INAUGURAL–DISSERTATION

zur

Erlangung der Doktorwürde

der

Naturwissenschaftlich–Mathematischen Gesamtfakultät

der

RUPRECHT–KARLS–UNIVERSITÄT

HEIDELBERG

vorgelegt von

Dipl.-Phys. Simon M. Lenz

aus Dortmund

Tag der mündlichen Prüfung

Impulsive Hybrid Discrete-Continuous Delay Differential Equations

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

Kurzbeschreibung

Thema dieser Arbeit ist ein neuer Typ von Differentialgleichungen, der aus zwei Gründen hochgradig anspruchsvoll ist. Einerseits werden Abhängigkeiten der rechten Seite von vergangenen Zuständen betrachtet, wobei die Zeitverzögerungen vom aktuellen Zustand abhängen. Andererseits werden Unstetigkeiten in der rechten Seite und in den Zuständen an implizit definierten Zeitpunkten zugelassen. Für den neuen Differentialgleichungstyp wird der englische Fachbegriff "impulsive hybrid discrete-continuous delay differential equation" (IHDDE) verwendet.

Die vorgestellten theoretischen Ergebnisse und numerischen Methoden stehen in Bezug zu drei Themengebieten: Lösung von Anfangswertproblemen (engl.: initial value problems, kurz IVPs) bei IHDDEs; Ableitungen von IVP Lösungen nach Parametern (auch "Sensitivitäten" genannt); und schließlich Schätzung von Parametern in IHDDE Modellen aus experimentellen Daten. Diese Arbeit liefert dabei unter anderem die folgenden Beiträge:

- Die theoretischen Grundlagen von IHDDE-IVPs werden bereitgestellt. Dies umfasst die Definition eines Lösungsbegriffs, die Existenz und Eindeutigkeit von Lösungen, sowie die Differenzierbarkeit von Lösungen nach Parametern.
- Ein neuer Ansatz zur numerischen Lösung von IVPs bei Differentialgleichungen mit Zeitverzögerungen wird vorgestellt, dessen Kernaspekt die Verwendung von Extrapolationen über vergangene Unstetigkeiten hinweg ist. Für stetige Runge-Kutta Verfahren, die im Rahmen des neuen Ansatzes realisiert sind, wird Konvergenz gezeigt. Ferner wird der Vorteil der Verwendung von Extrapolationen an einem praktischen Beispiel demonstriert.
- Zwei Ansätze zur Berechnung von Vorwärts-Sensitivitäten bei IHDDEs werden untersucht, die sich durch die Reihenfolge der Anwendung des Diskretisierungs- und des Differenzierungsoperators unterscheiden. Im Fall von stetigen Runge-Kutta Verfahren stellt sich heraus, dass Zeitverzögerungen zum Verlust der Kommutativität der beiden Operatoren führen.
- Eine Erweiterung des Konzepts der Internen Numerischen Differentiation für Differentialgleichungen mit Zeitverzögerungen wird vorgeschlagen. Die Anwendung des erweiterten Konzepts stellt sicher, dass die numerisch berechneten Sensitivitäten gegen die exakten Sensitivitäten mit derjenigen Ordnung konvergieren, mit der auch die Nominallösung konvergiert.
- Die ersten praktischen Schemata zur Berechnung von Vorwärts- und Rückwärtssensitivitäten bei IHDDEs werden vorgestellt, die dem Konzept der Internen Numerischen Differentiation folgen. Numerische Untersuchungen zeigen die drastisch höhere Effizienz der entwickelten Schemata im Vergleich zu klassischen Methoden der Sensitivitätsberechnung.
- Die neuen Methoden zur Lösung von IHDDE-IVPs und zur Sensitivitätsberechnung werden auf anspruchsvolle Probleme angewandt. Die Eigenschaften der Methoden werden analysiert.
- Es werden numerische Methoden für die Lösung von IHDDE-beschränkten Parameterschätzproblemen vorgestellt.
- Ein neues IHDDE-Modell der Epidemienausbreitung wird vorgestellt. Hierbei modelliert ein Impuls die Ankunft einer infizierten Bevölkerungsgruppe, und an den Nullstellen zustandsabhängiger Schaltfunktionen werden neue Medikamente bzw. Impfstoffe verfügbar.
- Eine Differentialgleichung mit Zeitverzögerungen zur Beschreibung der Wechselwirkung von zwei Zytokin-Signalkaskaden wird präsentiert. Im Vergleich zu einem Modell ohne Zeitverzögerungen wird einerseits eine bessere Anpassung an experimentelle Daten und andererseits eine Verkleinerung der Anzahl an differentiellen Zuständen erreicht.
- Das Abstimmungsverhalten von Zuschauern der 2012 ausgestrahlten Talentshow "Unser Star für Baku" wird modelliert. Numerische Untersuchungen zeigen, dass eine Zeitverzögerung im Modell wesentlich für eine qualitativ richtige Beschreibung ist. Durch Parameterschätzung wird zudem eine gute quantitive Übereinstimmung zwischen Modell und Daten erreicht.
- Die praktische Realisierung aller entwickelten Methoden in den Softwarepaketen Colsol-DDE und ParamEDE wird beschrieben.

