $a = 0.5$, $y_\mathrm{s} = 1$, $[t_\mathrm{s}, t_\mathrm{f}] = [0, 1]$ and the Catenary described in Section 10.1 (see also Example 7 in Section A.3.1). We verify that the defect integrals converge with order $k + 1$ to zero as demonstrated and utilized in the proof of Theorem 8.5 and furthermore that the relation between the defect integrals and the local error given by Lemma 8.2 holds. Therefore, we compute both quantities $\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t$ and $\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t + \mathbf{LE}(t_{n+1})$ at two different inner time points, for the Dahlquist equation $t = 0.5$, $t = 0.25$ and for the Catenary $t = 1.25$, $t = 0.5$, take the norms and decrease the stepsize successively. The results are visualized in Figure 10.4 and the theoretical findings, i.e. convergence of order $k + 1 = 3$ and $k + 2 = 4$, respectively, are confirmed also numerically.



(a) Dahlquist (b) Catenary

Figure 10.4: Norm of the defect integral $\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t$ (top) and the relation $\int_{t_n}^{t_{n+1}} \boldsymbol{r}_n(t)\,\mathrm{d}t + \mathbf{LE}(t_{n+1})$ of Lemma 8.2 (bottom) evaluated at two inner reference time points for BDF order $k = 2$.

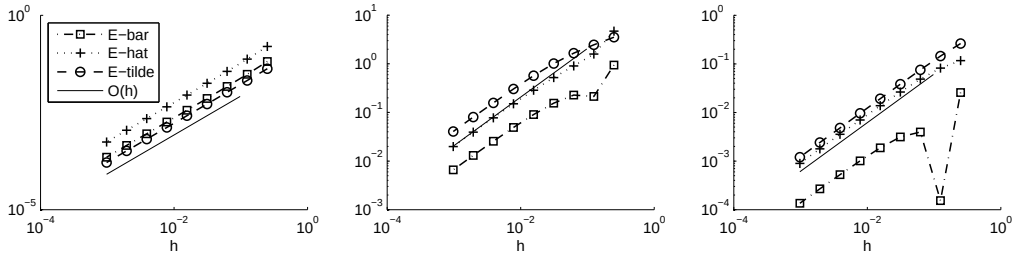
10.2.2 Accuracy of the goal-oriented error approximations

We investigate numerically the results of Section 8.4 on the asymptotic behavior of the three error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$. We use again the linear Dahlquist equation with $a = 0.5$, $y_\mathrm{s} = 1$, $[t_\mathrm{s}, t_\mathrm{f}] = [0, 1]$, the nonlinear Catenary and furthermore the nonlinear Example 2 described in Section A.3.1 as IVP test cases. The criterions of interest are $J(y(t_\mathrm{f})) = y(t_\mathrm{f})$ and $J(\boldsymbol{y}(t_\mathrm{f})) = y_1(t_\mathrm{f})$, respectively. Firstly, we focus on the one-step BDF method with order $k = 1$ and secondly we use the constant BDF method with order $k = 2$ as example for multistep BDF methods.

One-step method: BDF method of order $k = 1$

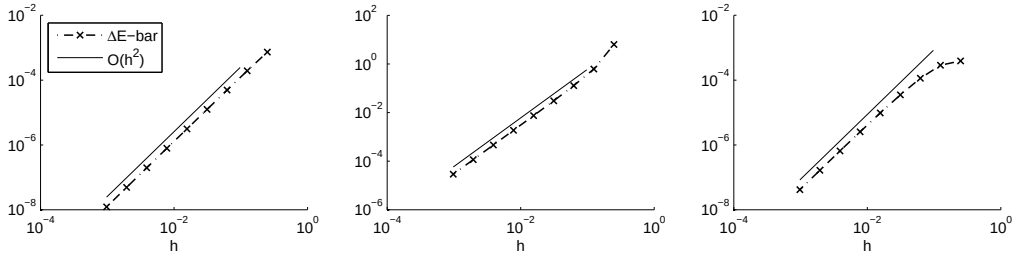
For a constant BDF method of order one the signed effectivity indices $I_\mathrm{eff}^\mathrm{s}$ defined by (8.11) of the three goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ to the true error $J(\boldsymbol{y}) - J(\boldsymbol{y}^h)$ are displayed in Figure 10.5. The indices of all three error approximations converge linearly to the desired value one, also for the nonlinear test cases. Hence, the numerical results confirm the theoretical findings given by Theorem 8.5 for $\bar{E}(\boldsymbol{y}^h)$, Corollary 8.6 for $\hat{E}(\boldsymbol{y}^h)$ and implicitly by Lemma 8.10 for $\tilde{E}(\boldsymbol{y}^h)$.

We also investigate the correction term defined by (8.19). According to Lemma 8.10 the correction terms $\Delta\bar{E}(\boldsymbol{y}^h)$ of $\bar{E}(\boldsymbol{y}^h)$ and $\Delta\hat{E}(\boldsymbol{y}^h)$ of $\hat{E}(\boldsymbol{y}^h)$ coincide for BDF order one and converge quadratically to zero. This is confirmed numerically as depicted in Figure 10.6.



(a) Dahlquist (b) Catenary (c) Example 2

Figure 10.5: $|I^{\mathrm{s}}_{\mathrm{eff}} - 1|$ of $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ for BDF order $k = 1$ and stepsize h.



(a) Dahlquist (b) Catenary (c) Example 2

Figure 10.6: $\left|\Delta\bar{E}(\boldsymbol{y}^h)\right|$ for BDF order $k = 1$ and stepsize h.

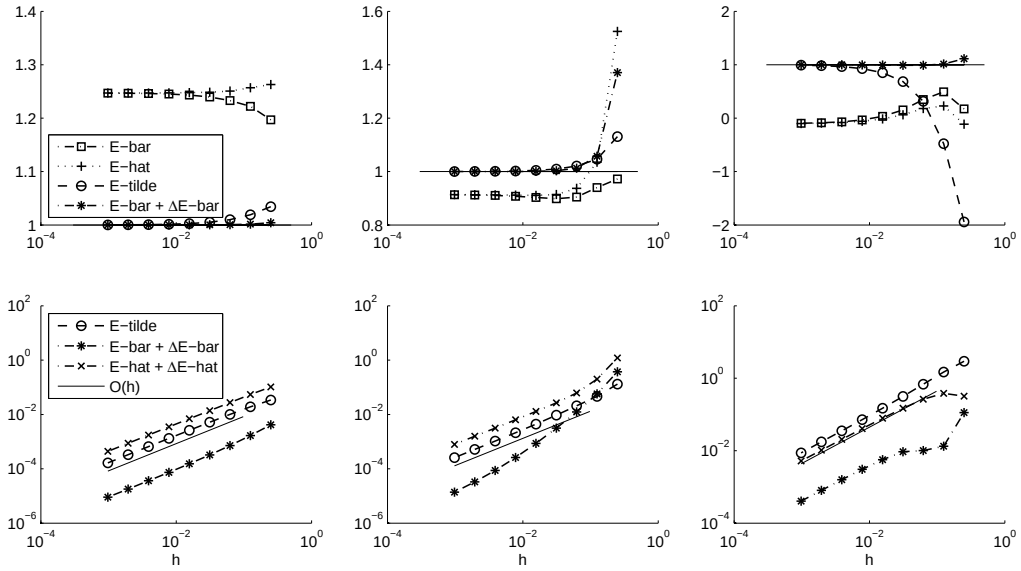
Multistep method: BDF method of order $k = 2$

For a constant BDF method of order two with two first-order steps of size $h/2$ the signed effectivity indices of the three goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ are visualized in the top row of Figure 10.7 for the three IVP examples. The effectivity indices of $\tilde{E}(\boldsymbol{y}^h)$ approach the desired value one whereas the other two error approximations $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ have effectivity indices that approach values different than one. In fact, they approach problem-dependent offset values. However, if $\bar{E}(\boldsymbol{y}^h)$ is corrected using $\Delta\bar{E}(\boldsymbol{y}^h)$ the effectivity indices of $\bar{E}(\boldsymbol{y}^h) + \Delta\bar{E}(\boldsymbol{y}^h)$ also approach one, cf. top row of Figure 10.7.

In the bottom row of Figure 10.7 the convergence of the signed effectivity indices of the error approximation $\tilde{E}(\boldsymbol{y}^h)$ and the corrected approximations $\bar{E}(\boldsymbol{y}^h)+\Delta\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h) + \Delta\hat{E}(\boldsymbol{y}^h)$ to the desired value one is depicted. In fact, the convergence to one is linear in all these error approximations and for all IVP examples.

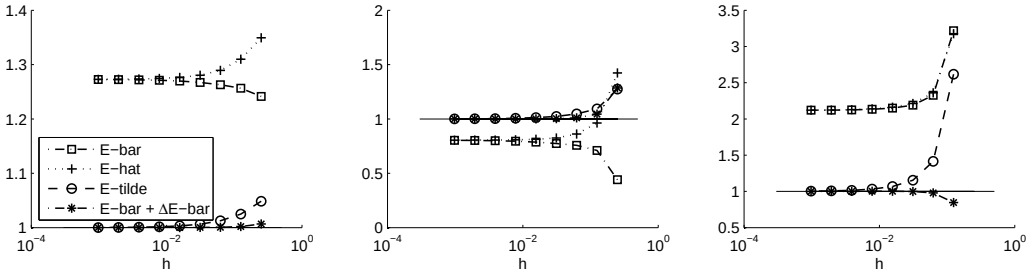
For another BDF method of order $k = 2$ that uses only one first-order step of size $h_0 = h$ as self-starter, and hence has all steps of the same size h, the signed effectivity indices of the three error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ are depicted in Figure 10.8.

Comparing Figure 10.8 to the first row of Figure 10.7 one observes that the



(a) Dahlquist (b) Catenary (c) Example 2

Figure 10.7: $I_{\text{eff}}^{\text{s}}$ of $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$, $\tilde{E}(\boldsymbol{y}^h)$ and $\bar{E}(\boldsymbol{y}^h) + \Delta\bar{E}(\boldsymbol{y}^h)$ (top) and $|I_{\text{eff}}^{\text{s}} - 1|$ of $\tilde{E}(\boldsymbol{y}^h)$, $\bar{E}(\boldsymbol{y}^h) + \Delta\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h) + \Delta\hat{E}(\boldsymbol{y}^h)$ (bottom) for BDF order $k = 2$ and stepsize h with two first-order steps.



(a) Dahlquist (b) Catenary (c) Example 2

Figure 10.8: $I_{\text{eff}}^{\text{s}}$ of $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$, $\tilde{E}(\boldsymbol{y}^h)$ and $\bar{E}(\boldsymbol{y}^h) + \Delta\bar{E}(\boldsymbol{y}^h)$ for BDF order $k = 2$ and stepsize h with one first-order step.

problem-dependent offset values actually also depend on the used nominal integration scheme. The convergence behavior of the error approximations and their corrected versions for this BDF method of order two are similar to the behavior for the method with two starting steps.

10.3 Goal-oriented global error estimation for variable BDF methods

In this section we investigate the accuracy of the error estimators $\bar{\eta}$, $\hat{\eta}$ and $\tilde{\eta}$ for variable order variable stepsize BDF methods. The nominal integration is done by the fully adaptive BDF-type method `DAESOL-II` with local stepsize and order strategies and the discrete adjoints are computed by iterative adjoint IND realized in `DAESOL-II`, cf. Section 2.4 and 3.4.2 or Albersmeyer [3]. We again use IVPs with analytic solutions as test cases. They are listed in Section A.3.1. To increase the number of tests we use different relative tolerance `RelTol` for the integration.

For each integration and each error estimator $\check{\eta}$ we compute the signed effectivity index $I_{\text{eff}}^{\text{s}}$ defined by (8.11). If the signed index is positive, the estimator reflects the right sign of the true error. If it is negative, the estimator was not able to reflect the right sign of the error. The closer the absolute value of the effectivity index is to one, the higher the accuracy of the estimator is. However, the estimator *overestimates* the true error if the index is greater than one and *underestimates* it if the index is less than one. Due to the fully-adaptive nominal integration with changing stepsizes and varying orders a reasonably bounded absolute value of the effectivity index is already satisfactory in practice. However, the quality of the estimator is considered to be the same regardless whether the effectivity index is e.g. 0.5 or 2. In the whole thesis we round the absolute value of the effectivity index of our goal-oriented error estimators to our disadvantage, i.e. an index of 0.367 becomes 0.36 and 3.451 becomes 3.46. We also compare our estimators to a corresponding a posteriori global error estimator proposed by Cao and Petzold [43] and denoted by η_{CP} for scalar criterions of interest and by $\boldsymbol{\eta}_{\text{CP}}$ for vector-valued criterions. The *error indices*

$$I_{\text{err}} := \frac{\|\boldsymbol{\eta}_{\text{CP}}\|}{\|\boldsymbol{J}(\boldsymbol{y}(t_{\text{f}})) - \boldsymbol{J}(\boldsymbol{y}_N)\|} \tag{10.3}$$

of the estimator $\boldsymbol{\eta}_{\text{CP}}$ by Cao and Petzold are provided by Tran and Berzins [118]. For scalar criterions of interest J both the effectivity and the error indices coincide (except sign). Unfortunately, for vector-valued criterions $\boldsymbol{J}$ the error index is not as precise as the effectivity indices of the single components of $\boldsymbol{J}$ to measure the accuracy of the estimator.

10.3.1 Accuracy of the goal-oriented error estimators

In contrast to the asymptotic investigations of Section 8.4.3 and 8.4 for the error approximations we examine here the accuracy of all three goal-oriented error estimators in practical use with variable BDF methods and possibly large stepsizes. As

Example	Estimator	RelTol 10^{-3}	10^{-4}	10^{-5}	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
3	$\bar{\boldsymbol{\eta}}$	0.62	0.87	0.97	1.02	1.03	1.03	1.01	1.02
		1.51	2.19	-17.27	-0.03	0.58	0.79	0.88	0.92
	$\hat{\boldsymbol{\eta}}$	0.37	0.43	0.40	0.47	0.45	0.46	0.47	0.45
		0.61	0.74	-10.50	-0.12	0.20	0.33	0.43	0.41
	$\tilde{\boldsymbol{\eta}}$	0.90	1.05	0.99	1.14	1.13	1.15	1.20	1.13
		1.48	1.76	-23.42	-0.29	0.53	0.82	1.05	1.01
	$\bar{\boldsymbol{\eta}}$	0.96	1.01	1.01	1.01	0.99	0.99	0.98	0.99
	$\hat{\boldsymbol{\eta}}$	0.45	0.45	0.42	0.46	0.44	0.45	0.46	0.44
	$\tilde{\boldsymbol{\eta}}$	1.11	1.11	1.05	1.16	1.09	1.12	1.17	1.11
	$\boldsymbol{\eta}_{\mathrm{CP}}$	13.58	13.02	13.66	13.00	11.59	10.92	10.77	11.35
Cat.	$\bar{\eta}$	0.53	0.47	0.76	1.02	1.02	0.99	0.98	0.99
	$\hat{\eta}$	0.33	0.34	0.24	0.45	0.47	0.42	0.55	0.41
	$\tilde{\eta}$	0.84	-0.07	0.83	1.20	1.13	1.06	1.36	1.02

Table 10.1: Signed effectivity indices $I_{\mathrm{eff}}^{\mathrm{s}}$ and error indices I_{err} (fourth and fifth part of Example 3) of $\bar{\eta}$, $\hat{\eta}$ and $\tilde{\eta}$ for variable BDF methods with decreasing relative tolerances `RelTol`. Error indices of goal-oriented estimator η_{CP} proposed by Cao and Petzold [43] are provided by Tran and Berzins [118].

test cases we use the linear IVP system of Example 3 with vector-valued criterion of interest $\boldsymbol{J}(\boldsymbol{y}(t_{\mathrm{f}})) = \boldsymbol{y}(t_{\mathrm{f}})$ and the nonlinear Catenary with nonlinear criterion of interest $J_3(\boldsymbol{y}(t_{\mathrm{f}})) = y_1(t_{\mathrm{f}}) \cdot y_2(t_{\mathrm{f}})$. The signed effectivity indices are given in Table 10.1 for both IVPs and criterions of interest including each component of the vector-valued criterion. For the first example we also list in Table 10.1 the error indices given by Tran and Berzins [118] of the estimator $\boldsymbol{\eta}_{\mathrm{CP}}$ by Cao and Petzold [43]. Since effectivity index and error index only coincide for scalar criterions we also give the error indices of our estimators for the vector-valued criterion.

All effectivity indices of our novel goal-oriented estimators derived in Section 8.5 remain bounded according to Table 10.1 also for fully adaptive BDF-type methods and a wide span of relative tolerances. Actually, the effectivity indices of $\bar{\eta}$ and $\tilde{\eta}$ are mostly near the desired value one and there are only very few that are outside the interval $[0.5, 2]$. The effectivity indices of $\hat{\eta}$ are not that good but remain bounded as well. Due to the fact that neither stepsize nor order are constant one can not expect that the effectivity indices approach one for increasing accuracy requirements, i.e. for decreasing relative tolerances `RelTol`. One should rather understand the variety of relative tolerances as augmentation of the test set.

We start the detailed examination of the results in Table 10.1 with Example 3.

Comparing the accuracy of our three estimators one recognizes that the estimator $\bar{\boldsymbol{\eta}}$ based on the defect integrals and the estimator $\tilde{\boldsymbol{\eta}}$ based on the estimated local truncation errors are better in componentwise absolute values than $\hat{\boldsymbol{\eta}}$ based on the local error estimates of Section 8.5.2. The local error based estimator $\hat{\boldsymbol{\eta}}$ exhibits an underestimation tendency since its effectivity indices are always less than one. The other two estimators $\bar{\boldsymbol{\eta}}$ and $\tilde{\boldsymbol{\eta}}$ neither show an underestimation nor an overestimation tendency. However, all three estimators give the correct sign of the true goal-oriented global error in most of the integrations. In order to compare our estimators to the corresponding estimator $\boldsymbol{\eta}_{\mathrm{CP}}$ by Cao and Petzold [43] Table 10.1 also contains the error indices defined by (10.3) of $\bar{\boldsymbol{\eta}}$, $\hat{\boldsymbol{\eta}}$ and $\tilde{\boldsymbol{\eta}}$. Overall, the error indices of our estimators are closer to one than those of $\boldsymbol{\eta}_{\mathrm{CP}}$. Hence, all our goal-oriented error estimators behave superior to that by Cao and Petzold.

Secondly, we have a look at the global error in a nonlinear criterion of interest evaluated in the final state of the Catenary. Again for this example the error estimators $\bar{\eta}$ and $\tilde{\eta}$ behave significantly better in absolute values than $\hat{\eta}$. However, all estimators give the correct error sign in nearly all cases.

Conclusions

Overall, the numerical experiments of this section indicate that the goal-oriented estimators $\bar{\eta}$ and $\tilde{\eta}$ are generally more accurate in estimating the true global error in J than the estimator $\hat{\eta}$. We recapitulate now all insights we have gained on the different goal-oriented error estimators so far. In Section 8.2 and 8.3 we have seen the derivation of the goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ in function spaces whereas a function space interpretation of $\tilde{E}(\boldsymbol{y}^h)$ is an open issue so far. In Section 10.2.2 we have learned that for constant multistep BDF methods the signed effectivity indices of $\tilde{E}(\boldsymbol{y}^h)$ converge to the desired value one whereas those of $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$ converge to a problem- and method-dependent value $\neq 1$. Furthermore, the corresponding goal-oriented error estimators $\bar{\eta}$, $\hat{\eta}$ and $\tilde{\eta}$ derived in Section 8.5 have different computational costs. The most efficient one is $\tilde{\eta}$, directly followed by $\hat{\eta}$ whereas $\bar{\eta}$ can be more expensive, cf. Section 8.5.4. The cost for $\bar{\eta}$ depends directly on the quadrature formula, the required tolerance and the evaluation cost for $\boldsymbol{f}(t, \boldsymbol{y})$. All these insights are summarized in Table 10.2.

In fact, the estimator $\hat{\eta}$ is not that good due to the rough approximation of the local errors by $\widehat{\mathbf{LE}}(t_{n+1})$, cf. Section 8.5.2. The local errors could be approximated more accurately using an integration method of higher order for the local IVPs of Definition 2.4. Unfortunately, this would increase the computational effort tremendously. Nevertheless, we recommend to use $\bar{\eta}$ instead since it provides good accuracy and numerical quadrature is significantly more efficient than integration. For all these reasons we discard the goal-oriented error estimator $\hat{\eta}$ at this point from further numerical testing.

Estimator	Derivation in function spaces	$I_{\text{eff}}^{\text{s}} \to 1$ for constant multistep BDF methods	Computation at low cost	$\lvert I_{\text{eff}}^{\text{s}} \rvert$ close to one for practical computations
$\bar{E}(\boldsymbol{y}^h)$, $\bar{\eta}$	yes	no (yes with correction)	depends on quadrature	yes
$\hat{E}(\boldsymbol{y}^h)$, $\hat{\eta}$	yes	no (yes with correction)	yes	no
$\tilde{E}(\boldsymbol{y}^h)$, $\tilde{\eta}$	no	yes	yes	yes

Table 10.2: Summary of all investigated properties of the novel goal-oriented global error estimators derived in Chapter 8.

10.3.2 Impact of residuals

For the remaining goal-oriented estimators from the practicable point of view, i.e. for $\bar{\eta}$ based on defect integrals and $\tilde{\eta}$ based on estimated local truncation errors, we now investigate the impact of the weighted sum $\eta_{\boldsymbol{\delta}}$ of the residuals of the nonlinear BDF equations given by (8.23). The residuals result from the iterative solution of the nonlinear BDF equations by a Newton-type method, cf. Section 2.4.3 and 8.5.3. As test cases we use the scalar Dahlquist equation of Example 1, the linear Harmonic Oscillator of Example 4 with vector-valued criterion of interest $\boldsymbol{J}(\boldsymbol{y}(t_{\text{f}})) = \boldsymbol{y}(t_{\text{f}})$, the nonlinear IVP system of Example 5 with $\boldsymbol{J}(\boldsymbol{y}(t_{\text{f}})) = \boldsymbol{y}(t_{\text{f}})$ and Example 2 with nonlinear criterion $J_2(y(t_{\text{f}})) = 1/y(t_{\text{f}}) \cdot \exp(y(t_{\text{f}}))$. We integrate again with different relative tolerances and compute the signed effectivity indices of $\bar{\eta}$ and $\tilde{\eta}$ including $\eta_{\boldsymbol{\delta}}$ and neglecting $\eta_{\boldsymbol{\delta}}$. The results are given in Table 10.3 for the scalar criterions of interest and some components of the vector-valued criterions. For the first three test cases the error indices given by Tran and Berzins [118] of the estimator $\boldsymbol{\eta}_{\text{CP}}$ by Cao and Petzold [43] are listed as well.