Abstract

This thesis deals with impulsive hybrid discrete-continuous delay differential equations (IHDDEs). This new class of differential equations is highly challenging for two reasons. First, because of a dependency of the right-hand-side function on past states, with time delays that depend on the current state. Second, because both the right-hand-side function and the state itself are discontinuous at implicitly defined time points.

The theoretical results and numerical methods presented in this thesis are related to the following subject areas: First, solutions of initial value problems (IVPs) in IHDDEs. Second, derivatives of IVP solutions with respect to parameters ("sensitivities"). Third, estimation of parameters in IHDDE models from experimental data. Amongst others, this thesis thereby makes the following contributions:

- The theoretical basis of IHDDE-IVPs is established. This includes the definition of a solution concept, the existence of solutions, the uniqueness of solutions, and the differentiability of solutions with respect to parameters.
- A new approach for numerically solving IVPs in differential equations with time delays is introduced. A key aspect is the use of extrapolations beyond past discontinuities. Convergence of continuous Runge-Kutta methods realized in the framework of the new approach is shown, and numerical results are presented that demonstrate the benefit of using extrapolations on a practical example.
- A "first discretize, then differentiate" approach and a "first differentiate, then discretize" approach for forward sensitivity computation in IHDDEs are investigated. It is revealed that the presence of time delays destroys commutativity of differentiation and discretization in the case of continuous Runge-Kutta methods.
- An extension of the concept of Internal Numerical Differentiation is proposed for differential equations with time delays. The use of the extended concept ensures that numerically computed sensitivities converge to the exact sensitivities, and that the convergence order is identical to the convergence order of the method that is used for solving the nominal IVP.
- The first practical forward and adjoint schemes are developed that realize Internal Numerical Differentiation for IHDDEs. Numerical investigations show that the developed schemes are drastically more efficient than classical methods for sensitivity computation.
- The new numerical methods for solving IVPs and for computing sensitivites are successfully applied to several challenging test cases, and the properties of the methods are analysed.
- Numerical methods are presented for solving nonlinear least-squares parameter estimation problems constrained by IHDDEs.
- A new epidemiological IHDDE model is developed. Therein, an impulse accounts for the arrival of an infected population. Further, the zeros of state-dependent switching functions characterize the time points at which new medical treatments become available.
- A delay differential equation model is presented for the crosstalk of the signaling pathways of two cytokines. In comparison to an ordinary differential equation model, a better fit to experimental data is obtained with a smaller number of differential states.
- A novel model is proposed to describe the voting behavior of the viewers of the TV singing competition "Unser Star für Baku" aired in 2012. Numerical results show that the use of a time delay is crucial for a qualitative correct description of the voting behavior. Furthermore, parameter estimation results yield a good quantitative agreeement with data from the TV show.
- The practical implementation of all developed methods in the new software packages Colsol-DDE and ParamEDE is described.

Contents

List of Figures

List of Tables

Introduction

The cooling of a hot cup of tea on the breakfast table, a football flying in mid air toward the goal, the replication of a hepatitis C virus in the cells of an infected host, the oscillations of the powertrain inside a car and a satellite moving in an orbit around the Earth have one thing in common: all these dynamic processes can successfully be described by a certain class of mathematical formulae, so-called *differential equations*. Given an initial state of the system under consideration, a set of differential equations – called the *(mathematical) model* – describes the rate of change of the system and hence the systems' evolution in the future.

Clearly, any of the above given examples requires different quantities as input variables for the differential equation. Depending on the nature of these variables, various subclasses of differential equations are obtained; some well-known classes are, e.g., ordinary differential equations, differential-algebraic equations, and partial differential equations.