Having a look at Example 1 in Table 10.3, the effectivity indices of the defect based estimator $\bar{\eta}$ with and without the residual term $\eta_{\boldsymbol{\delta}}$ differ for $\texttt{RelTol} < 10^{-3}$ only in the second decimal place. The same holds for the local truncation error based estimator $\tilde{\eta}$. Furthermore, all our goal-oriented estimators are much better than the estimator η_{CP} by Cao and Petzold. For Example 4 the discrepancy in the estimators with and without $\boldsymbol{\eta}_{\boldsymbol{\delta}}$ are at most in the second decimal places for both $\bar{\boldsymbol{\eta}}$ and $\tilde{\boldsymbol{\eta}}$. Moreover, the error indices of all our estimators are closer to one than those of $\boldsymbol{\eta}_{\text{CP}}$. For Example 5 with vector-valued criterion $\boldsymbol{J}(\boldsymbol{y}(t_{\text{f}})) = \boldsymbol{y}(t_{\text{f}})$, Table 10.3 only shows the effectivity indices of the global errors in the second and the fifth component of $\boldsymbol{J}$. In all components of $\boldsymbol{J}$ the discrepancy in the effectivity indices of $\bar{\boldsymbol{\eta}}$ with and without $\boldsymbol{\eta}_{\boldsymbol{\delta}}$ is at most in the second decimal place, expect in five of the twenty cases (five criterions, four relative tolerances) where it is in the first decimal place. The same holds true for $\tilde{\boldsymbol{\eta}}$. Comparing the error indices of our estimators to that of Cao and Petzold again our estimators are more accurate. For the nonlinear Example 2 with nonlinear criterion the discrepancy caused by $\eta_{\boldsymbol{\delta}}$ is once more in the

Example	Estimator	RelTol 10^{-3} w $\eta_{\boldsymbol{\delta}}$	RelTol 10^{-3} w/o $\eta_{\boldsymbol{\delta}}$	RelTol 10^{-5} w $\eta_{\boldsymbol{\delta}}$	RelTol 10^{-5} w/o $\eta_{\boldsymbol{\delta}}$	RelTol 10^{-7} w $\eta_{\boldsymbol{\delta}}$	RelTol 10^{-7} w/o $\eta_{\boldsymbol{\delta}}$	RelTol 10^{-9} w $\eta_{\boldsymbol{\delta}}$	RelTol 10^{-9} w/o $\eta_{\boldsymbol{\delta}}$
1	$\bar{\eta}$	1.083	0.877	1.002	0.966	1.105	1.071	1.120	1.072
	$\tilde{\eta}$	1.422	1.217	0.595	0.560	1.007	0.972	0.761	0.713
	η_{CP}	7.13		7.09		8.95		16.72	
4	$\bar{\boldsymbol{\eta}}$	-6.211	-6.204	0.706	0.680	0.914	0.909	0.968	0.967
		0.943	0.909	1.157	1.157	1.166	1.171	1.125	1.134
	$\tilde{\boldsymbol{\eta}}$	-7.336	-7.329	0.887	0.861	1.027	1.022	1.069	1.068
		1.247	1.213	1.337	1.338	1.282	1.287	1.249	1.258
	$\bar{\boldsymbol{\eta}}$	1.016	0.984	1.011	1.004	1.001	0.999	0.997	0.999
	$\tilde{\boldsymbol{\eta}}$	1.324	1.292	1.188	1.181	1.114	1.113	1.103	1.105
	$\boldsymbol{\eta}_{\text{CP}}$	4.13		15.04		1.45		8.64	
5	$\bar{\eta}_2$	1.283	1.254	1.390	1.279	1.301	1.223	1.494	1.342
	$\bar{\eta}_5$	1.028	1.022	0.997	0.954	0.974	0.968	1.002	0.990
	$\tilde{\eta}_2$	0.404	0.374	0.740	0.628	1.188	1.111	2.110	1.957
	$\tilde{\eta}_5$	0.560	0.554	0.926	0.883	1.047	1.041	1.156	1.146
	$\bar{\boldsymbol{\eta}}$	1.044	1.037	1.007	0.961	0.977	0.970	1.004	0.993
	$\tilde{\boldsymbol{\eta}}$	0.551	0.544	0.924	0.880	1.048	1.041	1.161	1.150
	$\boldsymbol{\eta}_{\text{CP}}$	6.14		14.31		12.94		8.35	
2	$\bar{\eta}$	1.214	1.183	0.907	0.858	1.118	1.174	0.858	0.909
	$\tilde{\eta}$	0.918	0.887	-0.590	-0.639	0.859	0.916	1.165	1.217

Table 10.3: Signed effectivity indices $I^{\text{s}}_{\text{eff}}$ and error indices I_{err} (third and fourth part of Example 4 and 5) of $\bar{\eta}$ and $\tilde{\eta}$ with and without $\eta_{\boldsymbol{\delta}}$ for variable BDF methods with decreasing relative tolerances `RelTol`. Error indices of goal-oriented estimator η_{CP} proposed by Cao and Petzold [43] are provided by Tran and Berzins [118].

second decimal place for both estimators $\bar{\eta}$ and $\tilde{\eta}$. To sum up, the numerical results confirm that the impact of $\eta_{\boldsymbol{\delta}}$ on the goal-oriented error estimators $\bar{\eta}$ and $\tilde{\eta}$ is quite small and there is no plain tendency that including $\eta_{\boldsymbol{\delta}}$ would yield a more accurate estimator. Hence, the computational cost for the evaluation of the goal-oriented error estimators $\bar{\eta}$ and $\tilde{\eta}$ can be further reduced by neglecting the residual term $\eta_{\boldsymbol{\delta}}$ which might be expensive to evaluate if the evaluation of the ODE right hand side $\boldsymbol{f}(t, \boldsymbol{y})$ is computationally expensive.

10.4 Summary

In the first part of this chapter we confirmed numerically the results on the FE approximations of weak adjoints using discrete adjoint IND values of multistep BDF methods derived in Part II. Firstly, the convergence results have been observed numerically using a constant multistep BDF method to solve the nonlinear Catenary. Secondly, we have given numerical evidence that the FE approximation serves as proper quantity to approximate the weak adjoints also in the case of fully adaptive BDF-type methods, i.e. also in areas of variable order and variable stepsize.

In Section 10.2 we demonstrated numerically that for constant multistep BDF methods the signed effectivity indices of the goal-oriented global error approximations $\bar{E}(\boldsymbol{y}^h)$, $\hat{E}(\boldsymbol{y}^h)$ and $\tilde{E}(\boldsymbol{y}^h)$ derived in Section 8.2 and 8.3 converge to different limit values. The indices of the approximation $\tilde{E}(\boldsymbol{y}^h)$ based on local truncation errors converge to the desired value one whereas the indices of $\bar{E}(\boldsymbol{y}^h)$ based on defect integrals and $\hat{E}(\boldsymbol{y}^h)$ based on local errors converge to problem- and method-dependent value $\neq 1$, respectively.

Finally, we sum up the results of the last part on goal-oriented global error estimation for fully variable BDF-type methods. The numerical experiments indicated the superiority of our novel goal-oriented error estimators $\bar{\eta}$, $\hat{\eta}$ and $\tilde{\eta}$ derived in Chapter 8 compared to the corresponding, existing estimator η_{CP} developed by Cao and Petzold [43] and investigated by Tran and Berzins [118]. Comparing our three estimators with each other we have learned that the estimator $\bar{\eta}$ based on defect integrals and $\tilde{\eta}$ based on estimated local truncation errors are more accurate than $\hat{\eta}$ in estimating the true error $J(\boldsymbol{y}(t_{\mathrm{f}})) - J(\boldsymbol{y}^h(t_{\mathrm{f}}))$. For this reason and some others, cf. Table 10.2, we have discarded the estimator $\hat{\eta}$ and remain with $\bar{\eta}$ and $\tilde{\eta}$. Furthermore, we observed that the residual term $\eta_{\boldsymbol{\delta}}$ is negligible in practical calculations, hence we may approximate it by zero. Overall, the signed effectivity indices of both estimators $\bar{\eta}$ and $\tilde{\eta}$ demonstrate that they give the correct sign of the true error in J in nearly all test cases. Furthermore, the indices show that both estimators exhibit neither an overestimation tendency nor an underestimation tendency. Thus, the constant c in the termination criterion (9.1) of goal-oriented adaption algorithms developed in Chapter 9 should be chosen to be one.

11 Integration with goal-oriented global error control

In this chapter we will examine the goal-oriented global error control strategies proposed in Chapter 9. They aim to give an approximation to the solution of an Initial Value Problem (IVP) with a goal-oriented global error that satisfies a user given tolerance `GTol`. To this end, the control strategies require that the goal-oriented error estimate $\check{\eta}$ fulfills (9.1), i.e. $|\check{\eta}| \leq c \cdot \texttt{GTol}$, for a prescribed positive constant c depending on the used estimator $\check{\eta}$. We utilize exemplarily the goal-oriented error estimator $\tilde{\eta}$ defined by (8.22) using estimated local truncation errors due to its favorable balancing of accuracy and computational efficiency, cf. Section 10.3. Furthermore, we suppose, as justified by Section 8.5.3 and 10.3.2, that the residual term $\eta_{\boldsymbol{\delta}}$ in $\tilde{\eta}$ is insignificant. The numerical examples of Section 10.3 have shown that our novel goal-oriented error estimator $\tilde{\eta}$ neither inclines to underestimate nor to overestimate the true goal-oriented error $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}^h(t_\mathrm{f}))$. Hence, we always take $c = 1$ in the termination criterion (9.1) as suggested in Section 10.4.

To get an impression of the Backward Differentiation Formula (BDF) method used to obtain the approximate solutions, we will state for every integration the number N of (successful) integration steps, the overall Newton iterations $\sum_{n=0}^{N-1} s_n$ as well as the matrix rebuilds and decompositions needed by the Newton-type method to solve the nonlinear BDF equations, cf. Section 2.4. Note that one rebuild is always caused by the initial setup of the iteration matrix.

11.1 Goal-oriented local tolerance adaption

In this section we examine the goal-oriented local tolerance adaption of Algorithm 1 proposed in Section 9.1. We take $c_\mathrm{red} = 0.2$ as factor to ensure reduction in the relative tolerance.

11.1.1 Linear IVP with time-varying coefficient matrix

We start with Example 3

$$\dot{\boldsymbol{y}}(t) = \begin{pmatrix} \frac{1}{2(1+t)} & -2t \\ 2t & \frac{1}{2(1+t)} \end{pmatrix} \boldsymbol{y}(t), \quad t \in (0, 10], \quad \boldsymbol{y}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

It should be solved with low accuracy such that the error in $J(\boldsymbol{y}(t_\mathrm{f})) = y_1(t_\mathrm{f})$ is less than the global tolerance $\texttt{GTol} = 4 \cdot 10^{-4}$. As relative tolerance for the first

j	RelTol	$\tilde{\eta}$	$J(\boldsymbol{y}(t_\mathrm{f}))-J(\boldsymbol{y}_N)$	$I_\mathrm{eff}^\mathrm{s}$	N	$\sum s_n$	reb/dec
0	2.000000e-04	7.287345e-02	7.028072e-02	1.037	313	687	2 / 4
1	1.097793e-06	7.928640e-04	7.075118e-04	1.121	663	1923	2 / 2
2	2.195587e-07	1.825734e-04	1.807485e-04	1.011	830	2233	2 / 2

Table 11.1: Results of the goal-oriented local tolerance adaption applied to Example 3 with $J(\boldsymbol{y}(t_\mathrm{f})) = y_1(t_\mathrm{f})$, $\texttt{RelTol} = 2 \cdot 10^{-4}$ and $\texttt{GTol} = 4 \cdot 10^{-4}$.

integration we choose $\texttt{RelTol} = 2 \cdot 10^{-4}$. Since $c = 1$ due to Section 10.4 the termination criterion on the goal-oriented error estimate $\tilde{\eta}$ is $|\tilde{\eta}| < c \cdot \texttt{GTol} = 4 \cdot 10^{-4}$. The results are given in Table 11.1.

Even for the simple IVP of Example 3 limiting the local truncation errors does not limit the global error to the same magnitude. As seen in the first row of Table 11.1, an integration with a relative tolerance of $2 \cdot 10^{-4}$ yields an approximation with an exact global error in $y_1(t_\mathrm{f})$ of magnitude $7.03 \cdot 10^{-2}$, that is an error accumulation factor of around 350. The error accumulation is caused by the instability of the IVP. The real parts of the eigenvalues of $\boldsymbol{f}_{\boldsymbol{y}}(t, \boldsymbol{y}(t))$ are $0.5/(1+t)$ and hence positive such that the IVP is unstable, cf. Section 1.3. Nevertheless, within three iterations, i.e. after three integrations with successively reduced relative tolerances, the estimated global error has been reduced below $c \cdot \texttt{GTol} = 4 \cdot 10^{-4}$, see Table 11.1. The true global errors $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}_N)$ stated in the fourth column confirm the suitability of the goal-oriented local tolerance adaption to reduce global error. To meet the global accuracy requirement a relative tolerance of $2.2 \cdot 10^{-7}$ has been necessary. The signed effectivity indices $I_\mathrm{eff}^\mathrm{s}$ of $\tilde{\eta}$ defined by (8.11) again show the suitability of $\tilde{\eta}$ in estimating the true global error in J in both magnitude and sign.

Already this small IVP example shows the difficulty in choosing the relative tolerance appropriately to meet a desired global integration accuracy.

11.1.2 Inhomogeneous linear IVP

Secondly, we consider Example 6

$$\dot{y}(t) = -L\,[y(t) - \sin(\pi t)] + \pi \cos(\pi t), \quad t \in (0, 1], \quad y(0) = 0$$

with $L = 50$. We solve this IVP very accurately such that the global error in $J(y(t_\mathrm{f})) = y(t_\mathrm{f})$ is less than $\texttt{GTol} = 2 \cdot 10^{-10}$. As relative tolerance for the first integration we choose $\texttt{RelTol} = 10^{-3}$ and might expect that many iterations are necessary. The results are stated in Table 11.2.

Already in the first integration with $\texttt{RelTol}^0 = 10^{-3}$ the local truncation errors are damped out such that a nominal approximation with exact global error $1.2 \cdot 10^{-4}$ has been computed, cf. Table 11.2. This damping of local inaccuracies is due to the asymptotic stability of the IVP since the eigenvalue of $f_y(t, y(t)) = -L$ has a negative real part, cf. Section 1.3. Although the second integration with adjusted relative

j	RelTol	$\tilde{\eta}$	$J(y(t_\mathrm{f}))-J(y_N)$	$I_\mathrm{eff}^\mathrm{s}$	N	$\sum s_n$	reb/dec
0	1.000000e-03	4.791418e-04	1.151663e-04	4.161	23	41	2 / 6
1	4.174130e-10	-3.657937e-10	-1.567932e-10	2.333	103	206	5 / 4
2	8.348260e-11	-5.776368e-11	-4.281387e-11	1.350	118	257	6 / 4

Table 11.2: Results of the goal-oriented local tolerance adaption applied to Example 6 with $J(y(t_\mathrm{f})) = y(t_\mathrm{f})$, $\texttt{RelTol} = 10^{-3}$ and $\texttt{GTol} = 2 \cdot 10^{-10}$.

tolerance yields an approximation with required exact global accuracy (fourth column in second row of Table 11.2), the goal-oriented local tolerance adaption has not terminated because the error estimator $\tilde{\eta}$ does not yet satisfy the termination criterion (9.1). The estimator slightly overestimates the true error. Nevertheless, the next adaption of the relative tolerance gives an approximation with desired accuracy. Furthermore, for decreasing relative tolerances the estimator $\tilde{\eta}$ becomes better in estimating the true error $J(y(t_\mathrm{f})) - J(y_N)$ again in both magnitude and sign, see fifth column of Table 11.2. This is caused by the smaller integration steps chosen in the nominal integrations since these smaller steps also imply an improvement in the approximation of the solution of the adjoint IVP.

11.2 Goal-oriented scheme adaption

To incorporate the conditioning of the IVP locally we do not only use the goal-oriented error estimator $\tilde{\eta}$ itself but also its local error indicators $\{\tilde{\eta}_n\}_{n=1}^N$. Hence, we examine the goal-oriented scheme adaption given by Algorithm 2 in conjunction with the indicator-based adaption of the integration scheme described by Algorithm 3, see Section 9.2. We consider the same test cases as in Section 11.1.

11.2.1 Linear IVP with time-varying coefficient matrix

We again compute an approximation to the solution of Example 3 with a global accuracy of $\texttt{GTol} = 4 \cdot 10^{-4}$ in $J(\boldsymbol{y}(t_\mathrm{f})) = y_1(t_\mathrm{f})$ and use $\texttt{RelTol} = 2 \cdot 10^{-4}$ for the first integration, cf. Section 11.1.1. For the indicator-based adaption of Algorithm 3 we choose p = 0.3. The results are summarized in Table 11.3.

In each iteration of the goal-oriented scheme adaption the error estimate of the approximation is successfully reduced, see second column of Table 11.3. After five iterations we are done since the estimated error $\tilde{\eta}$ satisfies the termination criterion $|\tilde{\eta}| \leq c \cdot \texttt{GTol} = 4 \cdot 10^{-4}$. The true errors $J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}_N)$ stated in the third column demonstrate the capability of the global error control strategy using indicator-based scheme adaption to reduce the true goal-oriented error. The overall number of iterations and the number of integration steps of each iteration depend directly on the choice of the refinement rate p. If p is small, more iterations with less integration steps are needed. If p is big, less iterations with more integration steps are necessary.

j	$\tilde{\eta}$	$J(\boldsymbol{y}(t_\mathrm{f})) - J(\boldsymbol{y}_N)$	$I_\mathrm{eff}^\mathrm{s}$	N	$\sum s_n$	reb/dec
0	7.287345e-02	7.028072e-02	1.037	313	687	2 / 4
1	2.925130e-02	3.085997e-02	0.947	416	1047	4 / 104
2	4.248216e-03	4.118479e-03	1.032	545	1522	3 / 85
3	6.939722e-04	7.408250e-04	0.936	727	1917	3 / 214
4	1.790707e-04	1.877929e-04	0.953	955	2490	2 / 298

Table 11.3: Results of the goal-oriented scheme adaption with the indicator-based scheme adaption applied to Example 3 with $J(\boldsymbol{y}(t_\mathrm{f})) = y_1(t_\mathrm{f})$, $\texttt{RelTol} = 2 \cdot 10^{-4}$, $\texttt{GTol} = 4 \cdot 10^{-4}$ and p = 0.3.

However, p should not be chosen too big to guarantee that the local error indicators are good approximtions to the true error contribution of each integration step.

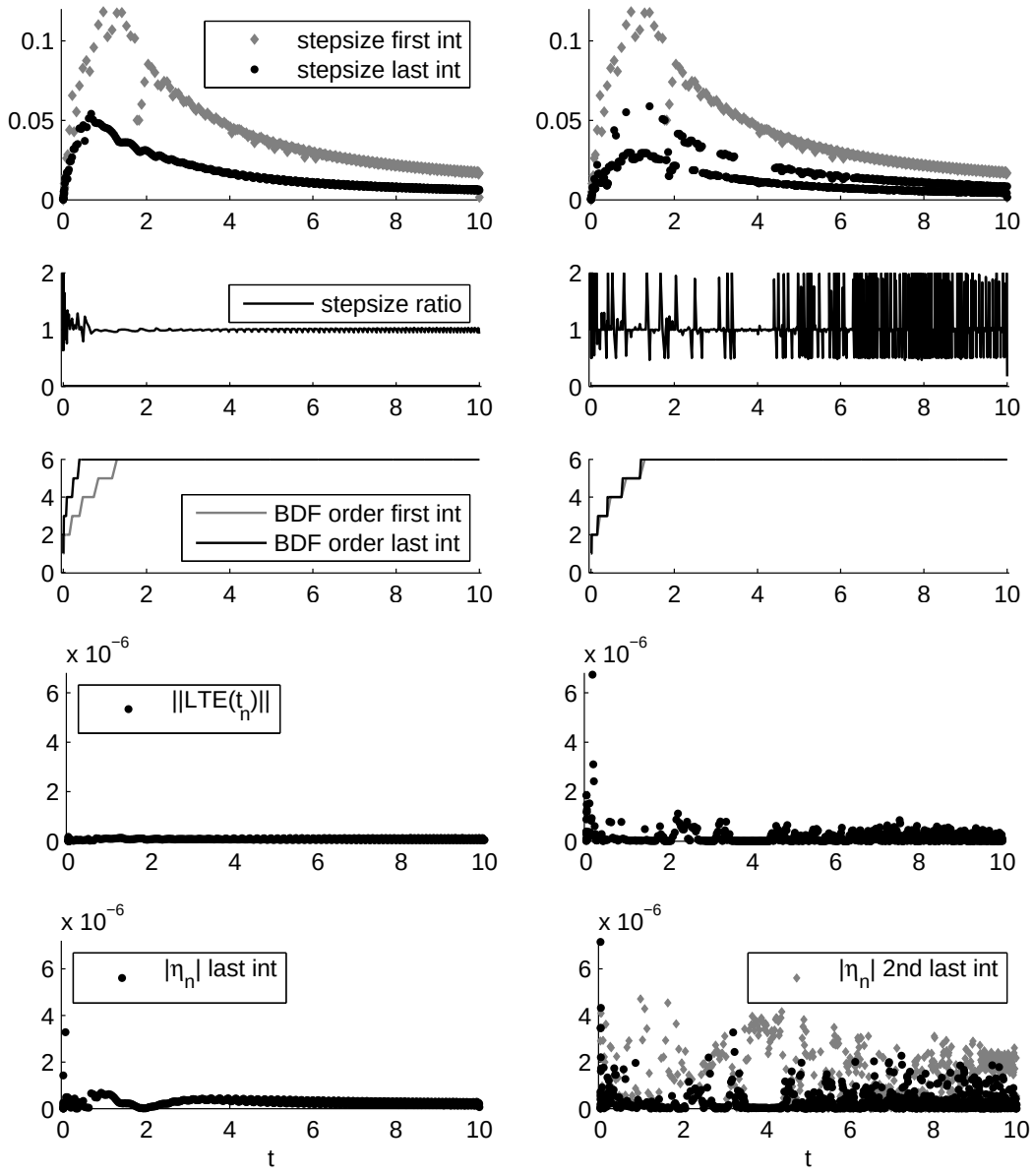
Comparing results of both goal-oriented adaption strategies

A comparison of the last iterations of both goal-oriented error control strategies shows that the goal-oriented scheme adaption (see Table 11.3) yields a slighly more expensive final integration in terms of integration steps and Newton iterations, and a much more expensive one in terms of matrix decompositions than the goal-oriented local tolerance adaption (see Table 11.1). The iteration matrix had to be decomposed that often since the stepsizes vary enormously as visualized in the second row of Figure 11.1(b) by the stepsize ratio defined on page 11.

Figure 11.1 displays the integration schemes and estimated quantities of the last iterations of both goal-oriented global error control strategies. The first row visualizes the stepsizes of the first integration with standard stepsize and order selection mechanism using $\texttt{RelTol}^0$ and those of the last integrations, respectively. In the second and the third row of Figure 11.1 the stepsize ratios and the BDF orders of the last integrations are depicted. In the penultimate row the norm of the estimated local truncation errors are depicted and in the last row the absolute values of the local error indicators of the goal-oriented global error estimator are given. The stepsize and order selection for the integration scheme of Figure 11.1(a) is based on the local truncation errors whereas the adaption for the integration scheme of Figure 11.1(b) relies on the local error indicators.

In fact, all estimated local truncation errors depicted in the fourth row of Figure 11.1(a), i.e. those on which the stepsize and order selection of Algorithm 1 is based, are bounded by the adapted relative tolerance $\texttt{RelTol}^2 = 2.195587 \cdot 10^{-7}$. Although, the estimated local truncation errors of Algorithm 2 are considerably greater according to the fourth row of Figure 11.1(b), the approximation fulfills the goal-oriented error bound as well. In this case the integration scheme is chosen by the indicator-based adaption using the local error indicators of $\tilde{\eta}^3$ depicted in gray in the last row of Figure 11.1(b).

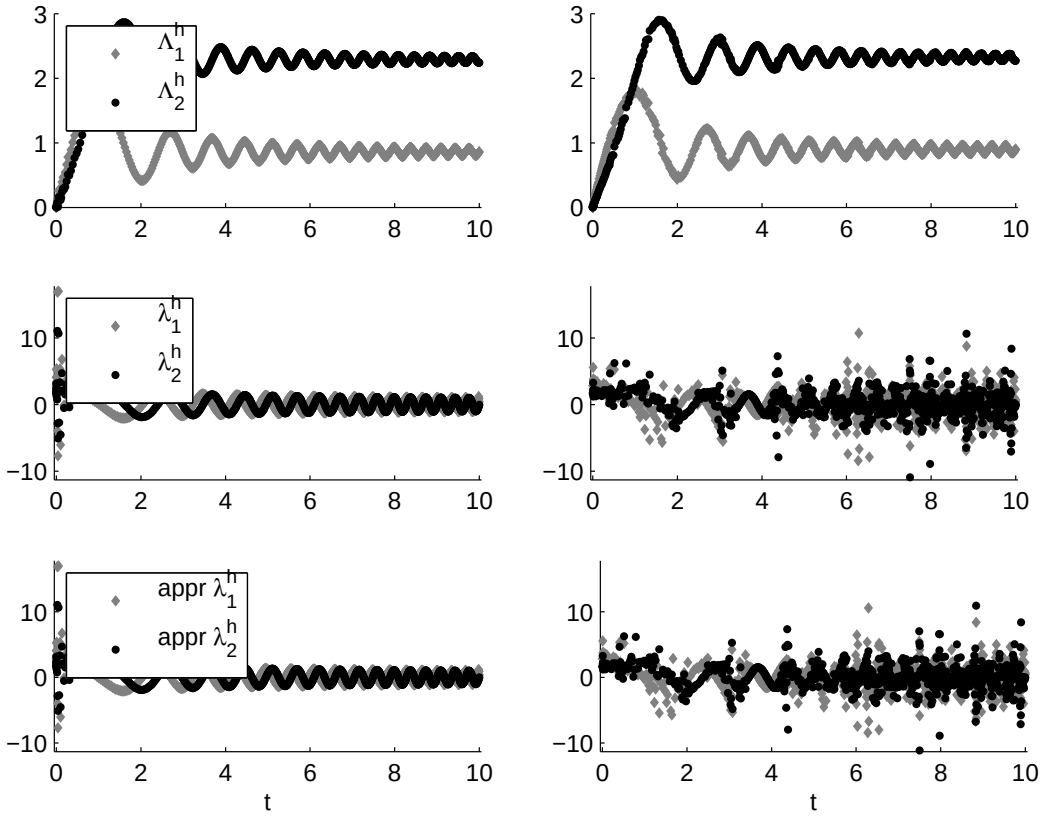
The stepsizes of the last integration of the scheme adaption strategy are generally



(a) Goal-oriented local tolerance adaption

(b) Goal-oriented scheme adaption

Figure 11.1: Comparison of the last iterations of both goal-oriented error control strategies applied to Example 3 with $J(\boldsymbol{y}(t_\mathrm{f})) = y_1(t_\mathrm{f})$, `RelTol` $= 2 \cdot 10^{-4}$, `GTol` $= 4 \cdot 10^{-4}$ and p $= 0.3$. Stepsizes (first row) and BDF orders (third row) of first (in gray) and last (in black) integration as well as stepsize ratios of last integration (second row) are given. The penultimate row shows the norm of the estimated local truncation errors and the last row the absolute value of the local error indicators.



(a) Goal-oriented local tolerance adaption

(b) Goal-oriented scheme adaption

Figure 11.2: FE weak adjoints $\mathbf{\Lambda}^h$ (first row), adjoint IND values $\boldsymbol{\lambda}^h = \{\boldsymbol{\lambda}_n\}_{n=0}^N$ (mid row) and approximated discrete adjoints $\widehat{\boldsymbol{\lambda}}^h = \{\widehat{\boldsymbol{\lambda}}_n\}_{n=0}^N$ (bottom row) of last iterations of both goal-oriented error control strategies applied to Example 3.

smaller at the left end of the interval $[0, 10]$ than those of the local tolerance adaption. The smaller steps seem to compensate on the one hand the slower increase of the BDF order, cf. third row of Figure 11.1. On the other hand, the adjoints indicate a worse conditioning of the IVP at the left interval end than on the orther parts of the interval. The adjoint Internal Numerical Differentiation (IND) values $\boldsymbol{\lambda}^h = \{\boldsymbol{\lambda}_n\}_{n=0}^N$, their approximations $\widehat{\boldsymbol{\lambda}}^h = \{\widehat{\boldsymbol{\lambda}}_n\}_{n=0}^N$ as well as their weak adjoints $\mathbf{\Lambda}^h$ are depicted in Figure 11.2 for both last integrations.

The huge fluctuation of the discrete adjoints of the scheme adaption strategy, see Figure 11.2(b), reflects the oscillations of the stepsizes. Nevertheless, as derived in Part II of this thesis, the approximated weak adjoints of both last integrations exhibit a smooth behavior on the entire time interval, see first row of Figure 11.2. Furthermore, they exhibit a remarkable gradient at the left interval end and a nearly zero gradient towards the right end which describes unstability with respect to errors at the beginning and an increasing stability towards the right end. Thus, in a huge

j	$\tilde{\eta}$	$J(y(t_\mathrm{f})) - J(y_N)$	$I_\mathrm{eff}^\mathrm{s}$	N	$\sum s_n$	reb/dec
0	4.791418e-04	1.151663e-04	4.161	23	41	2 / 6
1	1.887049e-05	7.612226e-06	2.479	30	57	1 / 8
2	4.389466e-07	2.561674e-07	1.714	36	74	1 / 7
3	3.012417e-08	1.886764e-08	1.597	43	89	1 / 13
4	3.703062e-09	2.612997e-09	1.418	51	110	1 / 14
5	9.928802e-10	7.962516e-10	1.247	61	138	1 / 15
6	1.939192e-10	1.564972e-10	1.240	72	167	1 / 18

Table 11.4: Results of the goal-oriented scheme adaption with the indicator-based scheme adaption applied to Example 6 with $J(y(t_\mathrm{f})) = y(t_\mathrm{f})$, $\texttt{RelTol} = 10^{-3}$, $\texttt{GTol} = 2 \cdot 10^{-10}$ and $\mathrm{p} = 0.18$.

area of the interval the local conditioning of the IVP has a rather small impact on the stepsize selection of the goal-oriented scheme adaption such that the latter strategy does not yield a better integration scheme than the standard strategies with adapted relative tolerance.

As already mentioned above, the oscillations in the stepsize sequence of integrations of Algorithm 2, depicted in the second row of Figure 11.1(b), are problematic in BDF methods since they cause a huge number of updates for the iteration matrix, cf. Table 11.3, and hence slow down the overall integration procedure.

11.2.2 Inhomogeneous linear IVP

We again compute an approximation to the solution of Example 6 with a global accuracy of $\texttt{GTol} = 2 \cdot 10^{-10}$ in $J(y(t_\mathrm{f})) = y(t_\mathrm{f})$, cf. Section 11.1.2. For the first integration we use again $\texttt{RelTol} = 10^{-3}$. For the indicator-based scheme adaption of Algorithm 3 we choose $\mathrm{p} = 0.18$. The results are depicted in Table 11.4.

In each iteration the estimated goal-oriented global error of the approximation is successfully reduced until it is below the required bound $c{\cdot}\texttt{GTol}$, cf. second column of Table 11.4. The same holds for the true goal-oriented error given in the third column. Due to the successive, implicit scheme adaption of only 18% of the integration steps, the scheme adaption needed quite a number of iterations, particularly in contrast to the goal-oriented local tolerance adaption, cf. Table 11.2. Nevertheless, in the last iteration the number of integration steps is reduced by more than one third compared to the last iteration of the local tolerance adaption strategy.

Comparing results of both goal-oriented adaption strategies

Comparing the last iterations of both goal-oriented error control strategies one notices that the goal-oriented scheme adaption (see Table 11.4) yields a more economic integration in contrast to the goal-oriented local tolerance adaption (see Table 11.2) in terms of integration steps and Newton iterations. But the iteration matrix of the

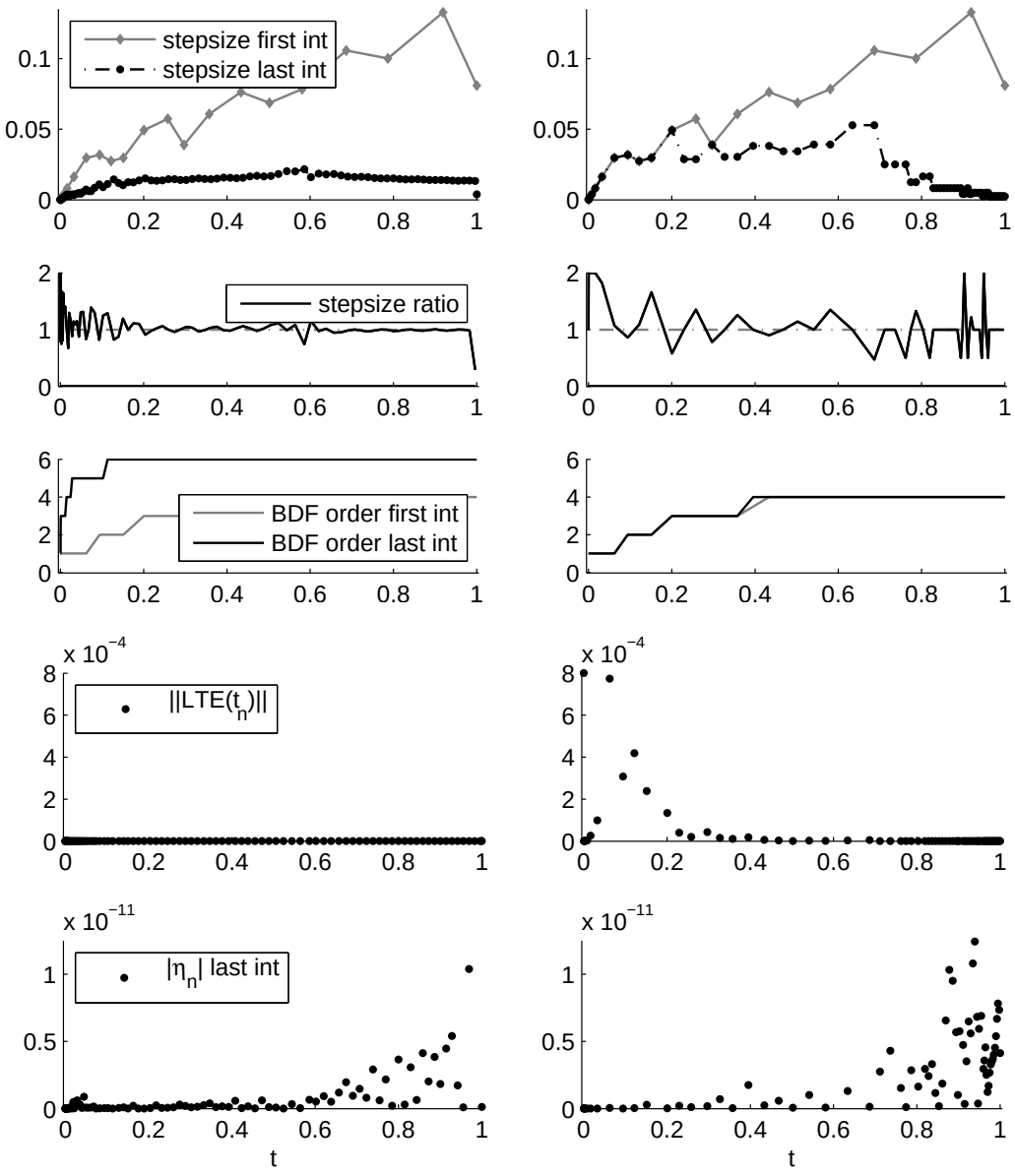
Newton-type method to solve the nonlinear BDF equations had to be decomposed more often. Nevertheless, only one rebuild which comprises the evaluation of the Jacobian of the Ordinary Differential Equation (ODE) right hand side $f(t, y)$ was necessary compared to six rebuilds in the local tolerance adaption. In Figure 11.3 the integration schemes and the estimated quantities of the last iterations of both goal-oriented error control strategies are visualized.

Comparing the stepsizes of the last iterations of both goal-oriented error control strategies, see first row of Figure 11.3, we notice that making tiny steps at the right interval end is more important for the reduction of the error in $y(t_\mathrm{f})$ than small steps at the left end and in the middle of the interval. In the solution of this IVP example the stepsize of the scheme adaption strategy is not used to compensate the BDF order depicted in the third row of Figure 11.3. The stepsize sequence is rather influenced by the conditioning of the IVP described by the adjoints. The approximated adjoints $\widehat{\lambda}^h = \{\widehat{\lambda}_n\}_{n=0}^N$ used for the last error estimation and the local error indicators are depicted at the bottom of Figure 11.4. In huge parts of the time interval $[0, 1]$ they are zero or at least extremely small and only at a small part towards the right interval end they increase rapidly. Hence, local inaccuracies at the beginning are insignificant whereas local inaccuracies towards the right interval end are weighted heavily. This behavior can not be detected by local tolerance adaption since within this strategy stepsizes and orders are chosen by the standard mechanism based on estimated local truncation errors, penultimate row of Figure 11.3(a). The estimated local truncation errors are all below the relative tolerance $\texttt{RelTol}^2 = 8.348260 \cdot 10^{-11}$. In contrast, the last estimated local truncation errors of the scheme adaption strategy are comparably huge at the left end of $[0, 1]$, see penultimate row of Figure 11.3(b). Due to the lack of a strategy for order adaption in the indicator-based scheme adaption, the orders of the scheme adaption strategy are still that of the first integration, third row of Figure 11.3(b), whereas in the tolerance adaption strategy the order is chosen adaptively in each integration, cf. third row of Figure 11.3(a). Nevertheless, the integration with goal-oriented scheme adaption yields a more economic integration scheme due to the incorporation of adjoint information.

Having a look at the weak adjoints Λ^h from the last iterations of both error control strategies, see top row of Figure 11.4, indicates the ability of the indicator-based scheme adaption to improve also the approximation of the weak adjoint solutions. However, further investigations in this direction are left for future research.

11.3 Summary

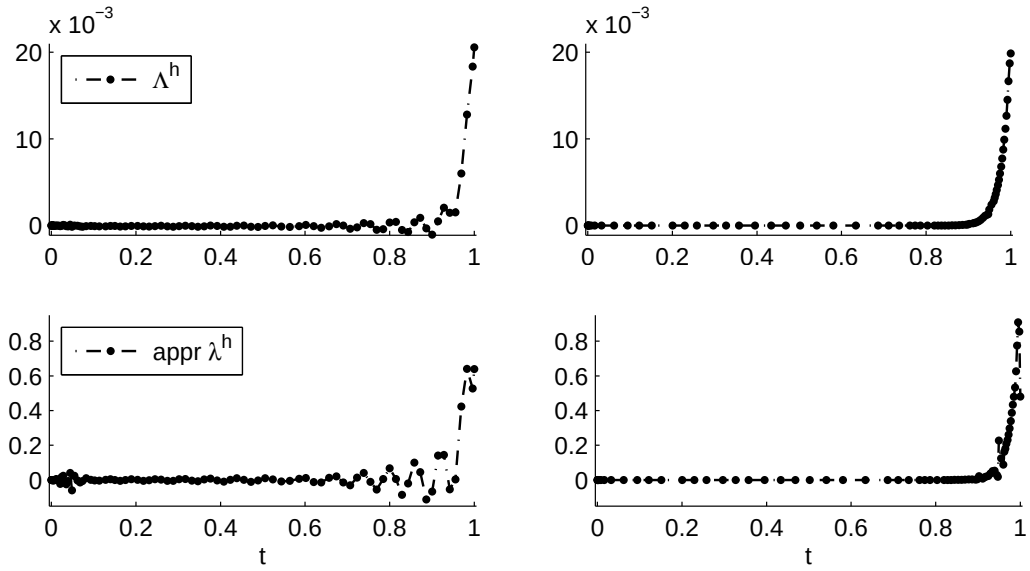
In this chapter we have investigated the usefulness of our novel goal-oriented error estimator $\tilde{\eta}$ given by (8.22) to control the nominal integration such that the goal-oriented global error of the nominal approximation is controlled. In fact, we successfully used the estimator to reduce the goal-oriented error of the approximation by adapting the local relative tolerance. This simple strategy of Algorithm 1 is



(a) Goal-oriented local tolerance adaption

(b) Goal-oriented scheme adaption

Figure 11.3: Comparison of the last iterations of both goal-oriented error control strategies applied to Example 6 with $J(y(t_\mathrm{f})) = y(t_\mathrm{f})$, `RelTol` $= 10^{-3}$, `GTol` $= 2 \cdot 10^{-10}$ and p = 0.18. Stepsizes (first row) and BDF orders (third row) of first (in gray) and last (in black) integration as well as stepsize ratios of last integration (second row) are given. The penultimate row shows the norm of the estimated local truncation errors and the last row the absolute value of the local error indicators.



(a) Goal-oriented local tolerance adaption

(b) Goal-oriented scheme adaption

Figure 11.4: FE weak adjoints Λ^h (top) and approximated discrete adjoints $\widehat{\lambda}^h = \{\widehat{\lambda}_n\}_{n=0}^N$ (bottom) of last iterations of both goal-oriented error control strategies applied to Example 6.

able to determine an appropriate local relative tolerance. Usually, there is no a priori knowledge for a suitable choice of the relative tolerance in order to obtain an approximation with bounded global error of desired size, see e.g. first integration in Section 11.1.1. Nevertheless, for subsequent integrations the relative tolerance could be adjusted appropriately. Furthermore, the local error indicators of estimator $\tilde{\eta}$ have been successfully used in indicator-based adaption of the integration scheme to give a nominal approximation with required global error bound. Although the scheme adaption of Algorithm 3 is quite simple so far, the numerical experiments already indicate the suitability of the information carried by the local error indicators and the applicability of the goal-oriented scheme adaption of Algorithm 2, see in particular Section 11.2.2. Concerning the goal-oriented error estimator $\tilde{\eta}$ itself the signed effectivity indices confirm again its good accuracy in magnitude and sign.

12 Hydrolysis of propionic anhydride in a Stirred Tank Reactor

In this chapter we focus on a real-world example from chemical engineering, a laboratory-scale discontinuous Stirred Tank Reactor (STR) carrying out the exothermic, self-accelerating hydrolysis of propionic anhydride. Firstly, we describe the chemical process, develop a new model in Ordinary Differential Equations (ODEs) using validated subcomponents and compare the simulation results to measurement data. Secondly, we investigate the weak adjoint solutions corresponding to a nonlinear criterion of interest and confirm numerically the results of Part II. Finally, we apply the goal-oriented global error control strategies developed in Chapter 9 to compute nominal approximations with guaranteed global error bounds.

12.1 General description

The operation of STRs in batch and semibatch mode is commonly used in the production of fine chemicals where only small amounts of one of numerous, highly specialized substances are produced. It allows not only the production of small amounts but also a rapid change from one reaction process to another. Unfortunately, these discontinuous reactors are prone to loss of thermal control and are more often involved in accidents than continuous reactors operating at steady states. The terms 'discontinuous' and 'continuous' refer to the operation mode of the STR: In discontinuous mode the products are completely removed from the tank after the reaction has finished whereas in continuous mode reactants are added and products are removed simultaneously. Many reaction processes for fine chemicals are heterogeneous liquid-liquid systems initiated by a catalyst. An example is the hydrolysis of propionic anhydride catalyzed by sulfuric acid. This self-accelerting reaction is strongly exothermic and can easily lead to thermal runaways. Nevertheless, it is allowed to be studied in a laboratory. For details on safety aspects of STRs we refer to Westerterp and Molga [124, 125].

Reaction

The hydrolysis of propionic anhydride $(CH_3CH_2CO)_2O$ (Ah) with water H_2O (w) to form propionic acid C_2H_5COOH (Ac) is described by the following stoichiometric equation

$$\mathrm{Ah} \quad + \quad \mathrm{H_2O} \quad \xrightarrow{H^+} \quad 2\,\mathrm{Ac}$$

catalyzed by hydrons H^+ that are provided in our case by sulfuric acid H_2SO_4 (S). This reaction has been studied by Molga and Cherbański [93, 94] and Cherbański [44]. It takes place in the aqueous phase although the solubility of propionic anhydride in water is limited. This limitation provokes that the mixture is heterogeneous with the propionic anhydride as organic phase, i.e. like oil droplets in water but the droplets outweighing the water. Generally, in such systems reaction and mass transfer occur simultaneously and the overall reaction rate is affected by the mass transfer. The reaction product propionic acid remains in the aqueous phase and increases, like sulfuric acid, the solubility of propionic anhydride. Therefore, the mixture fades to a homogeneous status. Due to the increasing mass transfer the reaction is self-accelerating. The reaction mixture is heterogeneous ('milky') as long as the propionic anhydride is not completely soluble in the aqueous phase and becomes homogeneous ('transparent') once all propionic anhydride is solved and the organic phase has disappeared.