From time to time, scientists direct their interest toward differential equations with special features – or combination of features – that have not been studied before. The interest in such "new equations" may be driven by a general theoretical curiosity or by the need to find mathematical models for specific real-world processes. If both is the case, the attention that these equations receive in the scientific community often becomes so great that it is justified to formally define a new subclass of differential equations.

In the past, this has happened for the study of dynamic processes that feature one of the following two characteristics.

Delayed Reactions

One peculiarity of dynamic processes that gave rise to a differential equation class whose study is nowadays a research field in its own right is that the rate of change of a system at the current time may depend on the state of the system in the past. An intuitive example for such a process is given by a car driver that attempts to follow another car in a prescribed distance. Clearly, whenever the leading car accelerates or decelerates, the following car should do likewise. However, the driver of the following car will need some time to react if, say, the red brake light of the leading car lights up. Hence, it is natural to introduce a *time delay* into the mathematical equations describing this system.

The formulation and mathematical analysis of differential equations with time delays dates back at least to the beginning of the 20'th century, see e.g. Schmidt [225], Hilb [146], Fite [106], and references therein. For some early works on real-world systems with time delays, see Callender [56], Rhodes [213], Sievert [237], and Schürer [228]. The interest in time delay systems has increased considerably in the 1950's, as can be concluded from the extensive bibliographies by Weiss [254] in 1959 and Choksy [64] in 1960. The monograph by Bellman and Cooke [28], published in 1963, can certainly be regarded as a very influential work for the study of time delay systems, and this research field has experienced a further and drastic increase of popularity since then.

Abrupt Changes

A second important feature of real-world dynamic processes that has attracted considerable interest by scientists is the effect of abrupt changes. Hereby, "abrupt" is meant in the sense of a *multi-scale problem*, i.e. the state of a system changes only slowly for a long time interval, and then undergoes a drastic change within a very short time interval. For example, a football may be flying through the air for several seconds before it changes its flight direction within milliseconds during a collision with a goal post. For the mathematical description of this system, the specification of two sets of equations is appropriate: one for the flight phase, and one for the collision phase.

Several mathematical models can be developed for the football example, see e.g. the introductory reading by Tolan [248]. In a simplified setting, the motion of the football could be described by the position and velocity of its center of mass, and the encounter with the goal post could be treated as an inelastic collision (i.e. by an immediate change of the football's velocity). Alternatively, one

may also regard the football as a three-dimensional object and describe, by a set of differential equations, how it is flattened during a finite, non-zero time interval of the collision.

The first option for the description of the collision of the ball with the goal post leads to differential equations in which the velocity experiences, at the time of the collision, a so-called *impulse*. Mathematical problems of this kind are the subject of many research works and have been analyzed at least since the 1960's, see Pavlidis and Jury [206], Schmaedeke [224], and Pavlidis [205]. Nowadays the book by Lakshmikantham, Bainov, and Simeonov [169] can be given as a standard reference.

The second option for the description of the ball with the goal post leads, instead, to a *switch* between two different differential equation models: one model for the flight in mid air and one for the duration of the collision. The study of problems with switches also has a long tradition, see e.g. Bocher's paper [34] from 1905, as well as the papers by Meissner [189] and Ziegler [270] in the 1930's, where discontinuous differential equations occur as models for oscillatory systems with friction. Early references for the systematic analysis of the closely related – but more general – class of "discontinuous differential equations" are Filippov [104] and Hájek [129]. A starting point for the study of such equations is the book by Filippov [105].

Modeling with Time Delays, Switches, and Impulses

For a single real-world process it is often possible to find several mathematical models, and the presence of time delays, switches, or impulses in the resulting equations typically depends on the level of abstraction that a modeler wishes to use.

The description of the football hitting the goal post can serve as an example. Here, the first model is "simpler" and "more abstract", because the ball is treated as a point in space in the mathematical equations, whereas the second model is "more elaborate" and "less abstract", because the ball is treated as a three-dimensional object. The collision with the goal post appears "more abrupt" in the first model, because the velocity of the ball is changed immediately. This is not the case in the second model, where "only" the employed set of differential equations – i.e. the rate of change – switches abruptly.

It is quite typical that transitions occur as "less abrupt" in the mathematical equations if more elaborate models are used. Similarly, more elaborate models may also help to avoid the use of time delays in the equations. For illustration, recall the driver of a car that attempts to follow another car. It might, in principle, be possible to find a set of equations that models the central nervous system of the driver of the second car, i.e. all biochemical reactions that take place inside the driver from the time that the red brake light falls on the eye until the muscles in the foot contract and relax in order to push the brake pedal. However, such a model can be regarded as too detailed if only the motion of the car is of interest. For this purpose, it is likely sufficient to take the more abstract viewpoint and use a time delay that represents the reaction time of the driver.

This discussion leads to the conclusion that it is, in many cases, possible to develop differential equation models without time delays and with short, but continuous transitions. However, these models may become overly large and complex and hence, it may not be appropriate to use them. In contrast, mathematical modeling with time delays, switches, and impulses often allows to find much smaller sets of equations that provide a sufficiently good description – and prediction – of the real-world process under consideration. This is the key argument for the theoretical study of differential equations with time delays, switches, and impulses, and also for the development of numerical methods for the approximate solution of these equations.

Mathematical Formalism

After this introductory motivation for the use of time delay, switches, and impulses in mathematical modeling, it is now appropriate to present and discuss the corresponding equations.

Therefore, let $t \in \mathbb{R}$ denote the *time* and let $y(t) \in \mathbb{R}^{n_y}$ be a vector describing the *state* of the dynamic system at the time t. The rate of change of the system is given by the derivative of y with respect to time and is in this thesis denoted by $\dot{y}(t) := dy(t)/dt$. As a basis for the discussion, consider the case that the rate of change $\dot{y}(t)$ is completely described in terms of the current time t and the current state $y(t)$:

$$\dot{y}(t) = f(t, y(t)). \tag{0.1}$$

This is the well-known standard form of an *ordinary differential equation*, with f being the so-called *right-hand-side function*, which is typically assumed to be continuous in both its arguments.

One generalization of equation (0.1) is to introduce a dependency of the right-hand-side function on the value of the state vector in the past, i.e. a dependency on the *past state* $y(t-\tau)$. The function $f(t, y(t))$ would then be replaced by $f(t, y(t), y(t-\tau))$, where τ is called the *time delay*. However, the description of a real-world process may require to consider dependencies on several past states $y(t-\tau_i)$, $1 \leqslant i \leqslant n_\tau$, and the time delays may themselves be time- or state-dependent, i.e. $\tau_i(t, y(t))$. This leads to the following differential equation:

$$\dot{y}(t) = f(t, y(t), \{y(t-\tau_i(t, y(t)))\}_{i=1}^{n_\tau}). \tag{0.2}$$

Another generalization of equation (0.1) is to stick to the arguments t and $y(t)$ of the right-hand-side function f but to drop the assumption of continuity. In a simple setting, the continuity assumption could only be dropped in the time argument, such that between two successive – and known – discontinuity points s_i and s_{i+1} the evolution of the system is still described by an equation of the form (0.1). In a more general setting, discontinuities in f could be allowed to occur along hypersurfaces in the space $\mathbb{R} \times \mathbb{R}^{n_y}$, where each hypersurface is described by an equation $\sigma_i(t, y(t)) = 0$ for $1 \leqslant i \leqslant n_\sigma$. The real-valued functions σ_i are called *switching functions*. Using the signs of the switching functions, $\zeta_i(t) = \operatorname{sign}(\sigma_i(t, y(t)))$, as an argument of the right-hand-side function f yields

$$\dot{y}(t) = f(t, y(t), \zeta(t)), \tag{0.3}$$

with $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))^T$. Any sign change in one of the switching functions σ_i leads to a "switch" in $\zeta_i(t)$ and – depending on the concrete definition of f – potentially to a discontinuity in f or in the partial derivatives of f with respect to its other two arguments. Hence, equation (0.3) represents an important special case of discontinuous ordinary differential equations.

In the earlier discussion of the football example it was mentioned that mathematical modeling of a real-world process with abrupt changes sometimes makes use of an immediate, impulsive change of the state vector. Such impulsive changes could, for example, be applied at a sequence of given discontinuity points s_i. Between two such time points, s_i and s_{i+1}, the system evolves as described by the differential equation (0.1). At the discontinuity points, e.g. at s_i, the *right-sided limit* of the state vector, written as $y^+(s_i) := \lim_{\epsilon \to 0^+} y(s_i + \epsilon)$, is given by

$$y^+(s_i) = y^-(s_i) + \omega(s_i, y^-(s_i)). \tag{0.4}$$

Herein, $y^-(s_i) := \lim_{\epsilon \to 0^+} y(s_i - \epsilon)$ is the *left-sided limit* of the state vector at s_i and ω is the *impulse function*.