Reactor

The reaction is carried out in a STR with cooling jacket, baffles and downward propeller stirrer. For semibatch operation, the vessel is charged with water and sulfuric acid whereas propionic anhydride is fed to the reactor for a certain time span. This discontinuous process operates far away from steady states and is characterized by strongly nonlinear dynamics with time varying coefficients. During the experiment the temperature of the cooling jacket is kept constant, called isoperibolic operation mode, and the temperature inside the reactor changes due to the heat generated by the reaction.

Concerning the optimal operation of semibatch reactors performance and safety are of great importance. Performance may mean to run the process at a minimal time or such that the products exhibit particular properties. Safety aims to reduce the risk for thermal runaways. Runaways may occur due to an accumulation of non-reacted substance or a malfunction of cooling or stirring system. A terrible experience of these effects has been the accident of 1976 in Seveso, Italy. Hence, for a safe operation the temperature rise due to a sudden reaction of unreacted substances inside the tank has to be kept bounded all the time. To achieve this usually the dosing rate of the added substance serves as control, see e.g. Kühl et al. [82, 81] and Ubrich et al. [119]. Most laboratory as well as industrial facilities of STRs allow only addition of substances at constant rates.

12.2 Modeling and simulation

In this section we develop a new mathematical model to describe the hydrolysis of propionic anhydride carried out in a semibatch STR. It is an empirical model without taking into account the reaction mechanism in all its details. We use validated

expressions for mass transport and reaction kinetics which have been investigated by Molga and Cherbański [93, 94] and Cherbański [44]. Finally, we observe that the newly composed model is able to describe experimental measurements taken during research stays at the Faculty of Chemical and Process Engineering of the Warsaw University of Technology. To perform the experiments an RC1 Mettler Toledo Reaction Calorimeter has been used. The reactants are of purity better than 97% in the case of propionic anhydride and 95% in that of sulfuric acid.

12.2.1 Mathematical modeling

Due to the stirring and the presence of the baffles the mixing can be assumed to be ideal and the mixture to be gradientfree in space. Based on this 'ideal mixing assumption' we model the process by a system of ODEs. Furthermore, we presume that the heat transfer between the phases is instanteneous. The dynamic states are the temperature of the reaction mixture and the mole numbers of each species where the propionic anhydride in the organic phase and in the aqueous phase are interpreted as different species. Thus, the resulting ODE system in five states reads

$$\dot{n}_{\mathrm{w}}(t) = -r(t) \cdot V^{\mathrm{aq}}(t) + [1 - p_{\mathrm{Ah}}] \cdot u(t)/M_{\mathrm{w}} \tag{12.1a}$$

$$\dot{T}(t) = \left[\Delta H_{\mathrm{Ah}} \cdot r(t) \cdot V^{\mathrm{aq}}(t) - q_{\mathrm{flow}}(t) - q_{\mathrm{loss}}(t) - q_{\mathrm{dos}}(t)\right] / (mC_p)_R(t) \tag{12.1b}$$

$$\dot{n}^{\mathrm{aq}}_{\mathrm{Ah}}(t) = -r(t) \cdot V^{\mathrm{aq}}(t) + Q(t) \tag{12.1c}$$

$$\dot{n}^{\mathrm{org}}_{\mathrm{Ah}}(t) = p_{\mathrm{Ah}} \cdot u(t)/M_{\mathrm{Ah}} - Q(t) \tag{12.1d}$$

$$\dot{n}_{\mathrm{Ac}}(t) = 2 \cdot r(t) \cdot V^{\mathrm{aq}}(t) \tag{12.1e}$$

where $u(t)$ [kg/s] describes the dosing rate of propionic anhydride (with purity p_{Ah}), $Q(t)$ [mol/s] the flow rate of propionic anhydride from the organic to the aqueous phase, $r(t)$ [mol/(m^3 s)] the reaction rate and $q_{\mathrm{flow}}(t)$, $q_{\mathrm{loss}}(t)$, $q_{\mathrm{dos}}(t)$ [J/s] the heat exchange with cooling jacket, surroundings and added substance, respectively. The molar mass M [kg/mol] of each substance is given in Table A.1. For simplification, we suppose that the heat capacity C_p [J/(kg K)] of each substance is constant, see Table A.1, and that the mixture has a uniform constant density $\rho = 991.014896$ [kg/m^3], i.e. that of water at a reference temperature of 313.15K. Furthermore, we assign the following purities $p_{\mathrm{Ah}} = 0.97$ and $p_{\mathrm{S}} = 0.95$.

Mass transport of propionic anhydride

The transport of propionic anhydride from the organic to the aqueous phase, i.e. from the droplets to the bulk of water containing also sulfuric acid, propionic acid and some propionic anhydride, has been investigated by Molga and Cherbański [93, 94] as well as Cherbański [44] and is modeled by

$$Q(t) = K^{\mathrm{aq}} \cdot a(t) \cdot \left\{ \widetilde{C}^{\mathrm{aq}}_{\mathrm{Ah}}(t) - C^{\mathrm{aq}}_{\mathrm{Ah}}(t) \right\} \cdot V^{\mathrm{aq}}(t).$$

The solubility of propionic anhydride in the aqueous phase depends on the concentration of propionic acid expressed as mass ratio to water as well as on the temperature of the mixture

$$\widetilde{C}_{\mathrm{Ah}}^{\mathrm{aq}}(t) = \frac{\rho}{M_{\mathrm{Ah}}} \left(U + V[T(t) - 273.15\mathrm{K}] + W \left[\frac{n_{\mathrm{Ac}}(t) M_{\mathrm{Ac}}}{n_{\mathrm{w}}(t) M_{\mathrm{w}}} \right]^{\chi} \right).$$

The parameter values listed at the left of Table 12.1 have been estimated for $T(t) \in [293.15\mathrm{K}, 313.15\mathrm{K}]$ and a mass ratio of propionic acid to water less than 0.25. The concentration $\widetilde{C}_{\mathrm{Ah}}^{\mathrm{aq}}(t)$ [mol/m^3] describes how many moles of propionic anhydride are soluble in the aqueous phase with volume $V^{\mathrm{aq}}(t)$ until it is saturated. Hence, the transport term $Q(t)$ is proportional to the difference of the concentration $\widetilde{C}_{\mathrm{Ah}}^{\mathrm{aq}}(t)$ of propionic anhydride that is soluble in aqueous phase and the concentration $C_{\mathrm{Ah}}^{\mathrm{aq}}(t) = n_{\mathrm{Ah}}^{\mathrm{aq}}(t)/V^{\mathrm{aq}}(t)$ [mol/m^3] of propionic anhydride that is available in the aqueous phase. The proportionality coefficient is composed of the overall mass transfer coefficient K^{aq} [m/s] reduced to the aqueous phase and the interfacial area $a(t)$ [m^2/m^3] per volume approximated as function in the volume fraction of the organic phase

$$a(t) = \frac{6}{d_{32}} \frac{V^{\mathrm{org}}(t)}{V^{\mathrm{aq}}(t) + V^{\mathrm{org}}(t)}$$

with Sauter mean diameter d_{32} [m] of the droplets. We take $K^{\mathrm{aq}} = 5 \cdot 10^{-4}$ and $d_{32} = 2 \cdot 10^{-4}$. The volume [m^3] of the aqueous and the organic phase are given by

$$\begin{aligned} V^{\mathrm{aq}}(t) &= \left\{ M_{\mathrm{Ah}} \cdot n_{\mathrm{Ah}}^{\mathrm{aq}}(t) + M_{\mathrm{w}} \cdot n_{\mathrm{w}}(t) + M_{\mathrm{S}} \cdot n_{\mathrm{S}} + M_{\mathrm{Ac}} \cdot n_{\mathrm{Ac}}(t) \right\} / \rho \\ V^{\mathrm{org}}(t) &= M_{\mathrm{Ah}} \cdot n_{\mathrm{Ah}}^{\mathrm{org}}(t) / \rho \end{aligned}$$

respectively, with constant number of moles n_{S} of sulfuric acid.

Reaction kinetics

The hydrolysis of propionic anhydride is a second-order reaction that takes place in the aqueous phase. Hence, its reaction rate reads

$$r(t) = k_{\mathrm{eff}}(t) \cdot C_{\mathrm{Ah}}^{\mathrm{aq}}(t) \cdot C_{\mathrm{w}}(t)$$

where $C_{\mathrm{Ah}}^{\mathrm{aq}}(t)$ and $C_{\mathrm{w}}(t) = n_{\mathrm{w}}(t)/V^{\mathrm{aq}}(t)$ are the concentrations of propionic anhydride and water in the aqueous phase, respectively. The kinetic expression for the effective reaction rate coefficient $k_{\mathrm{eff}}(t)$ has been studied by Cherbański [44]. The rate coefficient is of Arrhenius type

$$k_{\mathrm{eff}}(t) = A \cdot \exp\left(-\frac{E_a}{RT(t)} - H_R(t) \right)$$

where $R = 8.314472$ [J/(mol K)] is the universal gas constant and E_a the activation energy. The acidity function

$$H_R(t) = \left\{ BC_{\mathrm{Ac}}(t) + DC_{\mathrm{S}}(t) \right\} / T(t)$$

measures the acidity of the mixture caused by propionic acid and sulfuric acid and describes the catalyst transformation. The parameter values are listed at the right of Table 12.1.

Symbol	Value	Unit	Symbol	Value	Unit
U	0.00367	-	A	498670.82	$m^3/(mol\ s)$
V	$5.5 \cdot 10^{-4}$	1/K	E_a	78406.86	J/mol
W	0.3406	-	B	-0.934	m^3 K/mol
χ	1.751	-	D	0.0364	m^3 K/mol

Table 12.1: Model parameters for solubility (left part) and reaction kinetics (right part) of propionic anhydride.

Energy balance

The energy balance for reactor and reaction mixture yields the differential equation (12.1b) for the temperature inside the reactor. A detailed description of all heat transfers in STRs can be found e.g. in Zaldivar et al. [129]. The total heat capacity $(mC_p)_R(t)$ [J/K] of the mixture is approximated by

$$\begin{aligned}(mC_p)_R(t) = &\left\{n_{\mathrm{Ah}}^{\mathrm{aq}}(t) + n_{\mathrm{Ah}}^{\mathrm{org}}(t)\right\} \cdot M_{\mathrm{Ah}} \cdot C_{p,\mathrm{Ah}} + n_{\mathrm{w}}(t) \cdot M_{\mathrm{w}} \cdot C_{p,\mathrm{w}} \\ &+ n_{\mathrm{S}} \cdot M_{\mathrm{S}} \cdot C_{p,\mathrm{S}} + n_{\mathrm{Ac}}(t) \cdot M_{\mathrm{Ac}} \cdot C_{p,\mathrm{Ac}}\end{aligned} \tag{12.2}$$

and describes how much heat [J] is required to change the temperature. The heat exchange $q_{\mathrm{flow}}(t)$ with the heat transfer fluid of the cooling jacket is

$$q_{\mathrm{flow}}(t) = UA(t)\left\{T(t) - T_{\mathrm{j}}\right\}$$

where T_{j} [K] is the temperature of the fluid and $UA(t)$ [W/K] the heat transfer coefficient multiplied by the exchange area. A calibration before and after the reaction is performed to estimate $UA(t)$ at those times. During the reaction a linear interpolation is used

$$UA(t) = (UA_2 - UA_1)/(V_2 - V_1)(V(t) - V_1) + UA_1$$

as approximation where the volume of the whole mixture is $V(t) = V^{\mathrm{aq}}(t) + V^{\mathrm{org}}(t)$. The heat loss towards the surroundings (with ambient temperature T_{amb} [K])

$$q_{\mathrm{loss}}(t) = UA_0\left\{T(t) - T_{\mathrm{amb}}\right\}$$

depends on the transfer coefficient UA_0 [W/K] for heat losses through the top of the reactor estimated during calibration. The used values are given in Table A.2. More on the calibration procedure for the reaction calorimeter RC1 can be found in Milewska [92]. The heat absorbed by the added substance is given by

$$q_{\mathrm{dos}}(t) = \left(p_{\mathrm{Ah}} \cdot C_{p,\mathrm{Ah}} + [1 - p_{\mathrm{Ah}}] \cdot C_{p,\mathrm{w}}\right) u(t)\left\{T(t) - T_{\mathrm{dos}}\right\}$$

where T_{dos} [K] is the temperature of the dosed propionic anhydride which coincides in our experimental setup with the ambient temperature T_{amb}. The heat release due to the reaction depends on the overall conversion rate $r(t) \cdot V^{\mathrm{aq}}(t)$ and the reaction enthalpy $\Delta H_{\mathrm{Ah}} = 54885.7254$ [J/mol], cf. Cherbański [44].

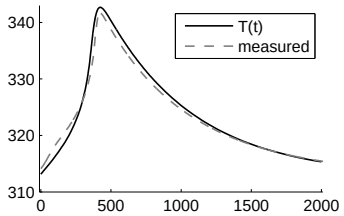


Figure 12.1: Comparison of measured temperature and simulated temperature of the IVP with ODE (12.1) and setup of Table A.3 on the time interval $[0, 2000]$ using `RelTol` $= 10^{-6}$.

Safety function

The temperature rise due to a sudden reaction of unreacted substance accumulated in the reactor is of great importance for safe operation of the process. In our case the amount of unreacted propionic anhydride $n_{\mathrm{Ah}}^{\mathrm{aq}}(t) + n_{\mathrm{Ah}}^{\mathrm{org}}(t)$ at time t is the limiting factor for the reaction. In the case of a sudden cooling failure, the process becomes adiabetic, i.e. there is nearly no heat exchange with the exterior, and the reaction of the accumulated substances accelerates quickly which may lead to dangerous situations. To avoid them, the following *safety function*

$$S(t) = T(t) + \left\{ n_{\mathrm{Ah}}^{\mathrm{aq}}(t) + n_{\mathrm{Ah}}^{\mathrm{org}}(t) \right\} \cdot \Delta H_{\mathrm{Ah}} / (mC_p)_R(t) \tag{12.3}$$

should remain bounded below a maximal temperature $T_{\max}$ during the whole reaction. Often in process optimization the safety function is added, apart from a bound on the reactor temperature $T(t)$ itself, as constraint to the Optimal Control Problem (OCP) formulation in order to run the process safely, see e.g. Kühl et al. [82, 81] and Ubrich et al. [119]. The safety function $S(t) = S(t, \boldsymbol{y}(t))$ itself is nonlinear in the states $\boldsymbol{y}(t) = [n_{\mathrm{w}}(t),\, T(t),\, n_{\mathrm{Ah}}^{\mathrm{aq}}(t),\, n_{\mathrm{Ah}}^{\mathrm{org}}(t),\, n_{\mathrm{Ac}}(t)]^{\intercal}$ of the Initial Value Problem (IVP) system due to the division by $(mC_p)_R(t)$ given by (12.2).

12.2.2 Simulation of experiments

The setup of a particular experiment defines the initial values and the experimental parameters for the ODE (12.1). Our model is capable to describe the batch experiment of Table A.3 where the propionic anhydride is added all at once. The measured and simulated temperature profiles are depicted in Figure 12.1. This is a typical temperature profile of a thermal runaway: Suddenly the temperature inside the reactor rises drastically which is at the one hand caused by an increasing reaction rate of accumulated substances and on the other hand accelerates the reaction as well, cf. Westerterp and Molga [124]. The temperature only declines if most of the substances has reacted. One can imagine the dangerous situation if such a temperature explosion would take place in a huge industrial STR.

Symbol	Value	Unit	Symbol	Value	Unit
$n_{\mathrm{w}}(t_{\mathrm{s}})$	$(1.02 + (1 - p_{\mathrm{S}})0.071)/M_{\mathrm{Ah}}$	mol	T_{j}	313.15	K
$T(t_{\mathrm{s}})$	313.15	K	T_{amb}	296.15	K
$n_{\mathrm{Ah}}^{\mathrm{aq}}(t_{\mathrm{s}})$	0	mol	u_{d}	0.4/1000	kg/s
$n_{\mathrm{Ah}}^{\mathrm{org}}(t_{\mathrm{s}})$	0	mol	t_{d}	1000	s
$n_{\mathrm{Ac}}(t_{\mathrm{s}})$	0	mol	n_{S}	$p_{\mathrm{S}} \cdot 0.071/M_{\mathrm{S}}$	mol

Table 12.2: Initial values and experimental parameters of the semibatch experiment with a dosing time of 1000s and a propionic anhydride amount of 0.4kg.

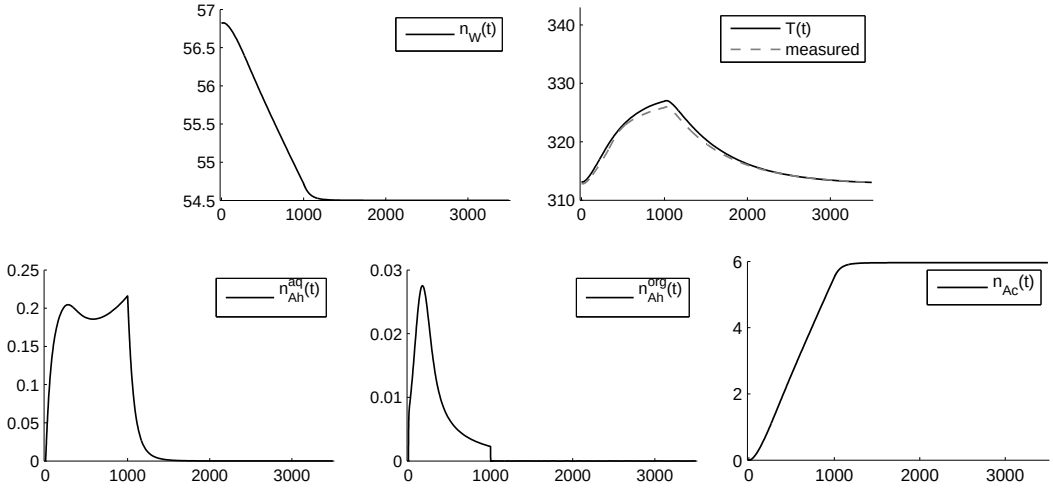


Figure 12.2: Simulation results of the IVP with ODE (12.1) and setup of Table 12.2 on the time interval $[0, 3500]$ using `RelTol` $= 10^{-6}$. Comparison of measured and simulated temperature at the upper right corner.

We furthermore perform a semibatch experiment where the propionic anhydride is fed within 1000s. After all substance is added, the system can be understood to operate in batch mode until the reaction is finished. The experimental setup is described in Table 12.2. The simulation results and the experimental measurements of this semibatch process are depicted in Figure 12.2. The maximal temperature inside the reactor is reduced compared to the batch experiment and the temperature profile is smooth which indicates a safe operation, cf. Westerterp and Molga [124]. Nevertheless, for four-fifth of the product amount much more time was needed.

We also used the derived model to plan an experiment where the amount of product should be doubled while the reaction should be fast and the operation safe. Performing some simulations we agreed on using a dosing time of 2000s. The proposed experimental setup is given in Table A.4. The simulated and measured temperature profiles are displayed in Figure 12.3. In fact, the overall progress of the reaction exhibits the desired properties of a quick onset, a fair conversion and

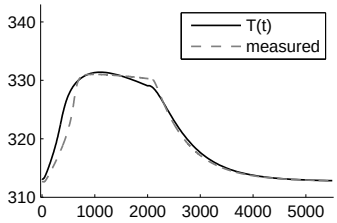


Figure 12.3: Comparison of measured temperature and simulated temperature of the IVP with ODE (12.1) and setup of Table A.4 on the time interval $[0, 5500]$ using $\texttt{RelTol} = 10^{-6}$.

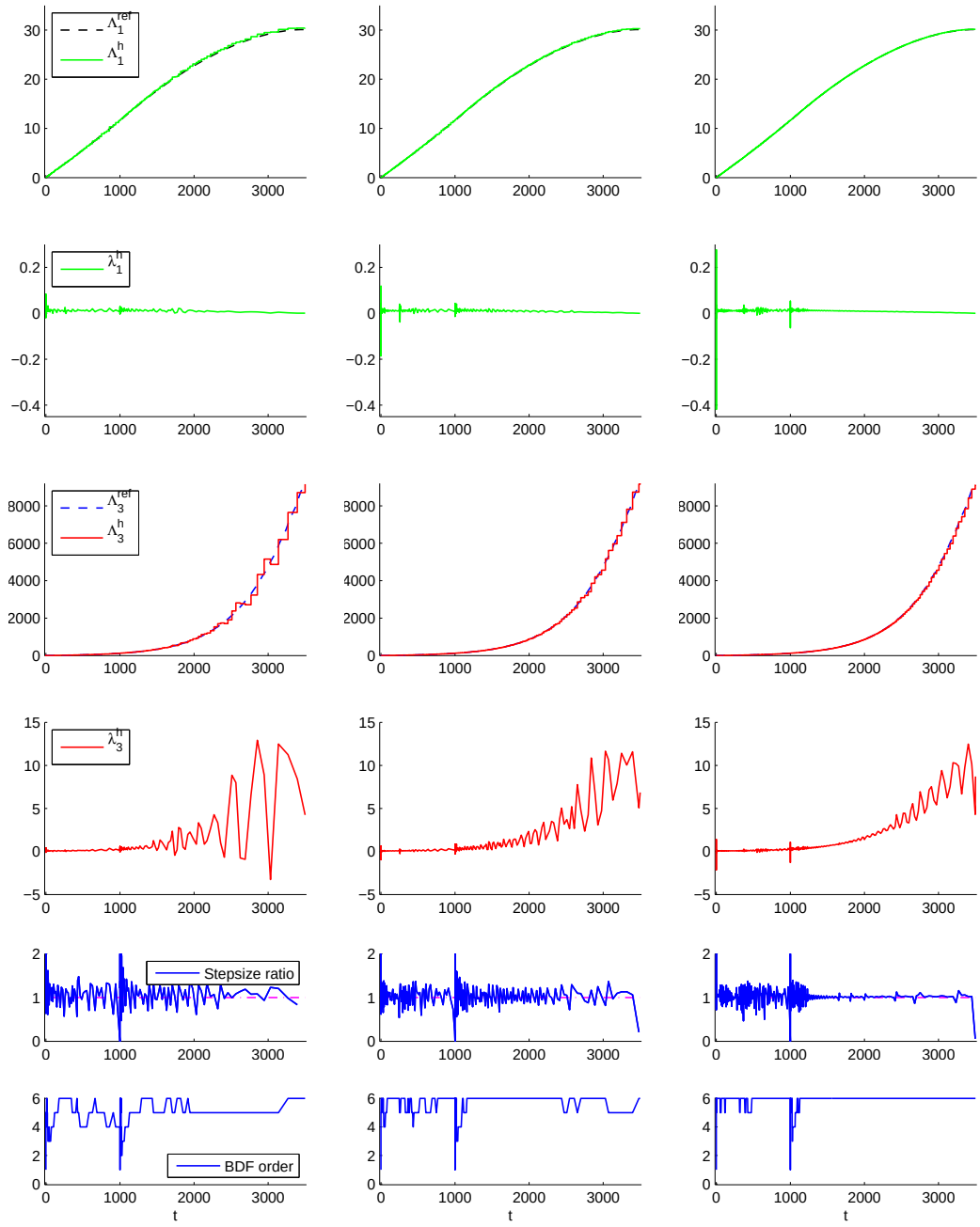
a smooth temperature profile, i.e. it is a so-called QFS reaction, cf. Westerterp and Molga [124].