The next straightforward extension of a differential equation with impulses is to allow that the impulses do not occur at a sequence of given time points s_i, but that they rather occur whenever the vector $(t, y(t))$ hits a specific hypersurface in $\mathbb{R} \times \mathbb{R}^{n_y}$. The description of the hypersurfaces could, as before, be done by means of switching functions.

Toward a New Class of Differential Equations

Real-world processes may exhibit both delayed reactions and abrupt changes. In order to describe such processes, differential equation models have been proposed in the literature that combine time delays with switches or impulses.

Probably the first model of this class is given in Sievert's paper [237] from 1941, where a differential equation with both a time delay and a switch is used to describe the damaging effect of gamma rays on biological cells. Further instances of differential equations with delays and switches are found in Bock and Schlöder [44], Kolmanovskii and Myshkis [163], Liu, Shen, and Zhang [180], Kim, Campbell, and Liu [159], Sieber et al. [235], and Simpson, Kuske, and Li [238]. Differential equations with delays and impulses are the subject of the works by Das and Sharma [73], Gopalsamy and Zhang [118], Chen, Yu, and Shen [62], Ballinger and Liu [15], Yan, Zhao, and Nieto [265], and Corwin, Thompson, and White [71]. Recently, also a limited number of works have considered differential equations with time delays and switches and impulses, see Wood [259], Li, Ma, and Feng [175], Yang and Zhu [266], Liu, Liu, and Xie [179], and Schnute, Couture-Beil, and Haigh [226].

The above list of references allows to conclude that the study of differential equations that incorporate time delays, switches, and impulses has reached a certain level of popularity in the scientific community. However, to the best knowledge of the author of this thesis, no general problem formulation has yet been proposed that allows for all of the following properties:

- The right-hand-side function depends on multiple past states, and the delay functions are time- and state-dependent.
- Switches and impulses occur at time points that are determined implicitly, i.e. they are characterized as zeros of state-dependent switching functions.
- The switching functions and impulse functions depend on past states.
- The right-hand-side function f, the switching functions σ_i, and the delay functions τ_i are nonlinear functions of their arguments.

The first contribution of this thesis is thus to formulate a general class of differential equations that covers all these aspects. For the *initial value problem* (IVP) corresponding to this newly established class of equations, this thesis then presents comprehensive theoretical results, computational methods, software, real-world applications and numerical investigations as outlined below. Several of the findings of this thesis are thereby novel also in the context of simpler subclasses of differential equations, in particular for differential equations with several state-dependent delays (see equation (0.2)).

Contributions of This Thesis and Related Work

This thesis is concerned, as mentioned before, with a class of differential equations and a corresponding class of IVPs that has not been formulated in this generality so far. Hence, it is obvious that this thesis is also the first work that is concerned with the theory of or with numerical methods for these very general equations.

There is, however, a wealth of existing literature that deals with simpler subclasses of differential equations and IVPs. In the following summary of the contributions of this thesis, only those works are cited that have a very immediate connection to the presented work. Many more references to other related papers and theses are given in the introductions to the individual chapters, in particular to the Chapters 4, 5, 6, 7, 8, 9, and 13.

Theoretical Foundations

For IVPs in differential equations with switches, e.g. equations of the form (0.3), the *classical* definition of a solution is unsuitable: It is evident that there might be no differentiable function $y(t)$ that fulfills the differential equation everywhere if there are time points where the right-hand-side function is allowed to change discontinuously. Hence, the first step of any theoretical analysis for these kind of equations is to define a more general notion of a solution, and many approaches have been made to address this issue, see Cortés [69].

In fact, a generalization of the concept of a solution is necessary for all classes of differential equations with discontinuities in the right-hand-side function f or in the state vector y. The problem class considered in this thesis gives rise to potentially many discontinuities: At all time points, where at least one of the switching functions becomes zero, because each zero may trigger discontinuities in either f or y (or both), and in addition at all time points where at least one *deviating argument* $t - \tau_i(t, y(t))$ crosses a discontinuity point in the past. Since this thesis is the first work to consider a new and very general class of differential equations, a simple generalization of the solution notion is employed. More precisely, a solution in the spirit of this thesis is a function $y(t)$ that fulfills the differential equation almost everywhere, with the exceptional set consisting of a finite number of time points.

With a definition of a solution at hand, three important topics in the theoretical analysis of IVPs in differential equations are addressed in this work: existence of solutions, uniqueness of solutions, and differentiability of solutions with respect to *parameters*. The parameters may thereby occur in the right-hand-side function, in the delay functions, in the switching functions, in the impulse function, and in the initial data of the IVP. The discussion of existence, uniqueness, and differentiable dependence is done separately for differential equations without state dependencies in the switching and delay functions, and for differential equations that feature such state dependencies.

For IVPs in differential equations without state dependencies in the switching and delay functions, existence, uniqueness, and differentiability results are given. The key idea of the proofs is to use the *method of steps*. With regard to existence and uniqueness of classical solutions of IVPs in differential equations of the form (0.2), this method has frequently been used, see e.g. El'sgol'ts and Norkin [92] or Smith [239]. In this thesis, it is shown how the method can also be used to obtain existence and uniqueness of solutions in a more general sense and in the context of more general IVPs. The method of steps can further be used to prove differentiability of IVP solutions with respect to parameters in the context of differential equations with constant delays, as shown very recently by Lenz, Schlöder, and Bock [173]. This thesis builds upon this work and contains a thorough discussion of sufficient differentiability conditions for the case of differential equations with time-dependent delays, also in combination with time-dependent switching functions and impulses.

For IVPs in differential equations with state dependencies in the switching or delay functions, a set of new definitions is introduced that allows to formulate, in a very concise way, sufficient conditions for the uniqueness of a given IVP solution. Furthermore, differentiability of IVP solutions with respect to parameters is discussed. The basis for this discussion is a differentiability theorem given in Bock [39] and Galán, Feheery, and Barton [111], which applies to IVPs in differential equations with switches and impulses. Here, the idea behind this theorem is transferred to differential equation that exhibit, in addition, time delays. As a special case, a theorem is presented for differentiability of solutions of IVPs in differential equations of the form (0.2), which allows that discontinuities are present in the initial function. Contrariwise, all previously established differentiability theorems in the context of equation (0.2) known to the author – e.g. Hartung et al. [137] and Hartung [135] – assume continuity of the initial function.

Numerical Methods and Their Analysis

For the numerical solution of IVPs in differential equations with time delays, the so-called *standard approach* – see e.g. Bellen and Zennaro [26] – relies on methods that provide a continuous approximation of the solution. In the standard approach, the computation of past states is carried out by evaluating either the initial function or the continuous approximation of the IVP solution provided by the numerical method. This standard approach has one disadvantage: If the current integration step is such that a deviating argument crosses a discontinuity point in the past, then the smoothness assumptions of the employed numerical method are typically violated.

As a remedy, extrapolations beyond past discontinuities have been employed in REBUS by Bock and Schlöder [43, 44] and in RADAR5 by Guglielmi and Hairer [124], and detailed descriptions of the use of extrapolations can be found in ZivariPiran [271], ZivariPiran and Enright [272], and Ernst [101]. To date, however, there exists no theoretical basis for the use of extrapolations beyond past discontinuity points. In this thesis, the use of extrapolations is part of the formal definition of the *modified standard approach*, and the properties of this approach are analyzed. More precisely, novel well-posedness and convergence results are presented for continuous Runge-Kutta methods in the framework of the modified standard approach. In addition, it is observed that a previously given proof for the convergence of numerical methods realized in the framework of the standard approach (see Bellen and Zennaro [26]) does not hold. Hence, only the modified standard approach has a rigorous theoretical basis.

Several contributions of this thesis are related to the numerical computation of *sensitivities* in the context of differential equations with time delays, i.e. of the derivatives of IVP solutions with respect to parameters. More precisely, continuous Runge-Kutta methods are regarded and it is investigated how the numerical integration of a suitably-defined *variational initial value problem* (a "first differentiate, then discretize" approach) – as proposed, e.g., by ZivariPiran and Enright [273] – relates to the differentiation of the integration scheme that was used for the solution of the original IVP (a "first discretize, then differentiate" approach). The convergence properties of both approaches are analyzed, and based on this analysis an extension of the principle of Internal Numerical Differentiation (see Bock [36, 38]) for differential equations with time delays is proposed.