We now focus again on the achievements of Part II and III of this thesis. To this end, the IVP model of the semibatch experiment with 1000s as dosing time serves as highly nonlinear real-world test case. Hence, subsequently we focus on the IVP with ODE (12.1) and setup of Table 12.2 and use the safety function $S(t)$ defined by (12.3) and (12.2) as nonlinear criterion of interest $J(\boldsymbol{y}(t_\mathrm{f})) = S(t_\mathrm{f})$.

12.3 Computation of weak adjoints

In semibatch STRs usually the addition of substances is done at a constant rate due to the available facilities, cf. Section 12.1. The resulting piecewise constant dosing rates cause discontinuities, also called switches, in the right hand side of the ODE. For the particular semibatch process modeled by (12.1) and the setup of Table 12.2 the switching time $t_\mathrm{d} = 1000\mathrm{s}$ is explicitly known. In our setting, the solution trajectories as well as the (forward and adjoint) sensitivities with respect to initial values are continuous at t_d but not differentiable with respect to time, cf. Bock [31]. Since a polynomial of higher order, like the Backward Differentiation Formula (BDF) polynomials of Chapter 2, cannot be used across a kink in the trajectories, the BDF integration has to be stopped and re-started at t_d. We compute the Finite Element (FE) weak adjoints of the semibatch IVP and the safety function given by (12.3) as criterion of interest, i.e. $J(\boldsymbol{y}(t_\mathrm{f})) = S(t_\mathrm{f})$. Exemplarily, the first and the third adjoints, i.e. the derivatives of J with respect to the discrete approximations of the reactants $n_\mathrm{w}(t)$ and $n_\mathrm{Ah}^\mathrm{aq}(t)$, are depicted in Figure 12.4 for nominal integrations with the three decreasing relative tolerances $\texttt{RelTol} = 10^{-4},\ 10^{-6},\ 10^{-8}$. As reference weak adjoints $\boldsymbol{\Lambda}^\mathrm{ref}(t)$ we use those of a more accurate nominal integration with $\texttt{RelTol} = 10^{-9}$. The derivatives of $J(\boldsymbol{y}(t_\mathrm{f}))$ with respect to the reaction product $n_\mathrm{Ac}(t)$ exhibit the same shape as that with respect to $n_\mathrm{w}(t)$ whereas those with respect to $T(t)$ and $n_\mathrm{Ah}^\mathrm{org}(t)$ are similar to that of $n_\mathrm{Ah}^\mathrm{aq}(t)$.

Again the discrete adjoint Internal Numerical Differentiation (IND) values $\boldsymbol{\lambda}^h = \{\boldsymbol{\lambda}_n\}_{n=0}^N$ show huge oscillations, see second and fourth row of Figure 12.4. This



(a) `RelTol` $= 10^{-4}$ (b) `RelTol` $= 10^{-6}$ (c) `RelTol` $= 10^{-8}$

Figure 12.4: Results for the variable BDF-type method `DAESOL-II` applied to the IVP with ODE (12.1), setup of Table 12.2, time interval $[0, 3500]$ and $J(\boldsymbol{y}(t_\mathrm{f})) = S(t_\mathrm{f})$. First and the third FE weak adjoints $\boldsymbol{\Lambda}^h$ compared to the reference weak adjoints $\boldsymbol{\Lambda}^{\mathrm{ref}}$ (first and third row) and corresponding discrete adjoints $\boldsymbol{\lambda}^h$ (second and fourth row). Stepsize ratio (penultimate row) and BDF order (last row) of the integration scheme are also depicted.

is due to the inconsistency of the adjoint IND schemes with the adjoint IVP, cf. Section 3.6. Nevertheless, using them in (6.2) to obtain FE approximations $\mathbf{\Lambda}^h$ of the unknown exact weak adjoints $\mathbf{\Lambda}$, a smooth behavior is observed once again, see first and third row of Figure 12.4. The smoothness of $\mathbf{\Lambda}^h$ also appears in areas of variable stepsize and variable order, see last two rows of Figure 12.4, as well as in conjunction with efficient Newton-type methods and iterative adjoint IND as used to full capacity by `DAESOL-II`. Moreover, for decreasing relative tolerances the FE approximations $\mathbf{\Lambda}^h$ approach the reference weak adjoints $\mathbf{\Lambda}^{\mathrm{ref}}$ on the whole time interval. These observations for a real-world example coincide with the results of Section 10.1.2 for an academic test case. Furthermore, the suitability of the novel functional-analytic framework and the Petrov-Galerkin FE interpretation of BDF methods and their adjoint IND schemes developed in Part II of this thesis are confirmed again numerically with the help of a challenging real-world IVP from chemical engineering.

From the nominal integrations above and further integrations with $\texttt{RelTol} = 10^{-3}, 10^{-5}, 10^{-7}$ we have been able to obtain reference solutions for the trajectory values at t_{d} and t_{f}, see Section A.3.3. They will be used in the next section to quantify the results from goal-oriented global error control.

12.4 Goal-oriented global error control

In this section we consider the global error of (12.1) with setup from Table 12.2 in the safety function defined by (12.3), i.e. in $J(\boldsymbol{y}(t_{\mathrm{f}})) = S(t_{\mathrm{f}})$. We aim to reduce the goal-oriented error below $\texttt{GTol} = 10^{-6}$ using the goal-oriented global error control strategies proposed in Chapter 9. As goal-oriented global error estimator we use exemplarily the estimator $\tilde{\eta}$ given in (8.22) which is based on the estimated local truncation errors as nominal error quantities. Furthermore, we suppose again that the residual term $\eta_{\boldsymbol{\delta}}$ is insignificant. For the first integration of both strategies, i.e. the goal-oriented local tolerance adaption of Algorithm 1 and the goal-oriented scheme adaption of Algorithm 2 together with the indicator-based scheme adaption of Algorithm 3, we use the relative tolerance $\texttt{RelTol} = 5 \cdot 10^{-4}$. For Algorithm 1 we set $c_{\mathrm{red}} = 0.5$ and for Algorithm 3 we use $\mathrm{p} = 0.08$. According to Section 10.4 we set again $c = 1$ for the termination criterion (9.1) since $\tilde{\eta}$ neither inclines to underestimate nor to overestimate the true goal-oriented global error. Hence, the goal-oriented error estimate has to satisfy $|\tilde{\eta}| \leq c \cdot \texttt{GTol} = 10^{-6}$. To investigate also the goal-oriented global error estimator $\tilde{\eta}$ itself we compare the estimated value to the goal-oriented difference $J(\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{f}})) - J(\boldsymbol{y}_N)$ where $\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{f}})$ is the reference solution explained in Section 12.3 and written down in Section A.3.3. In the first subsection we consider the simulation of the whole reaction whereas in the second subsection we simulate only the semibatch part of the experiment, i.e. the time interval on which propionic anhydride is dosed through the reactor.

j	RelTol	$\tilde{\eta}$	$J(\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{f}}))-J(\boldsymbol{y}_N)$	$I_{\mathrm{eff}}^{\mathrm{s}}$	N	$\sum s_n$	reb/dec
0	5.000000e-04	6.647517e-05	6.200526e-05	1.073	247	489	17 / 36
1	7.521605e-06	-6.573493e-07	-7.781984e-07	0.844	381	800	18 / 39
0	5.000000e-04	6.647517e-05	6.200526e-05	1.073	247	489	17 / 36
1		4.429414e-06	4.497608e-06	0.984	261	525	10 / 39
2		7.619366e-07	2.552993e-07	2.985	277	544	11 / 42

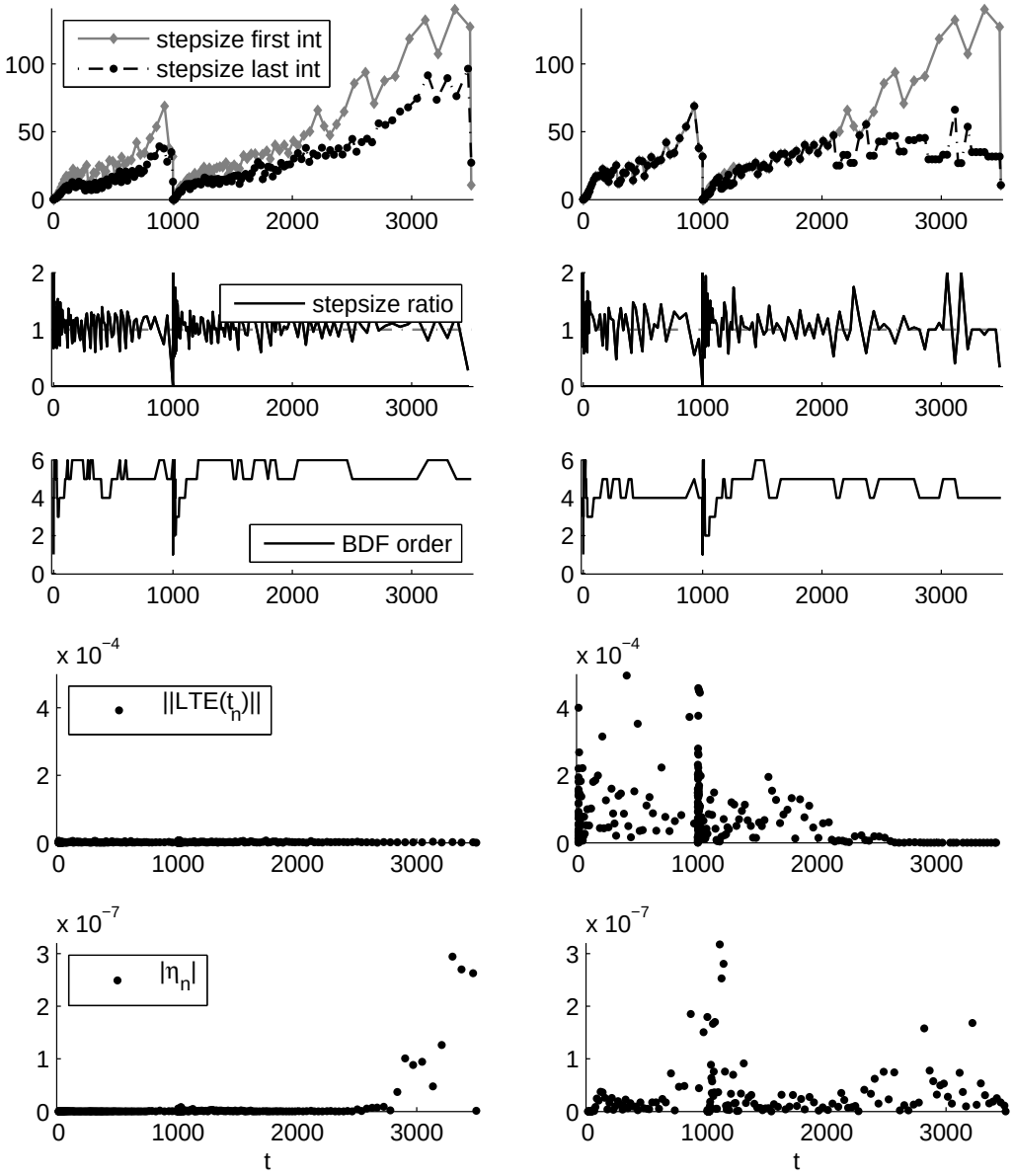
Table 12.3: Results of the goal-oriented local tolerance adaption (first part) and the goal-oriented scheme adaption with the indicator-based scheme adaption (second part) applied to the IVP with ODE (12.1), setup of Table 12.2, time interval $[0, 3500]$, $J(\boldsymbol{y}(t_{\mathrm{f}})) = S(t_{\mathrm{f}})$, $\texttt{RelTol} = 5 \cdot 10^{-4}$, $\texttt{GTol} = 10^{-6}$ and $\mathrm{p} = 0.08$.

12.4.1 Global error controlled simulation of the whole reaction

The results obtained by the goal-oriented local tolerance adaption are listed in the upper part of Table 12.3. The first integration with $\texttt{RelTol}^0 = 5 \cdot 10^{-4}$ yields an approximation with $6.2 \cdot 10^{-5}$ as goal-oriented error computed using the reference solution $\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{f}}) = \boldsymbol{y}^{\mathrm{r}}(3500)$ of Section A.3.3. Hence, local inaccuracies are damped out during the integration at least in the nonlinear safety function used here as criterion of interest. For subsequent integrations the relative tolerance is adapted based on the error estimate and the termination criterion $c \cdot \texttt{GTol}$, cf. Section 9.1. After two iterations the estimated global error is below the required tolerance $c \cdot \texttt{GTol}$. The goal-oriented errors computed with the reference solution, see fourth column of Table 12.3, confirm the suitability of the tolerance adaption strategy. The last three columns of Table 12.3 give an impression of the computational effort caused by the standard stepsize and order selection with monitor strategy, cf. Section 2.4. Furthermore, the signed effectivity indices $I_{\mathrm{eff}}^{\mathrm{s}}$ computed with reference solution and listed in the fifth column of Table 12.3 indicate once again the good accuracy in sign and magnitude of our novel estimator $\tilde{\eta}$ also for variable BDF-type methods applied to a real-world IVP with nonlinear criterion of interest.

The second part of Table 12.3 displays the results obtained by goal-oriented scheme adaption together with indicator-based adaption of the scheme. Within three iterations also this strategy has been successful in reducing the goal-oriented error estimate below $c \cdot \texttt{GTol} = 10^{-6}$. Its suitability is again confirmed by the goal-oriented errors computed with the reference solution. Overall, the last integration of the scheme adaption strategy is more efficient than that of the local tolerance adaption strategy. The former required only three-fourths of the integration steps of the latter, around two-thirds of the Newton iterations and nearly half of the costly matrix rebuilds while the number of matrix decompositions increased slightly.

In Figure 12.5 the integration schemes and the estimated error quantities of the last iterations of both goal-oriented error control strategies are depicted, respectively.



(a) Goal-oriented local tolerance adaption

(b) Goal-oriented scheme adaption

Figure 12.5: Comparison of the last iterations of both goal-oriented error control strategies applied to the IVP with ODE (12.1), setup of Table 12.2, time interval $[0, 3500]$, $J(\boldsymbol{y}(t_\mathrm{f})) = S(t_\mathrm{f})$, `RelTol` $= 5 \cdot 10^{-4}$, `GTol` $= 10^{-6}$ and $\mathrm{p} = 0.08$. Stepsizes (first row) of first (in gray) and last (in black) integration, stepsize ratios (second row) and BDF orders (third row) of last integration are given. The penultimate row shows the norm of the estimated local truncation errors and the last row the absolute value of the local error indicators.

The last integration of the local tolerance adaption uses smaller stepsizes on the whole time interval than its first integration, see first row of Figure 12.5. In the last integration of the scheme adaption strategy only the steps at the second half of the time interval $[0, 3500]$ are downsized. The reduction of these integration steps is caused by the local error indicators which incorporate adjoint information. The indicators of both last integrations are depicted at the bottom of Figure 12.5. According to the left plot also for the integration with adapted relative tolerance the biggest contributions to the error in J come from the last integration steps. The approximated weak adjoints look similar, but not identical, to the FE weak adjoints $\mathbf{\Lambda}^h$ displayed in Figure 12.4 for other integration schemes. Since the gradients of the weak adjoints are very small at the first part of the time interval, the contribution of the comparably big estimated local truncation errors (penultimate row of Figure 12.5(b)) of that part on the goal-oriented error is small and a reduction of these integration steps is not necessary. However, this behavior could not be detected by the tolerance adaption strategy using the standard selection mechanism for stepsize and order based only on the estimated local truncation errors depicted in the penultimate row of Figure 12.5(a). In fact, all these estimated local truncation errors are below the required tolerance $\texttt{RelTol}^1 = 7.521605 \cdot 10^{-6}$.

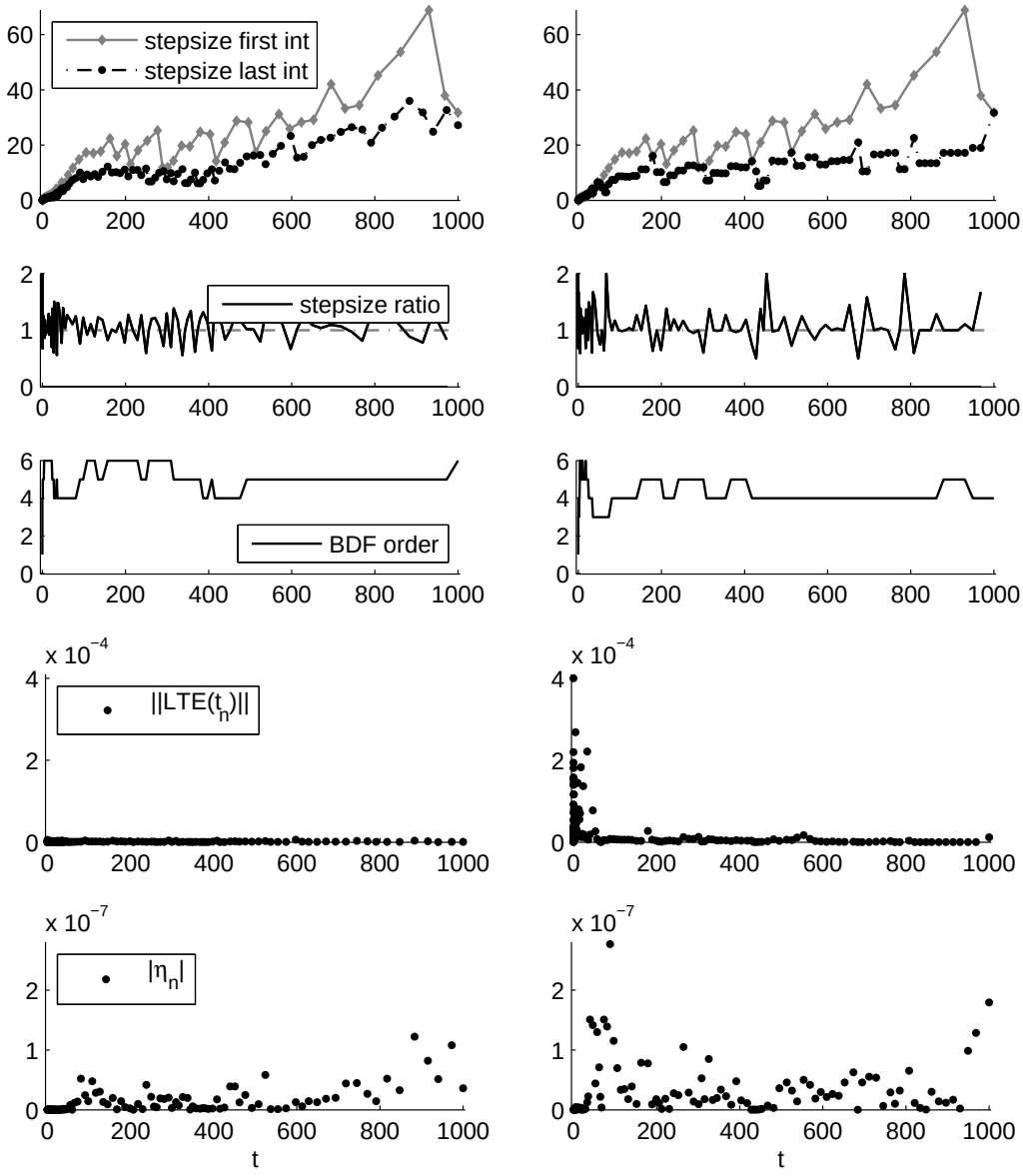
In summary, the error controlled simulations of the whole hydrolysis reaction on $[0, 3500]$ with a dosing time of 1000s exhibit an analogous behavior as observed for the academic test case with analytic solutions provided by Example 6, see Section 11.2.2.

12.4.2 Global error controlled simulation of the semibatch part

In this section we have a look at the global error controlled simulation with the same accuracy requirements as in Section 12.4.1 but only of the first part of the reaction, i.e. of that part until the switch in the ODE right hand side occurs due to the termination of dosing, cf. Section 12.3. The results of the goal-oriented tolerance adaption and the goal-oriented scheme adaption together with the indicator-based scheme adaption are listed in the first and the second part of Table 12.4, respectively.

On the integration interval $[0, 1000]$ of the dosing time the last integrations of both goal-oriented adaption strategies do not differ much in terms of computational effort, see Table 12.4. The last integration of the scheme adaption strategy is only slightly more efficient than that of the local tolerance adaption. The integration schemes and the estimated error quantities of both last iterations are depicted in Figure 12.6, respectively.

Compared to the stepsize sequence of the first integration which is the same for both adaption strategies, displayed in gray at the top of Figure 12.6, both strategies refine the stepsizes over the whole time interval, depicted in black. The stepsizes of the last iteration of the tolerance adaption strategy are slightly smaller at the beginning and some coarser at the end of $[0, 1000]$ compared to those of the scheme adaption strategy. In this case, the local conditioning of the IVP is not crucial for



(a) Goal-oriented local tolerance adaption

(b) Goal-oriented scheme adaption

Figure 12.6: Comparison of the last iterations of both goal-oriented error control strategies applied to the IVP with ODE (12.1), setup of Table 12.2, time interval $[0, 1000]$, $J(\boldsymbol{y}(t_\mathrm{f})) = S(t_\mathrm{f})$, $\texttt{RelTol} = 5 \cdot 10^{-4}$, $\texttt{GTol} = 10^{-6}$ and $\mathrm{p} = 0.08$. Stepsizes (first row) of first (in gray) and last (in black) integration, stepsize ratios (second row) and BDF orders (third row) of last integration are given. The penultimate row shows the norm of the estimated local truncation errors and the last row the absolute value of the local error indicators.

j	RelTol	$\tilde{\eta}$	$J(\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{f}}))-J(\boldsymbol{y}_N)$	$I_{\mathrm{eff}}^{\mathrm{s}}$	N	$\sum s_n$	reb/dec
0	5.000000e-04	-3.722629e-05	-3.546574e-05	1.050	104	187	12 / 22
1	1.343137e-05	1.581686e-06	8.809566e-07	1.796	137	267	12 / 16
2	6.715684e-06	5.559073e-07	4.866295e-07	1.143	156	294	13 / 23
0	5.000000e-04	-3.722629e-05	-3.546574e-05	1.050	104	187	12 / 22
1		-5.413436e-06	-2.400686e-06	2.255	115	216	7 / 28
2		-3.344130e-06	-2.547972e-06	1.313	132	263	7 / 25
3		-3.936634e-07	-2.016050e-07	1.953	145	290	5 / 25

Table 12.4: Results of the goal-oriented local tolerance adaption (first part) and the goal-oriented scheme adaption with the indicator-based scheme adaption (second part) applied to the IVP with ODE (12.1), setup of Table 12.2, time interval $[0, 1000]$, $J(\boldsymbol{y}(t_{\mathrm{f}})) = S(t_{\mathrm{f}})$, $\texttt{RelTol} = 5 \cdot 10^{-4}$, $\texttt{GTol} = 10^{-6}$ and $\mathrm{p} = 0.08$.

an efficient integration and the goal-oriented local tolerance adaption already yields a good result as in the academic test case provided by Example 3, see Section 11.2.1.

12.5 Summary

In the first part of this chapter we have constituted a dynamic model for the hydrolysis of propionic anhydride in a tank reactor. This is a representative for a wide class of fine chemical reactions mostly carried out in STRs due to their high specializations and small production amounts. A phenomenological comparison to experimental measurements indicate the ability of the model to describe real laboratory experiments.

In the second part we confirmed again that the FE approximations based on the discrete adjoint IND values serve as proper quantities to approximate the weak adjoints also in the case of fully variable BDF-type methods applied to a challenging real-world IVP. This furthers the results of Section 10.1.2 on academic IVP test cases.

In the last part of this chapter we have been able to confirm the results of Chapter 11 obtained with academic test cases also in the context of a real-world IVP example. It turned out that for the error controlled integration of the semibatch IVP on the time interval where propionic anhydride is fed to the reactor the goal-oriented local tolerance adaption already gave an efficient integration scheme. For the integration of the whole reaction the goal-oriented scheme adaption gave a more efficient scheme.

Furthermore, investigating the accuracy of our novel goal-oriented error estimator $\tilde{\eta}$ with the help of the signed effectivity index and the reference solution indicates the good accuracy of $\tilde{\eta}$ in magnitude and sign for variable BDF-type methods and real-world IVPs. All effectivity indices lie in $[0.844, 2.985]$ which is a good result

in the context of fully variable BDF-type methods. Due to the correctness in sign the estimator $\tilde{\eta}$ could be used within the context of OCPs to decide whether an inequality constraint in the unknown true IVP solution $\boldsymbol{y}(t_\mathrm{f})$ is fulfilled or not. Considering, for example, the safety constraint $c(t, \boldsymbol{y}(t)) := S(t, \boldsymbol{y}(t)) - T_\mathrm{max} \leq 0$ of Section 12.2.1 and estimating the global error in $c(t_\mathrm{f}, \boldsymbol{y}(t_\mathrm{f}))$ by $\tilde{\eta}$, then one may decide by evaluating $c(t_\mathrm{f}, \boldsymbol{y}(t_\mathrm{f})) = c(t_\mathrm{f}, \boldsymbol{y}^h(t_\mathrm{f})) + \tilde{\eta}$ whether the unknown true solution $\boldsymbol{y}(t_\mathrm{f})$ is inside or outside the feasible region of the OCP.

Conclusions and perspectives

In this thesis we have developed a novel functional-analytic framework for Initial Value Problems (IVPs) in Ordinary Differential Equations (ODEs) in Banach spaces. With the proposed Petrov-Galerkin Finite Element (FE) discretization the discrete adjoints computed by adjoint Internal Numerical Differentiation (IND) of multistep Backward Differentiation Formula (BDF) methods have been related to the solution of the classical adjoint IVP via weak adjoint solutions. Using this bridge between BDF methods and Petrov-Galerkin FE methods together with the well-established Dual Weighted Residual (DWR) methodology we have derived novel goal-oriented global error estimators for BDF methods using adjoint IND. In fact, our novel error estimators are superior to a corresponding existing one and have been successfully used to compute global error controlled approximations to IVP solutions, also for a real-world example from chemical engineering which we have modeled during research stays in Warsaw.

The achievements of this thesis give inspirations for future research directions. Concerning the goal-oriented global error estimators, the most evident of them are

- the functional-analytic interpretation of the implicit correction term (8.19) for the goal-oriented error approximations $\bar{E}(\boldsymbol{y}^h)$ and $\hat{E}(\boldsymbol{y}^h)$.
- the numerical and theoretical investigation if the adaption of the integration scheme based on local error indicators also improves the FE approximation of the weak adjoints.
- the reduction of the computational cost for the approximation of the defect integrals in the goal-oriented estimator $\bar{\eta}$ by specially tailored numerical quadrature formulas.
- a strategy for the indicator-based scheme adaption to adapt also the orders of the BDF scheme according to the local error indicators.
- a strategy to utilize the residual term $\eta_{\boldsymbol{\delta}}$ and its indicators to suitably choose the termination tolerance for the Newton-type method used to solve the nonlinear BDF equations.
- an approach to make the novel goal-oriented global error estimators and in particular their weights based on adjoints accessible within the standard stepsize and order selection of a subsequent integration.

Concerning the practical application of the goal-oriented global error information, the most evident future research directions include

- the utilization of the information gain by the global error estimators in the solution of Optimal Control Problems (OCPs) by integrator-based methods. During optimization the integration accuracy should be chosen adaptively according to the distance to the optimum in order to increase the overall accuracy while the computational effort is reduced. Moreover, having an efficient and accurate integration scheme it should be reused for several optimization iterations to increase, for example, the accuracy of low rank updates in quasi-Newton methods.
- the choice of suitable criterions of interest specific to the particular applications of the novel goal-oriented error estimators.

Concerning the real-world example, the most evident issue for future research is

- the usage of the ODE model in the context of optimal control and nonlinear model predictive control to optimize the hydrolysis of propionic anhydride with regard to performance and safety.

Acknowledgments

At this place I wish to express my gratitude to my advisor Prof. Dr. Dr. h.c. Hans Georg Bock for giving me the possibility to do research on this exciting, wide-ranging subject in such a friendly and cooperative atmosphere as present in his Simulation & Optimization group at the Faculty of Mathematics and Computer Sciences and the Interdisciplinary Center for Scientific Computing (IWR) of the Heidelberg University. Hence, my thanks also go to the members of the group and in particular to Dr. Johannes P. Schlöder whom I like to thank for numerous discussions and his helpful advices in general science-related questions. I especially appreciate their support in presenting my research at international workshops and conferences in Darmstadt, Hanoi, München, Pavia and Toulouse.

My sincere thanks are addressed to Prof. Dr. Leon Gradoń of the Faculty of Chemical and Process Engineering (ICHIP) of the Warsaw University of Technology and Prof. Dr. Marek Niezgódka of the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) of the University of Warsaw for the opportunity to do research in their groups. For the very good organization of my stays in Warsaw I like to thank Dr. Anna Trykozko of ICM. Moreover, I am most grateful to Prof. Dr. Eugeniusz Molga and Dr. Michał Lewak of ICHIP for their fascinating practical introduction in laboratory research and the opportunity to mathematically model real-world processes.

Furthermore, I like to thank Prof. Dr. Dr. h.c. Rolf Rannacher heading the Numerical Analysis group of the Heidelberg University for his scientific support and the members of his group Dr. Dominik Meidner, Prof. Dr. Thomas Richter and Dr. Thomas Wick for numerous fruitful discussions.

Several colleagues from the Simulation & Optimization group have contributed significantly to the progress and success of this work. Among them, I like to especially thank Dr. Mario S. Mommer, as well as Dr. Christian Kirches and Dr. Andreas Potschka. Moreover, I like to express my gratitude to Leonard Wisching who has been always open for discussions, has asked numerous key questions and has given valuable suggestions concerning my work and this thesis. For proofreading parts of the thesis I am indebted to Dr. Falk Hante, Simon Lenz, Mario S. Mommer, Johannes P. Schlöder and Leonard Wirsching.

I am especially grateful to Dr. Jan Albersmeyer, Christian Hoffmann, Christian Kirches, Peter Kühl, Simon Lenz, Andreas Potschka, Prof. Dr. Sebastian Sager and Leonard Wirsching for extensive scientific as well as non-scientific discussions during the last years. My sincere thanks go to Jan Albersmeyer for the `DAESOL-II`

implementation and to Christian Kirches, Andreas Potschka, Andreas Schmidt and Leonard Wirsching for our joint effort in maintaining `DAESOL-II` after Jan's departure. My special thanks also go to Margret Rothfuß and Thomas Klöpfer for their kind support in organizational and technical matters. Furthermore, I cordially thank Holger Diedam, Kathrin Hatz, Dennis Janka and Andreas Sommer for our successful and enjoyable teamwork during the foundation and management of the Heidelberg Chapter of SIAM (Society for Industrial and Applied Mathematics).

Scientific and financial support by the Heidelberg University, the IWR, the International Graduiertenkolleg 710 "Complex processes: Modeling, Simulation and Optimization" at IWR Heidelberg and ICM Warsaw, the Graduate School 220 "Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences" established in the course of the German Excellence Initiative, the NOVOEXP project in the programme "Mathematik für Innovationen in Industrie und Dienstleistungen" of the German Federal Ministry of Education and Research, the SBCancer network in the Helmholtz Alliance on Systems Biology, and the project "Embedded Optimization for Resource Constrained Platforms" of the European Commission is gratefully acknowledged.

I am deeply grateful to Vera and Nina for countless hours that brightened up even the most challenging time and to Christiane for the enduring friendship over years. It is my desire to warmly thank Felix for the feeling of security he is giving me and for his tranquility that helped me so much in the last year. Last but not least I like to thank my parents and my siblings for their never-ending support and the knowledge that they are always on my side.

Thanks to all of you!

A Appendix

A.1 Useful definitions and theorems

For the reader's convenience we recall here some frequently used definitions and theorems.

Definition A.1 (Landau symbol $\mathcal{O}$) *If there exists $c > 0$ such that for two functions f and g holds*

$$\lim_{h \to 0} \left| \frac{f(h)}{g(h)} \right| < c,$$

we write $f(h) = \mathcal{O}(g(h))$.

Occasionally, we use the symbol " $\doteq$ " to indicate that a function f is approximated by a function g up to first order in $x - x_0$, i.e. $f(x) \doteq g(x)$ means

$$f(x) = g(x) + \mathcal{O}(|x - x_0|) \text{ for } x \to x_0.$$

Theorem A.2 (Interpolation/Extrapolation error) *Let $\mathcal{P}(t; t_0, \ldots, t_k)$ be the interpolation polynomial through $g(t_0), \ldots, g(t_k)$ evaluated at t with $t_j \neq t_i$ for $i, j = 0, \ldots, k$. If the function g is $(k+1)$-times differentiable, then the error of the polynomial at t is*

$$g(t) - \mathcal{P}(t; t_0, \ldots, t_k) = \prod_{j=0}^{k} (t - t_j)\, \nabla^{k+1}[g(t), g(t_k), \ldots, g(t_0)].$$

Proof *See Stoer and Bulirsch [116].* □

Definition A.3 *A function f is called to be* continuous from the left *at t if*

$$\lim_{\varepsilon \searrow 0} f(t - \varepsilon) = f(t)$$

and continuous from the right *at t if*

$$\lim_{\varepsilon \searrow 0} f(t + \varepsilon) = f(t).$$

Theorem A.4 (Neumann series) *If $\|\boldsymbol{T}\| < 1$, then the matrix $\boldsymbol{A} := \boldsymbol{I} - \boldsymbol{T}$ is nonsingular and its inverse $\boldsymbol{A}^{-1}$ is given by the Neumann series*

$$\boldsymbol{A}^{-1} = (\boldsymbol{I} - \boldsymbol{T})^{-1} = \sum_{j=0}^{\infty} \boldsymbol{T}^j = \boldsymbol{I} + \boldsymbol{T} + \sum_{j=2}^{\infty} \boldsymbol{T}^j.$$

Proof *See Werner [123].* □

A.2 Additional proofs

A.2.1 Proofs of Lemma 3.1, 3.2 and 3.3

This section contains the technical details corresponding to Section 3.4. As described there, the form of the adjoint Internal Numerical Differentiation (IND) scheme depends on the representation of the Backward Differentiation Formula (BDF) method itself. We first consider the BDF method in its standard formulation (2.2).

Proof (of Lemma 3.1) *First, all terms in the BDF method* (2.2) *are written on the left hand side to give root finding formulations. Then, each equation is multiplied by an arbitrary prefactor* $\boldsymbol{\lambda}_n^\intercal$ *and added to result in a variational formulation of the BDF method*

$$0 = -\boldsymbol{\lambda}_0^\intercal (\boldsymbol{y}_0 - \boldsymbol{y}_\mathrm{s}) - \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^\intercal \left(\sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{y}_{n+1-i} - h_n \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1}) \right). \tag{A.1}$$

A variation $\boldsymbol{w}_\mathrm{s}$ *in the initial value, i.e. a differentiation with respect to* $\boldsymbol{y}_\mathrm{s}$ *in direction* $\boldsymbol{w}_\mathrm{s}$*, gives*

$$0 = -\boldsymbol{\lambda}_0^\intercal (\boldsymbol{w}_0 - \boldsymbol{w}_\mathrm{s}) - \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^\intercal \left(\sum_{i=0}^{k_n} \alpha_i^{(n)} \boldsymbol{w}_{n+1-i} - h_n \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{y}_{n+1}) \boldsymbol{w}_{n+1} \right). \tag{A.2}$$

This equation is now rearranged according to the variations $\boldsymbol{w}_n$ *in the discrete solutions* $\boldsymbol{y}_n$*. We use the convention that* $\alpha_i^{(n)} = 0$ *for* $i > k_n$ *and* $k_\mathrm{max} = \max_n\{k_n\}$*. Note that* $k_\mathrm{max} \leq 6$ *due to Theorem 2.19. Due to the self-starter it is in particular* $\alpha_i^{(i-1)} = 0$ *for* $i = 2, \ldots, k_\mathrm{max}$*. We start with the double sum and use* $m := n - i$ *to obtain*

$$\begin{aligned} \sum_{n=0}^{N-1} \sum_{i=0}^{k_\mathrm{max}} \alpha_i^{(n)} \boldsymbol{\lambda}_{n+1}^\intercal \boldsymbol{w}_{n+1-i} &= \sum_{i=0}^{k_\mathrm{max}} \sum_{m=-i}^{N-1-i} \alpha_i^{(m+i)} \boldsymbol{\lambda}_{m+1+i}^\intercal \boldsymbol{w}_{m+1} \\ &= \sum_{m=0}^{N-1} \alpha_0^{(m)} \boldsymbol{\lambda}_{m+1}^\intercal \boldsymbol{w}_{m+1} + \sum_{i=1}^{k_\mathrm{max}} \sum_{m=-i}^{N-1-i} \alpha_i^{(m+i)} \boldsymbol{\lambda}_{m+1+i}^\intercal \boldsymbol{w}_{m+1}. \end{aligned} \tag{A.3}$$

For steps beyond the integration interval $[t_\mathrm{s}, t_\mathrm{f}] = [t_0, t_N]$ *we set* $\alpha_i^{(n)} := 0$ *for* $n \geq N$ *and* $i = 0, \ldots, k_\mathrm{max}$*. To ease the notion we omit the scalars* $\boldsymbol{\lambda}_{m+1+i}^\intercal \boldsymbol{w}_{m+1}$*, use the convention that the empty sum is zero, and consider for* $i = 1, \ldots, k_\mathrm{max}$

$$\sum_{m=-i}^{N-1-i} \alpha_i^{(m+i)} = \sum_{m=-i}^{-2} \alpha_i^{(m+i)} + \sum_{m=-1}^{N-2} \alpha_i^{(m+i)} - \sum_{m=N-i}^{N-2} \alpha_i^{(m+i)} = \sum_{m=-1}^{N-2} \alpha_i^{(m+i)}.$$

Using the above relation we interchange the sums in (A.3) *to obtain*

$$\begin{aligned}
&\sum_{m=0}^{N-1}\alpha_0^{(m)}\boldsymbol{\lambda}_{m+1}^{\intercal}\boldsymbol{w}_{m+1}+\sum_{m=-1}^{N-2}\sum_{i=1}^{k_{\max}}\alpha_i^{(m+i)}\boldsymbol{\lambda}_{m+1+i}^{\intercal}\boldsymbol{w}_{m+1}\\
&\qquad=\alpha_0^{(N-1)}\boldsymbol{\lambda}_N^{\intercal}\boldsymbol{w}_N+\sum_{i=1}^{k_{\max}}\underbrace{\alpha_i^{(-1+i)}}_{=0,\,i\geq 2}\boldsymbol{\lambda}_i^{\intercal}\boldsymbol{w}_0+\sum_{m=0}^{N-2}\sum_{i=0}^{k_{\max}}\alpha_i^{(m+i)}\boldsymbol{\lambda}_{m+1+i}^{\intercal}\boldsymbol{w}_{m+1}.
\end{aligned}$$

Altogether, the system becomes

$$\begin{aligned}
0=\boldsymbol{\lambda}_0^{\intercal}\boldsymbol{w}_{\mathrm{s}}-&\boldsymbol{\lambda}_0^{\intercal}\boldsymbol{w}_0-\alpha_0^{(N-1)}\boldsymbol{\lambda}_N^{\intercal}\boldsymbol{w}_N-\alpha_1^{(0)}\boldsymbol{\lambda}_1^{\intercal}\boldsymbol{w}_0\\
&-\sum_{n=0}^{N-2}\sum_{i=0}^{k_{\max}}\alpha_i^{(n+i)}\boldsymbol{\lambda}_{n+1+i}^{\intercal}\boldsymbol{w}_{n+1}+\sum_{n=0}^{N-1}\boldsymbol{\lambda}_{n+1}^{\intercal}h_n\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1},\boldsymbol{y}_{n+1})\boldsymbol{w}_{n+1}\\
\Leftrightarrow 0=\boldsymbol{\lambda}_0^{\intercal}\boldsymbol{w}_{\mathrm{s}}-&\boldsymbol{\lambda}_0^{\intercal}\boldsymbol{w}_0-\alpha_0^{(N-1)}\boldsymbol{\lambda}_N^{\intercal}\boldsymbol{w}_N-\alpha_1^{(0)}\boldsymbol{\lambda}_1^{\intercal}\boldsymbol{w}_0+\boldsymbol{\lambda}_N^{\intercal}h_{N-1}\boldsymbol{f}_{\boldsymbol{y}}(t_N,\boldsymbol{y}_N)\boldsymbol{w}_N\\
&-\sum_{n=0}^{N-2}\left\{\sum_{i=0}^{k_{\max}}\alpha_i^{(n+i)}\boldsymbol{\lambda}_{n+1+i}^{\intercal}-\boldsymbol{\lambda}_{n+1}^{\intercal}h_n\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1},\boldsymbol{y}_{n+1})\right\}\boldsymbol{w}_{n+1}\\
\Leftrightarrow 0=\boldsymbol{\lambda}_0^{\intercal}\boldsymbol{w}_{\mathrm{s}}-&\left[\boldsymbol{\lambda}_0+\alpha_1^{(0)}\boldsymbol{\lambda}_1\right]^{\intercal}\boldsymbol{w}_0-\left[\alpha_0^{(N-1)}\boldsymbol{\lambda}_N-h_{N-1}\boldsymbol{f}_{\boldsymbol{y}}^{\intercal}(t_N,\boldsymbol{y}_N)\boldsymbol{\lambda}_N\right]^{\intercal}\boldsymbol{w}_N\\
&-\sum_{n=0}^{N-2}\left[\sum_{i=0}^{k_{\max}}\alpha_i^{(n+i)}\boldsymbol{\lambda}_{n+1+i}-h_n\boldsymbol{f}_{\boldsymbol{y}}^{\intercal}(t_{n+1},\boldsymbol{y}_{n+1})\boldsymbol{\lambda}_{n+1}\right]^{\intercal}\boldsymbol{w}_{n+1}
\end{aligned}$$

Now, requiring that $\{\boldsymbol{\lambda}_n\}_{n=0}^N$ *solve* (3.2) *we obtain for the adjoint direction* $r=J'(\boldsymbol{y}_N)$ *that*

$$0=\boldsymbol{\lambda}_0^{\intercal}\boldsymbol{w}_{\mathrm{s}}-\boldsymbol{r}^{\intercal}\boldsymbol{w}_N \tag{A.4}$$

which describes the relation between the forward and the adjoint IND scheme (for the forward IND scheme see Remark A.5). □

Remark A.5 *If we would vary all* $\boldsymbol{\lambda}_n$ *in* (A.2) *to define* $\boldsymbol{w}_n$*, this would yield the forward IND scheme*

$$\boldsymbol{w}_0=\boldsymbol{w}_{\mathrm{s}} \tag{A.5a}$$

$$\sum_{i=0}^{k_n}\alpha_i^{(n)}\boldsymbol{w}_{n+1-i}=h_n\boldsymbol{f}_{\boldsymbol{y}}(t_{n+1},\boldsymbol{y}_{n+1})\boldsymbol{w}_{n+1},\ n=0,\dots,N-1. \tag{A.5b}$$

This scheme together with the nominal BDF method is again a BDF method applied to the augmented system (1.1) *and* (1.4)*. Hence, the convergence behavior of the forward IND scheme* (A.5) *is the same as that of the nominal BDF method.*

Remark A.6 *The results* $\boldsymbol{w}_N$ *and* $\boldsymbol{\lambda}_0$ *of the forward and adjoint IND scheme, respectively, are related by* (A.4)*. This relation also proves their (transposed) similarity if initialized with* $\boldsymbol{w}_{\mathrm{s}}=\boldsymbol{I}$ *and* $\boldsymbol{r}=\boldsymbol{I}$ *and the convergence behavior of* $\boldsymbol{\lambda}_0$ *towards* $\boldsymbol{\lambda}(t_{\mathrm{s}})$ *to be the same as that of the nominal BDF method, cf. Theorem 1.9 and Remark A.5.*

We now focus on the fomulation (3.3) of the BDF method where each new approximation is given as solution of an implicit function. For the adjoint IND values we use the same notation $\bar{\boldsymbol{y}}_{n+1}$ like in Albersmeyer and Bock [5] and Albersmeyer [3].

Proof (of Lemma 3.2) *The proof follows mainly that of Lemme 3.1. Starting with*

$$0 = -\bar{\boldsymbol{y}}_0^\intercal(\boldsymbol{y}_0 - \boldsymbol{y}_\mathrm{s}) - \sum_{n=0}^{N-1} \bar{\boldsymbol{y}}_{n+1}^\intercal \left(\boldsymbol{y}_{n+1} - \boldsymbol{\theta}_{n+1}(\boldsymbol{y}_n, \dots, \boldsymbol{y}_{n+1-k_n})\right)$$

a variation $\boldsymbol{w}_\mathrm{s}$ *in the initial value leads to*

$$0 = -\bar{\boldsymbol{y}}_0^\intercal(\boldsymbol{w}_0 - \boldsymbol{w}_\mathrm{s}) - \sum_{n=0}^{N-1} \bar{\boldsymbol{y}}_{n+1}^\intercal \left(\boldsymbol{w}_{n+1} - \sum_{i=1}^{k_n} \frac{\partial \boldsymbol{\theta}_{n+1}}{\partial \boldsymbol{y}_{n+1-i}}(\boldsymbol{y}_n, \dots, \boldsymbol{y}_{n+1-k_n}) \cdot \boldsymbol{w}_{n+1-i} \right).$$

According to the Implicit Function Theorem and (3.3a) *it is*

$$\begin{aligned} \frac{\partial \boldsymbol{\theta}_{n+1}}{\partial \boldsymbol{y}_{n+1-i}}(\boldsymbol{y}_n, \dots, \boldsymbol{y}_{n+1-k_n}) &= -\left(\alpha_0^{(n)} \boldsymbol{I} - h_n \boldsymbol{f}_{\boldsymbol{y}}(t_{n+1}, \boldsymbol{\theta}_{n+1})\right)^{-1} \alpha_i^{(n)} \\ &= -\alpha_i^{(n)} \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})^{-1} \end{aligned}$$

for $i = 0, \dots, k_n$. *Inserting in the above system yields*

$$0 = -\bar{\boldsymbol{y}}_0^\intercal(\boldsymbol{w}_0 - \boldsymbol{w}_\mathrm{s}) - \sum_{n=1}^{N-1} \bar{\boldsymbol{y}}_{n+1}^\intercal \left(\boldsymbol{w}_{n+1} + \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1})^{-1} \sum_{i=1}^{k_n} \alpha_i^{(n)} \boldsymbol{w}_{n+1-i} \right). \quad \text{(A.6)}$$

With the same transformations and assumptions like in the proof of Lemma 3.1 we end up with

$$\begin{aligned} 0 =& \bar{\boldsymbol{y}}_0^\intercal \boldsymbol{w}_\mathrm{s} - \left[\bar{\boldsymbol{y}}_0 + \alpha_1^{(0)} \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(0)}(\boldsymbol{y}_1)^{-\intercal} \bar{\boldsymbol{y}}_1\right]^\intercal \boldsymbol{w}_0 - \bar{\boldsymbol{y}}_N^\intercal \boldsymbol{w}_N \\ & - \sum_{n=0}^{N-2} \left[\bar{\boldsymbol{y}}_{n+1} + \sum_{i=1}^{k_{\max}} \alpha_i^{(n+i)} \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n+i)}(\boldsymbol{y}_{n+1+i})^{-\intercal} \bar{\boldsymbol{y}}_{n+1+i}\right]^\intercal \boldsymbol{w}_{n+1}. \end{aligned}$$

Now, requiring that $\{\bar{\boldsymbol{y}}_n\}_{n=0}^N$ *solve* (3.4) *we again obtain for the adjoint direction* $\boldsymbol{r} = J'(\boldsymbol{y}_N)$ *that* $0 = \bar{\boldsymbol{y}}_0^\intercal \boldsymbol{w}_\mathrm{s} - \boldsymbol{r}^\intercal \boldsymbol{w}_N$. □

The adjoint IND scheme (3.4) is the same as the direct adjoint IND scheme presented in Algorithm 6 of Albersmeyer and Bock [5] and Algorithm 6.6 of Albersmeyer [3].

Remark A.7 *Varying all* $\bar{\boldsymbol{y}}_n$ *in* (A.6) *would yield the same forward IND scheme* (A.5) *since the equations are linear in* $\boldsymbol{w}_n$.

Finally, we prove Lemma 3.3 of Section 3.4.

Proof (of Lemma 3.3) *Expressing* (3.2a) *in terms of* $\boldsymbol{\lambda}_N$ *is*

$$\left(\alpha_0^{(N-1)}\boldsymbol{I} - h_{N-1}\boldsymbol{f}_{\boldsymbol{y}}(t_N, \boldsymbol{y}_N)\right)^{\mathsf{T}} \boldsymbol{\lambda}_N = J'(\boldsymbol{y}_N)^{\mathsf{T}} \Leftrightarrow \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(N-1)}(\boldsymbol{y}_N)^{\mathsf{T}}\boldsymbol{\lambda}_N = J'(\boldsymbol{y}_N)^{\mathsf{T}}.$$

Furthermore, by (3.4a) *we have* $\bar{\boldsymbol{y}}_N = J'(\boldsymbol{y}_N)^{\mathsf{T}}$ *such that the assertion is shown for* $n = N-1$. *For* $n = N-2$ (3.2b) *expressed in terms of* $\boldsymbol{\lambda}_{N-1}$ *reads*

$$\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(N-2)}(\boldsymbol{y}_{N-1})^{\mathsf{T}}\boldsymbol{\lambda}_{N-1} = -\alpha_1^{(N-1)}\boldsymbol{\lambda}_N = -\alpha_1^{(N-1)}\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(N-1)}(\boldsymbol{y}_N)^{-\mathsf{T}}\bar{\boldsymbol{y}}_N$$

where the relation between $\boldsymbol{\lambda}_N$ *and* $\bar{\boldsymbol{y}}_N$ *is used. The above right hand side is exactly the right hand side of* (3.4b) *for* $n = N-2$ *such that* $\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(N-2)}(\boldsymbol{y}_{N-1})^{\mathsf{T}}\boldsymbol{\lambda}_{N-1} = \bar{\boldsymbol{y}}_{N-1}$ *and the assertion is shown for* $n = N-2$. *Continuing in this way the assertion is shown for all* $n = N-3, \ldots, 0$. *Finally,* (3.2c) *gives* $\boldsymbol{\lambda}_0 = -\alpha_1^{(0)}\boldsymbol{\lambda}_1 = -\alpha_1^{(0)}\boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(0)}(\boldsymbol{y}_1)^{-\mathsf{T}}\bar{\boldsymbol{y}}_1$ *where the relation between* $\boldsymbol{\lambda}_1$ *and* $\bar{\boldsymbol{y}}_1$ *is used. With* (3.4c) *the last assertion is shown.* □

A.2.2 Proofs of Lemma 8.7 and 8.8

This section contains the technical details corresponding to two lemmas stated in Section 8.4.4.

Proof (of Lemma 8.7) *For a constant BDF method of order* k *with* m *variable starting steps the first integration step with constant BDF coefficients is the* l*-th step with* $l = m + k - 1$. *We start with* $\tilde{E}(\boldsymbol{y}^h)$ *given by* (8.9) *and replace all local truncation errors using* (8.15)

$$\begin{aligned}
\tilde{E}(\boldsymbol{y}^h) &= \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\mathbf{LTE}(t_{n+1}) + \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\boldsymbol{\delta_{n+1}} \\
&= \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\left[\alpha_0^{(n)}\mathbf{LE}(t_{n+1}) - \mathcal{O}(h_n)\mathbf{LE}(t_{n+1})\right] + \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\boldsymbol{\delta_{n+1}} \\
&= \sum_{n=0}^{l-1} \alpha_0^{(n)}\boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\mathbf{LE}(t_{n+1}) + \alpha_0^{(l)}\sum_{n=l}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\mathbf{LE}(t_{n+1}) + \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\boldsymbol{\delta_{n+1}} \\
&\quad - \sum_{n=0}^{l-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\mathcal{O}(h_n)\mathbf{LE}(t_{n+1}) - \mathcal{O}(h^{k+1})
\end{aligned}$$

since $\mathbf{LE}(t_{n+1}) = \mathcal{O}(h^{k+1})$ *for* $n \geq l$ *due to Lemma 2.8 and the consistency order* k *of the constant BDF method, cf. Section 2.3.1. On the other hand,* $\hat{E}(\boldsymbol{y}^h)$ *given by* (8.8) *is equivalent to*

$$\sum_{n=l}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\mathbf{LE}(t_{n+1}) = \hat{E}(\boldsymbol{y}^h) - \sum_{n=0}^{l-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\mathbf{LE}(t_{n+1}) - \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\mathsf{T}}\boldsymbol{\delta_{n+1}}.$$

Both together give

$$\begin{aligned}\tilde{E}(\boldsymbol{y}^h) =& \sum_{n=0}^{l-1} \alpha_0^{(n)} \boldsymbol{\lambda}_{n+1}^{\intercal} \mathbf{LE}(t_{n+1}) + \alpha_0^{(l)} \left[\hat{E}(\boldsymbol{y}^h) - \sum_{n=0}^{l-1} \boldsymbol{\lambda}_{n+1}^{\intercal} \mathbf{LE}(t_{n+1}) - \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\intercal} \boldsymbol{\delta_{n+1}} \right] \\ &+ \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\intercal} \boldsymbol{\delta_{n+1}} - \sum_{n=0}^{l-1} \boldsymbol{\lambda}_{n+1}^{\intercal} \mathcal{O}(h_n) \mathbf{LE}(t_{n+1}) - \mathcal{O}(h^{k+1}) \\ =& \alpha_0^{(l)} \hat{E}(\boldsymbol{y}^h) + \sum_{n=0}^{l-1} (\alpha_0^{(n)} - \alpha_0^{(l)}) \boldsymbol{\lambda}_{n+1}^{\intercal} \mathbf{LE}(t_{n+1}) + (1 - \alpha_0^{(l)}) \sum_{n=0}^{N-1} \boldsymbol{\lambda}_{n+1}^{\intercal} \boldsymbol{\delta_{n+1}} \\ &- \sum_{n=0}^{l-1} \boldsymbol{\lambda}_{n+1}^{\intercal} \mathcal{O}(h_n) \mathbf{LE}(t_{n+1}) - \mathcal{O}(h^{k+1})\end{aligned}$$

and the assertion is shown. □

In Lemma 2.8 the relation between the local error and the local truncation error is described provided that the Localizing Assumption holds. If it does not hold, the relation is described by Lemma 8.8.

Proof (of Lemma 8.8) *Substracting* (2.1) *from* (2.8) *with the non-zero global errors* $\mathbf{GE}(t_{n+1-i}) = \boldsymbol{y}(t_{n+1-i}) - \boldsymbol{y}_{n+1-i} \neq \mathbf{0}$*, since the Localizing Assumption of Definition 2.7 is not satisfied here, yields*

$$\begin{aligned}\mathbf{LTE}(t_{n+1}) =& \sum_{i=0}^{k_n} \alpha_i^{(n)} \mathbf{GE}(t_{n+1-i}) - h_n [\boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1})) - \boldsymbol{f}(t_{n+1}, \boldsymbol{y}_{n+1})] \\ =& \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1}) \mathbf{GE}(t_{n+1}) + \sum_{i=1}^{k_n} \alpha_i^{(n)} \mathbf{GE}(t_{n+1-i}) - h_n \mathcal{O}(\|\mathbf{GE}(t_{n+1})\|^2)\end{aligned}$$

by the Taylor series expansion of $\boldsymbol{f}(t_{n+1}, \boldsymbol{y}(t_{n+1}))$ *around* $\boldsymbol{y}_{n+1}$*. With the BDF Jacobian and the relation* $\mathbf{GE}(t_{n+1}) = \boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1}) + \mathbf{LE}(t_{n+1})$ *we obtain*

$$\begin{aligned}\mathbf{LTE}(t_{n+1}) =& \alpha_0^{(n)} \mathbf{LE}(t_{n+1}) - \mathcal{O}(h_n) \mathbf{LE}(t_{n+1}) + \boldsymbol{\mathcal{J}}_{\mathrm{BDF}}^{(n)}(\boldsymbol{y}_{n+1}) [\boldsymbol{y}(t_{n+1}) - \boldsymbol{u}_n(t_{n+1})] \\ &+ \sum_{i=1}^{k_n} \alpha_i^{(n)} \mathbf{GE}(t_{n+1-i}) - h_n \mathcal{O}(\|\mathbf{GE}(t_{n+1})\|^2).\end{aligned}$$

Finally, by the observation that $\mathcal{O}(h_n \|\mathbf{LE}(t_{n+1})\|)$ *dominates* $\mathcal{O}(h_n \|\mathbf{GE}(t_{n+1})\|^2)$ *the assertion is shown.* □

A.3 Supplementary material for Part IV

A.3.1 Test set

The first three examples are originally chosen by Cao and Petzold [43] and the subsequent three examples by Tran and Berzins [118]. Finally, we also state here the Catenary of Section 10.1.

Example 1 (Dahlquist equation)

$$\dot{y}(t) = ay(t), \quad t \in (0, 10], \quad y(0) = 10^{-4}, \quad a = 1.$$

The analytic solution is given by $y(t) = y_{\mathrm{s}} e^{at}$ *and the locally analytic solution for* $y(t_{\mathrm{s}}) = y_{\mathrm{s}}$ *by* $y(t) = y_{\mathrm{s}} e^{a(t-t_{\mathrm{s}})}$.

Example 2

$$\dot{y}(t) = -\left[0.25 + \sin(\pi t)\right] y(t)^2, \quad t \in (0, 1], \quad y(0) = 1.$$

The analytic solution is given by $y(t) = \pi/(\pi + 1 + 0.25\pi t - \cos(\pi t))$ *and the locally analytic solution for* $y(t_{\mathrm{s}}) = y_{\mathrm{s}}$ *by* $y(t) = y_{\mathrm{s}}\pi/(\pi + y_{\mathrm{s}}\cos(\pi t_{\mathrm{s}}) - 0.25\pi y_{\mathrm{s}}(t_{\mathrm{s}} - t) - y_{\mathrm{s}}\cos(\pi t))$.

Example 3

$$\dot{\boldsymbol{y}}(t) = \begin{pmatrix} \frac{1}{2(1+t)} & -2t \\ 2t & \frac{1}{2(1+t)} \end{pmatrix} \boldsymbol{y}(t), \quad t \in (0, 10], \quad \boldsymbol{y}(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

The analytic solution is given by $\boldsymbol{y}(t) = [(1+t)^{1/2}\cos(t^2), (1+t)^{1/2}\sin(t^2)]^\intercal$.

Example 4 (Harmonic oscillator)

$$\dot{\boldsymbol{y}}(t) = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \boldsymbol{y}(t), \quad t \in (0, 50], \quad \boldsymbol{y}(0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

The analytic solution is given by $\boldsymbol{y}(t) = [\sin(t), \cos(t)]^\intercal$.

Example 5

$$\begin{cases} \dot{y}_1 = y_1 \\ \dot{y}_2 = y_2 + y_1 y_1 \\ \dot{y}_3 = y_3 + y_1 y_2 \\ \dot{y}_4 = y_4 + y_1 y_3 + y_2 y_2 \\ \dot{y}_5 = y_5 + y_1 y_4 + y_2 y_3 \end{cases}, \quad t \in (0, 1], \quad \boldsymbol{y}(0) = \begin{pmatrix} 1 \\ 1 \\ 0.5 \\ 0.5 \\ 0.25 \end{pmatrix}.$$

The analytic solution is given by $\boldsymbol{y}(t) = [e^t, e^{2t}, 0.5e^{3t}, 0.5e^{4t}, 0.25e^{5t}]^\intercal$.

Example 6

$$\dot{y}(t) = -L\left[y(t) - \sin(\pi t)\right] + \pi\cos(\pi t), \quad t \in (0, 1], \quad y(0) = 0.$$

The analytic solution is given by $y(t) = \sin(\pi t)$ *and* L *is positive and may be large. Exemplarily we take* $L = 50$.

Example 7 (Catenary)

$$\dot{\boldsymbol{y}}(t) = \begin{pmatrix} y_2(t) \\ p\sqrt{1+y_2(t)^2} \end{pmatrix}, \quad t \in (0,2], \quad p = 3, \quad \boldsymbol{y}(0) = \begin{pmatrix} 1/3\cos(-3) \\ \sin(-3) \end{pmatrix}.$$

The analytic solution is given by $\boldsymbol{y}(t) = [B + 1/p\cosh(pt + A), \sinh(pt + A)]^\intercal$ *with* $A = -p$ *and* $B = 0$*, cf. Section 10.1. The locally analytic solution for* $\boldsymbol{y}(t_\mathrm{s}) = \boldsymbol{y}_\mathrm{s}$ *is given by* $\boldsymbol{y}(t) = [B + 1/p\cosh(p(t - t_\mathrm{s}) + A), \sinh(p(t - t_\mathrm{s}) + A)]^\intercal$ *where* $A = \ln\left((\boldsymbol{y}_\mathrm{s})_2 + \sqrt{1 + [(\boldsymbol{y}_\mathrm{s})_2]^2}\right)$ *and* $B = (\boldsymbol{y}_\mathrm{s})_1 - 1/p \cdot \cosh(A)$.

A.3.2 Additional data for the IVP model of the hydrolysis

For the mathematical model of the hydrolysis of propionic anhydride carried out in a discontinuous Stirred Tank Reactor (STR) further chemical quantities and equipment parameters are required. The modeling Initial Value Problem (IVP) in Ordinary Differential Equations (ODEs) is developed in Section 12.2.1.

Symbol	Value	Unit	Symbol	Value	Unit
M_Ah	0.130150	kg/mol	$C_{p,\mathrm{Ah}}$	1822.316117	J/(kg K)
M_w	0.0180150	kg/mol	$C_{p,\mathrm{w}}$	4176.665782	J/(kg K)
M_Ac	0.0740790	kg/mol	$C_{p,\mathrm{Ac}}$	2111.839763	J/(kg K)
M_S	0.098080	kg/mol	$C_{p,\mathrm{S}}$	1480.0	J/(kg K)

Table A.1: Molar mass M and heat capacity C_p of the chemical substances. C_p is given at a reference temperature of 313.15K.

Symbol	Value	Unit	Symbol	Value	Unit
UA_1	6.712368215195024	W/(m^2 K)	V_1	0.001100891625830	m^3
UA_2	7.852551350287481	W/(m^2 K)	V_2	0.001496613831028	m^3
UA_0	0.207160211598949	W/(m^2 K)			

Table A.2: Values used for the heat transfer coefficients of the heat flow $q_\mathrm{flow}(t)$ and the heat loss $q_\mathrm{loss}(t)$ and obtained by calibration of the semibatch experiment with a dosing time of 1000s.

The setups of two more experiments are given below.

Symbol	Value	Unit	Symbol	Value	Unit
$n_{\mathrm{w}}(t_{\mathrm{s}})$	$(1.016+[1-p_{\mathrm{Ah}}]0.49+$ $[1-p_{\mathrm{S}}]0.072)/M_{\mathrm{Ah}}$	mol	T_{j}	313.15	K
$T(t_{\mathrm{s}})$	313.15	K	T_{amb}	296.15	K
$n_{\mathrm{Ah}}^{\mathrm{aq}}(t_{\mathrm{s}})$	0	mol	u_{d}	0	kg/s
$n_{\mathrm{Ah}}^{\mathrm{org}}(t_{\mathrm{s}})$	$p_{\mathrm{Ah}} \cdot 0.49/M_{\mathrm{Ah}}$	mol	t_{d}	0	s
$n_{\mathrm{Ac}}(t_{\mathrm{s}})$	0	mol	n_{S}	$p_{\mathrm{S}} \cdot 0.072/M_{\mathrm{S}}$	mol

Table A.3: Initial values and experimental parameters of the batch experiment with a propionic anhydride amount of 0.49kg.

Symbol	Value	Unit	Symbol	Value	Unit
$n_{\mathrm{w}}(t_{\mathrm{s}})$	$(0.76257+[1-p_{\mathrm{S}}]0.0571)/M_{\mathrm{Ah}}$	mol	T_{j}	313.15	K
$T(t_{\mathrm{s}})$	313.15	K	T_{amb}	293.65	K
$n_{\mathrm{Ah}}^{\mathrm{aq}}(t_{\mathrm{s}})$	0	mol	u_{d}	0.95/2000	kg/s
$n_{\mathrm{Ah}}^{\mathrm{org}}(t_{\mathrm{s}})$	0	mol	t_{d}	2000	s
$n_{\mathrm{Ac}}(t_{\mathrm{s}})$	0	mol	n_{S}	$p_{\mathrm{S}} \cdot 0.0571/M_{\mathrm{S}}$	mol

Table A.4: Initial values and experimental parameters of the semibatch experiment with a dosing time of 2000s and a propionic anhydride amount of 0.95kg.

A.3.3 Reference solution for the hydrolysis IVP

By the computations for the weak adjoints of Section 12.3 with nominal integrations using `RelTol` $= 10^{-3}, \ldots, 10^{-9}$ we obtain the following reference solutions for the trajectory values $\boldsymbol{y}(t_{\mathrm{d}})$ at $t_{\mathrm{d}} = 1000$

$$\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{d}}) = \begin{pmatrix} 54.7198483238 \\ 326.93545950 \\ 0.216095521 \\ 0.0022750243 \\ 5.5256100411 \end{pmatrix}$$

and for $\boldsymbol{y}(t_{\mathrm{f}})$ at $t_{\mathrm{f}} = 3500$

$$\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{f}}) = \begin{pmatrix} 54.5014779465 \\ 313.04440465 \\ 0.00000016877 \\ 0.0000000000000000 \\ 5.96235079575 \end{pmatrix}.$$

The criterion of interest at these reference solutions takes the values $J(\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{d}})) = S(t_{\mathrm{d}}) = 329.0855962924183586$ and $J(\boldsymbol{y}^{\mathrm{r}}(t_{\mathrm{f}})) = S(t_{\mathrm{f}}) = 313.0444063117229234$.

List of acronyms

AD	Algorithmic Differentiation
BDF	Backward Differentiation Formula
BVP	Boundary Value Problem
CVP	Constrained Variational Problem
DAE	Differential Algebraic Equation
DWR	Dual Weighted Residual
END	External Numerical Differentiation
FE	Finite Element
IND	Internal Numerical Differentiation
IVP	Initial Value Problem
LMM	Linear Multistep Method
NLP	Nonlinear Program
OCP	Optimal Control Problem
ODE	Ordinary Differential Equation
PDE	Partial Differential Equation
SQP	Sequential Quadratic Programming
STR	Stirred Tank Reactor

Bibliography

[1] R.A. Adams and J.F. Fournier. *Sobolev Spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] J. Albersmeyer. Effiziente Ableitungserzeugung in einem adaptiven BDF-Verfahren. Diploma thesis, Universität Heidelberg, 2005.

[3] J. Albersmeyer. *Adjoint based algorithms and numerical methods for sensitivity generation and optimization of large scale dynamic systems*. PhD thesis, Ruprecht–Karls–Universität Heidelberg, 2010.

[4] J. Albersmeyer, D. Beigel, C. Kirches, L. Wirsching, H.G. Bock, and J.P. Schlöder. Fast nonlinear model predictive control with an application in automotive engineering. In L. Magni, D.M. Raimondo, and F. Allgöwer, editors, *Lecture Notes in Control and Information Sciences*, volume 384, pages 471–480. Springer Verlag Berlin Heidelberg, 2009.

[5] J. Albersmeyer and H.G. Bock. Efficient sensitivity generation for large scale dynamic systems. Technical report, SPP 1253 Preprints, University of Erlangen, 2009.

[6] J. Albersmeyer and C. Kirches. The `SolvIND` webpage. http://www.solvind.org, 2007.

[7] H.W. Alt. *Lineare Funktionalanalysis*. Springer-Verlag Berlin Heidelberg, 4 edition, 2002.

[8] U.M. Ascher and L.R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential–Algebraic Equations*. SIAM, Philadelphia, 1998.

[9] I. Babuška and A. Miller. The post-processing approach in the finite element method–part 1: Calculation of displacements, stresses and other higher derivatives of the displacements. *Int. J. Numer. Meth. Eng.*, 20(6):1085–1109, 1984.

[10] I. Babuška and A. Miller. The post-processing approach in the finite element method–part 2: The calculation of stress intensity factors. *Int. J. Numer. Meth. Eng.*, 20(6):1111–1129, 1984.

[11] I. Babuška and A. Miller. The post-processing approach in the finite element method–part 3: A posteriori error estimates and adaptive mesh selection. *Int. J. Numer. Meth. Eng.*, 20(12):2311–2324, 1984.

[12] I. Babuška and W.C. Rheinboldt. A-posteriori error estimates for the finite element method. *Int. J. Numer. Meth. Eng.*, 12:1597–1615, 1978.

[13] I. Babuška and W.C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.

[14] I. Babuška and T. Strouboulis. *The Finite Element Method and its Reliability*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2001.

[15] W. Bangerth and R. Rannacher. *Adaptive Finite Element Methods for Differential Equations*. Birkhäuser Verlag, 2003.

[16] I. Bauer. *Numerische Verfahren zur Lösung von Anfangswertaufgaben und zur Generierung von ersten und zweiten Ableitungen mit Anwendungen bei Optimierungsaufgaben in Chemie und Verfahrenstechnik*. PhD thesis, Universität Heidelberg, 1999.

[17] I. Bauer, H.G. Bock, S. Körkel, and J.P. Schlöder. Numerical methods for optimum experimental design in DAE systems. *J. Comput. Appl. Math.*, 120(1-2):1–15, 2000.

[18] R. Becker and R. Rannacher. A feed-back approach to error control in finite element methods: Basic analysis and examples. *East-West J. Numer. Math*, 4:237–264, 1996.

[19] R. Becker and R. Rannacher. Weighted a posteriori error control in FE methods. In *Enumath 1997: Proceedings of the 2nd European Conference on Numerical Mathematics and Advanced Applications, Heidelberg, Germany 28 September-3 October 1997*, pages 621–637, Singapore, 1998. World Scientific Publ. Preprint 96-1 (SFB 359), Universität Heidelberg.

[20] R. Becker and R. Rannacher. An optimal control approach to error estimation and mesh adaptation in finite element methods. *Acta Numerica 2000*, pages 1–101, 2001.

[21] D. Beigel, M.S. Mommer, L. Wirsching, and H.G. Bock. Approximation of weak adjoints by reverse automatic differentiation of BDF methods. Technical report, arXiv.org, Cornell University Library, 2011. Available at http://arxiv.org/abs/1109.3061.

[22] L.D. Berkovitz. *Optimal Control Theory*, volume 12 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1974.

[23] J.T. Betts. *Practical Methods for Optimal Control Using Nonlinear Programming*. SIAM, Philadelphia, 2001.

[24] T. Binder, L. Blank, H.G. Bock, R. Bulirsch, W. Dahmen, M. Diehl, T. Kronseder, W. Marquardt, J.P. Schlöder, and O.v. Stryk. Introduction to model based optimization of chemical processes on moving horizons. In M. Grötschel, S.O. Krumke, and J. Rambau, editors, *Online Optimization of Large Scale Systems: State of the Art*, pages 295–340. Springer, 2001.

[25] G. Bleser. Eine effiziente Ordnungs– und Schrittweitensteuerung unter Verwendung von Fehlerformeln für variable Gitter und ihre Realisierung in Mehrschrittverfahren vom BDF–Typ. Diploma thesis, Universität Bonn, 1986.

[26] H.G. Bock. Numerical solution of nonlinear multipoint boundary value problems with applications to optimal control. *Zeitschrift für Angewandte Mathematik und Mechanik*, 58:407, 1978.

[27] H.G. Bock. *Numerische Berechnung zustandsbeschränkter optimaler Steuerungen mit der Mehrzielmethode*. Carl-Cranz-Gesellschaft, Heidelberg, 1978.

[28] H.G. Bock. Numerical treatment of inverse problems in chemical reaction kinetics. In K.H. Ebert, P. Deuflhard, and W. Jäger, editors, *Modelling of Chemical Reaction Systems*, volume 18 of *Springer Series in Chemical Physics*, pages 102–125. Springer, Heidelberg, 1981.

[29] H.G. Bock. Numerische Behandlung von zustandsbeschränkten und Chebyshev-Steuerungsproblemen. Technical Report R106/81/11, Carl Cranz Gesellschaft, Heidelberg, 1981.

[30] H.G. Bock. Recent advances in parameter identification techniques for ODE. In P. Deuflhard and E. Hairer, editors, *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, pages 95–121. Birkhäuser, Boston, 1983.

[31] H.G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*, volume 183 of *Bonner Mathematische Schriften*. Universität Bonn, Bonn, 1987.

[32] H.G. Bock, M. Diehl, E.A. Kostina, and J.P. Schlöder. Constrained Optimal Feedback Control for DAE. In L. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, and B. van Bloemen Waanders, editors, *Real-Time PDE-Constrained Optimization*, chapter 1, pages 3–24. SIAM, 2007.

[33] H.G. Bock and K.J. Plitt. A Multiple Shooting algorithm for direct solution of optimal control problems. In *Proceedings of the 9th IFAC World Congress*, pages 242–247, Budapest, 1984. Pergamon Press. Available at http://www.iwr.uni-heidelberg.de/groups/agbock/FILES/Bock1984.pdf.

[34] H.G. Bock and J.P. Schlöder. Numerical solution of retarded differential equations with state-dependent time lags. *Zeitschrift für Angewandte Mathematik und Mechanik*, 61:269, 1981.

[35] H.G. Bock, J.P. Schlöder, and V.H. Schulz. Numerik großer Differentiell-Algebraischer Gleichungen – Simulation und Optimierung. In H. Schuler, editor, *Prozeßsimulation*, pages 35–80. VCH Verlagsgesellschaft mbH, Weinheim, 1994.

[36] K. Böttcher and R. Rannacher. Adaptive error control in solving ordinary differential equations by the discontinuous galerkin method. Preprint 96-53, SFB 359, University of Heidelberg, 1996.

[37] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, 3rd edition, 2007. Theory, fast solvers, and applications in elasticity theory.

[38] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*. SIAM, Philadelphia, 1996. Classics in Applied Mathematics 14.

[39] S.C. Brenner and L.R. Scott. *The mathematical theory of finite element methods*. Springer, 2008.

[40] R. Bulirsch. Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung. Technical report, Carl-Cranz-Gesellschaft, Oberpfaffenhofen, 1971.

[41] J.C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge–Kutta and General Linear Methods*. Wiley, 1987. ISBN 0-471-91046-5 (paperback).

[42] Y. Cao, S. Li, and L. Petzold. Adjoint sensitivity analysis for differential-algebraic equations: algorithms and software. *Journal of Computational and Applied Mathematics*, 149:171–191, 2002.

[43] Y. Cao and L. Petzold. A posteriori error estimation and global error control for ordinary differential equations by the adjoint method. *SIAM J.Sci. Comput.*, 26:359–374, 2004.

[44] R. Cherbański. *Zastosowanie sieci neuronowych do opisu kinetyki złożonych reakcji chemicznych*. PhD thesis, Warsaw University of Technology, 2002.

[45] C.W. Cryer. On the instability of high order backward-difference multistep methods. *BIT*, 12:17–25, 1972.

[46] C.F. Curtiss and J.O. Hirschfelder. Integration of stiff equations. *Proc. Nat. Acad. Sci*, 38:235–243, 1952.

[47] G. Dahlquist. Convergence and stability in numerical integration of ordinary differential equations. *Math. Scand.*, 4:33–53, 1956.

[48] M. Diehl. *Real-Time Optimization for Large Scale Nonlinear Processes*. PhD thesis, Universität Heidelberg, 2001.

[49] M. Diehl, H.G. Bock, and E. Kostina. An approximation technique for robust nonlinear optimization. *Mathematical Programming*, 107:213–230, 2006.

[50] M. Diehl, H.G. Bock, J.P. Schlöder, R. Findeisen, Z. Nagy, and F. Allgöwer. Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. *J. Proc. Contr.*, 12(4):577–585, 2002.

[51] P. Eberhard and C. Bischof. Automatic differentiation of numerical integration algorithms. *Mathematics of Computation*, 68(226):717–731, 1999.

[52] E. Eich. Numerische Behandlung semi-expliziter differentiell-algebraischer Gleichungssysteme vom Index I mit BDF Verfahren. Diploma thesis, Universität Bonn, 1987.

[53] E. Eich. *Projizierende Mehrschrittverfahren zur numerischen Lösung von Bewegungsgleichungen technischer Mehrkörpersysteme mit Zwangsbedingungen und Unstetigkeiten*. PhD thesis, University of Augsburg, 1991.

[54] K. Enke. Ein effizientes Rückwärtsdifferentiationsverfahren mit variabler Schrittweite und Ordnung zur Lösung steifer Anfangswertaufgaben. Diploma thesis, Universität Bonn, 1984.

[55] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Introduction to Adaptive Methods for Differential Equations. *Acta Numerica*, pages 105–158, 1995.

[56] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. *Computational Differential Equations*. Cambridge University Press, 1996.

[57] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*. Springer, 2004.

[58] D. Estep. A posteriori error bounds and global error control for approximation of ordinary differential equations. *SIAM Journal on Numerical Analysis*, 32(1):1–48, 1995.

[59] R. Fletcher. *Practical Methods of Optimization*. Wiley, Chichester, 2nd edition, 1987. ISBN 0-471-49463-1 (paperback).

[60] H. Gajewski, K. Gröger, and K. Zacharias. *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*. Akademie-Verlag, Berlin, 1974.

[61] C.W. Gear. *Numerical initial value problems in ordinary differential equations*. Prentice Hall, Englewood Cliffs, NJ, 1971.

[62] C.W. Gear. Asymptotic estimation of errors and derivatives for the numerical solution of ordinary differential equations. *IFIP 74*, pages 447–451, 1974.

[63] C.W. Gear and D.S. Watanabe. Stability and convergence of variable order multistep methods. *SIAM Journal on Numerical Analysis*, 11(5):1044–1058, 1974.

[64] M. Gerdts. *Optimal Control of ODEs and DAEs*. De Gruyter, 2012.

[65] A. Griewank. *Evaluating Derivatives, Principles and Techniques of Algorithmic Differentiation*. Number 19 in Frontiers in Applied Mathematics. SIAM, Philadelphia, 2000.

[66] I.E. Grossmann, P.A. Aguirre, and M. Barttfeld. Optimal synthesis of complex distillation columns using rigorous models. *Computers and Chemical Engineering*, 29:1203–1215, 2005.

[67] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I*, volume 8 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 1993.

[68] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II – Stiff and Differential-Algebraic Problems*. Springer, Berlin Heidelberg, 1991.

[69] P. Hartman. *Ordinary differential equations*, volume 38 of *Classics in Applied Mathematics*. SIAM, Philadelphia, PA, 2002. Corrected reprint of the second (1982) edition [Birkhäuser, Boston, MA; MR0658490 (83e:34002)].

[70] R. Hartmann. Adjoint consistency analysis of Discontinuous Galerkin discretizations. *SIAM J. Numer. Anal.*, 45:2671–2696, 2007.

[71] M.T. Heath. *Scientific Computing: An Introductory Survey*. McGraw-Hill Higher Education, second edition, 2002.

[72] P. Henrici. *Discrete Variable Methods in Ordinary Differential Equations*. John Wiley and Sons, New York, 1962.

[73] P. Henrici. *Error Propagation for Difference Methods*. Robert E. Krieger Publishing Co., Huntington, N. Y., 1970. Reprint of the 1963 edition.

[74] A.C. Hindmarsh, P.N. Brown, K.E. Grant, S.L. Lee, R. Serban, D.E. Shumaker, and C.S. Woodward. SUNDIALS: Suite of Nonlinear and Differential/Algebraic Equation Solvers. *ACM Transactions on Mathematical Software*, 31(3):363–396, September 2005.

[75] A.D. Ioffe and V.M. Tihomirov. *Theory of extremal problems*, volume 6 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1979.

[76] C. Johnson. *Numerical solutions of partial differential equations by the finite element method*. Cambridge University Press, Cambridge, 1987.

[77] C. Johnson. Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations. *SIAM Journal on Numerical Analysis*, 25(4):908–926, 1988.

[78] C. Kirches, L. Wirsching, S. Sager, and H.G. Bock. Efficient numerics for nonlinear model predictive control. In M. Diehl, F. Glineur, E. Jarlebring, and W. Michiels, editors, *Recent Advances in Optimization and its Applications in Engineering*, pages 339–359. Springer, 2010. ISBN 978-3-6421-2597-3.

[79] A.N. Kolmogorov and S.V. Fomin. *Introductory real analysis*. Revised English edition. Translated from the Russian and edited by Richard A. Silverman. Prentice-Hall Inc., Englewood Cliffs, N.Y., 1970.

[80] S. Körkel, E. Kostina, H.G. Bock, and J.P. Schlöder. Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes. *Optimization Methods and Software*, 19:327–338, 2004.

[81] P. Kühl, M. Diehl, A. Milewska, E. Molga, and H.G. Bock. Robust NMPC for a benchmark fed-batch reactor with runaway conditions. In R. Findeisen, F. Allgoewer, and L.T. Biegler, editors, *Assessment and Future Directions of Nonlinear Model Predictive Control*, volume 358 of *Lecture Notes in Control and Information Sciences*, pages 455–464. Springer Berlin/Heidelberg, 2007.

[82] P. Kühl, A. Milewska, M. Diehl, E. Molga, and H.G. Bock. NMPC for runaway-safe fed-batch reactors. In *Proc. Int. Workshop on Assessment and Future Directions of NMPC*, pages 467–474, 2005.

[83] J. D. Lambert. *Computational Methods in Ordinary Differential Equations*. Wiley, New York, 1973.

[84] J. Lang and J.G. Verwer. On global error estimation and control for initial value problems. *SIAM J. Sci. Comput.*, 29:1460–1475, 2007.

[85] R.J. LeVeque. *Finite difference methods for ordinary and partial differential equations*. SIAM, Philadelphia, PA, 2007. Steady-state and time-dependent problems.

[86] S. Li, L. Petzold, and W. Zhu. Sensitivity analysis of differential-algebraic equations: A comparison of methods on a special problem. *Applied Numerical Mathematics*, 32(2):161 – 174, 2000.

[87] A. Logg. Multi-adaptive galerkin methods for ODEs II: Implementation and applications. *SIAM J. Sci. Comput.*, 25(4):1119–1141, 2003.

[88] D.G. Luenberger. *Optimization by vector space methods*. Wiley Professional Paperback Series. John Wiley & Sons, Inc., New York, NY, 1969. ISBN 0471-18117-X (paperback).

[89] J.N. Lyness and C.B. Moler. Numerical differentiation of analytic functions. *SIAM Journal on Numerical Analysis*, 4:202–210, 1967.

[90] Maplesoft. *Maple 13*. Maplesoft, Inc., 2009.

[91] D. Meidner. *Adaptive Space-Time Finite Element Methods for Optimization Problems Governed by Nonlinear Parabolic Systems*. PhD thesis, University of Heidelberg, 2008.

[92] A. Milewska. *Modelling of batch and semibatch chemical reactors – safety aspects*. PhD thesis, Warsaw University of Technology, 2006.

[93] E. Molga and R. Cherbański. Catalytic reaction performed in the liquid-liquid system at batch and semibatch operating mode. *Catalysis Today*, 66:325–333, 2001.

[94] E. Molga and R. Cherbański. Catalytic reaction performed in the liquid-liquid system: comparison of conventional and neural networks modelling methods. *Catalysis Today*, 79–80:241–247, 2003.

[95] K.-S. Moon, A. Szepessy, R. Tempone, and G. Zouraris. Convergence rates for adaptive approximation of ordinary differential equations. *Numerische Mathematik*, 96:99–129, 2003.

[96] K.-S. Moon, A. Szepessy, R. Tempone, and G. Zouraris. A variational principle for adaptive approximation of ordinary differential equations. *Numerische Mathematik*, 96:131–152, 2003.

[97] I.P. Natanson. *Theorie der Funktionen einer reellen Veränderlichen*. Akademie-Verlag, Berlin, 1975. Übersetzung nach der zweiten russischen Auflage von 1957.

[98] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer Verlag, Berlin Heidelberg New York, 2nd edition, 2006. ISBN 0-387-30303-0 (hardcover).

[99] T.A. Oliver and D.L. Darmofal. Analysis of dual consistency for discontinuous Galerkin discretization of source terms. *SIAM J. Num. Anal.*, 47(5):3507–3525, 2009.

[100] M.R. Osborne. On shooting methods for boundary value problems. *Journal of Mathematical Analysis and Applications*, 27:417–433, 1969.

[101] A. Potschka. *A direct method for the numerical solution of optimization problems with time-periodic PDE constraints*. PhD thesis, Universität Heidelberg, 2012.

[102] Wolfram Research. *Mathematica Edition: Version 7.0*. Wolfram Research, Inc., 2008.

[103] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1966.

[104] A. Sandu. On the properties of Runge–Kutta discrete adjoints. In V. Alexandrov, G. van Albada, P. Sloot, and J. Dongarra, editors, *Computational Science – ICCS 2006*, volume 3994 of *Lecture Notes in Computer Science*, pages 550–557. Springer Berlin/Heidelberg, 2006.

[105] A. Sandu. On consistency properties of discrete adjoint linear multistep methods. Technical report, Virginia Polytechnic Institute and Sate University, Blacksburg, 2007.

[106] A. Sandu. Reverse automatic differentiation of linear multistep methods. In C.H. Bischof, H.M. Bücker, P. Hovland, U. Naumann, and J. Utke, editors, *Advances in Automatic Differentiation*, volume 64 of *Lecture Notes in Computational Science and Engineering*, pages 1–12. Springer-Verlag, Berlin, 2008.

[107] A.A.S. Schäfer. *Efficient reduced Newton-type methods for solution of large-scale structured optimization problems with application to biological and chemical processes*. PhD thesis, Universität Heidelberg, 2005.

[108] R. Serban and A.C. Hindmarsch. CVODES: An ODE solver with sensitivity capabilities. Technical Report UCRL-JP-200039, LLNL, 2003.

[109] L.F. Shampine. *Numerical solution of ordinary differential equations*. Chapman & Hall, New York, 1994.

[110] L.F. Shampine. Erros estimation and control for ODEs. *J. Sci. Comp.*, 25(1):3–16, 2005.

[111] L.F. Shampine and M.K. Gordon. *Computer Solution of Ordinary Differential Equations*. Freeman, San Francisco, 1975.

[112] L.F. Shampine and W. Zhang. Rate of convergence of multistep codes started by variation of order and stepsize. *SIAM J. Numer. Anal.*, 27(6):1506–1518, 1990.

[113] R.D. Skeel. Thirteen ways to estimate global error. *Numerische Mathematik*, 48:1–20, 1986.

[114] R.D. Skeel. Global error estimation and the backward differentiation formulas. *Applied Mathematics and Computation*, 31:197 – 208, 1989.

[115] H.J. Stetter. Economical global error estimation. In R.A. Willoughby, editor, *Stiff differential systems*, The IBM Research Symposia Series, pages 245–258, New York, 1974. Plenum Press.

[116] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer, 1992.

[117] K. Strehmel and R. Weiner. *Numerik gewöhnlicher Differentialgleichungen*. Teubner, Stuttgart, 1995.

[118] L.T. Tran and M. Berzins. Defect Sampling in Global Error Estimation for ODEs and Method-Of-Lines PDEs Using Adjoint Methods. SCI Technical Report UUSCI-2011-006, SCI Institute, University of Utah, 2011.

[119] O. Ubrich, B. Srinivasan, F. Stoessel, and D. Bonvin and. Optimisation of a semi-batch reaction system under safety constraints. In *5th European Control Conference*, volume 12, Karlsruhe, 1999.

[120] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Advances in numerical mathematics. Wiley, Teubner, 1996.

[121] R. von Schwerin. *Numerical methods, algorithms, and software for higher index nonlinear differential-algebraic equations in multibody system simulation*. PhD thesis, Universität Heidelberg, 1997.

[122] A. Walther. Automatic differentiation of explicit Runge-Kutta methods for optimal control. *Comput. Optim. Applic.*, 36:83–108, 2007.

[123] D. Werner. *Funktionalanalysis*. Springer-Verlag, Berlin, 2000.

[124] K. Westerterp and E. Molga. No more runaways in fine chemical reactors. *Ind. Eng. Chem. Res.*, 43(16):4585–4594, 2004.

[125] K.R. Westerterp and E.J. Molga. Safety and runaway prevention in batch and semibatch reactors – a review. *Chemical Engineering Research and Design*, 84(7):543–552, 2006.

[126] J. Wloka. *Funktionalanalysis und Anwendungen*. Walter de Gruyter, Berlin-New York, 1971. de Gruyter Lehrbuch.

[127] P.E. Zadunaisky. A method for the estimation of errors propagated in the numerical solution of a system of ordinary differential equations. In *Proc. Intern. Astron. Union, Symposium No. 25, Thessaloniki, 1964*. Academic Press, 1966.

[128] P.E. Zadunaisky. On the estimation of errors propagated in the numerical integration of ordinary differential equations. *Numer. Math.*, 27:21–39, 1976.

[129] J.M. Zaldivar, H. Hernándes, and C. Barcons. Development of a mathematical model and a simulator for the analysis and optimisation of batch reactors: experimental model characterisation using a reaction calorimeter. *Thermochimica Acta*, 289:267–302, 1996.

[130] E. Zeidler. *Nonlinear Functional Analysis and its Application I.* Springer-Verlag, New York, 1989.