In the case that there is a large number of parameters in the IVP whose sensitivities are of interest, it is known that so-called *adjoint* (or "backward") methods for sensitivity computation are more efficient, see e.g. Bock [39], Albersmeyer and Bock [3]. This thesis introduces a new numerical method for the computation of adjoint sensitivities in the context of differential equations with time delays. The presented method relies on the development of a *discrete adjoint scheme*, which exhibits the same convergence properties as the corresponding forward scheme for sensitivity

computation. The proposed discrete adjoint scheme can also be used to compute the sensitivities of the state at time points that are not part of the mesh of the integration method.

A frequent situation in scientific and engineering applications is that a differential equation model is available for a particular process, but that some of the parameters in the model are unknown and have to be estimated from experimental data. This leads to so-called *parameter estimation problems*. Under certain standard assumptions, a maximum likelihood estimate for the parameters is given by the solution of an infinite-dimensional least-squares optimization problem, in which the differential equation occurs as an infinite-dimensional equality constraint. The task considered in this thesis is to estimate parameters in differential equations with time delays, switches, and impulses.

As a solution approach, the single shooting parameterization is employed in order to reduce the problem to finite dimension. Furthermore, inspired by the works of Bock [36, 38, 39] and Bock, Kostina, and Schlöder [41, 42], a damped Generalized Gauss-Newton method based on the restrictive monotonicity test for solving the resulting finite-dimensional nonlinear constrained least-squares problem is proposed and realized.

Software

This thesis contains a detailed description of the numerical methods that are employed in Colsol-DDE, a novel software package for solving IVPs in differential equations that exhibit time delays, switches, and impulses. In contrast to the few existing solvers for this purpose – e.g. DDE_SOLVER by Thompson and Shampine [246] and Solv95 by Wood [259] – Colsol-DDE is based on implicit methods (more precisely, implicit Runge-Kutta methods) and is thus suitable for stiff IVPs.

Furthermore, Colsol-DDE is the first program that allows the computation of forward and adjoint sensitivities for the considered very general class of IVPs by means of Internal Numerical Differentiation. In particular, this makes Colsol-DDE also the first solver that provides adjoint sensitivities for simpler subclasses of differential equations, e.g. for those that feature only switches or only impulses or only time delays. It should further be noted that also the automated computation of forward sensitivities of IVP solutions in the context of differential equations with time delays is rarely found in existing solvers. To the knowledge of the author, only DDEM by ZivariPiran [271] includes this feature, but – in contrast to Colsol-DDE – this code is unsuitable for stiff IVPs.

This thesis further presents the numerical methods realized ParamEDE, which is the first software that solves nonlinear constrained least-squares parameter estimation problems in differential equations with time delays, switches, impulses. It makes use of several sophisticated numerical techniques that go back to Bock [36, 38, 39], and Bock, Kostina, and Schlöder [41, 42]. The use of the Generalized Gauss-Newton method as basic optimization method ensures that ParamEDE converges only to those solutions that are statistically stable against small perturbations of the measurement data. ParamEDE further makes use of the restrictive monotonicity test, which is a particularly well-suited strategy for globalizing the convergence of (Generalized) Gauss-Newton methods. An internal regularization is implemented in ParamEDE in order to deal with singular or ill-conditioned problems. Last but not least, Colsol-DDE is used as underlying IVP solver in ParamEDE, which guarantees an efficient and accurate computation of the required sensitivities by Internal Numerical Differentiation.

Both Colsol-DDE and ParamEDE are designed to be used in conjunction with the Automatic Differentiation tool Tapenade (see Hascoët and Pascual [140, 141]). In particular, the combined use of Internal Numerical Differentiation and Automatic Differentiation in Colsol-DDE and ParamEDE allows to compute sensitivities with very high accurarcy, and, in addition, makes the codes entirely derivative-free for the user.

Mathematical Models for Applications

This thesis introduces new mathematical models for three applications.

Epidemiology is considered as one application area in which the use of time delays is very popular. A model by Cooke and van den Driessche [68] is considered, which involves two time delays representing the latency period of the disease and the immunization period after recovery from an infection. For this model, three extensions are proposed, which demonstrate exemplarily how impulses and switches can be used in epidemiology. The extensions allow for an invasion of a healthy population by an infected population, and for the development of a new drug and a

vaccine a certain time after the total number of casualties due to the disease has reached a given threshold.

A second model developed in this thesis describes the interaction between the signaling pathways of two cytokines, Interleukin-6 (IL-6) and granulocyte macrophage colony-stimulating factor (GM-CSF). The work presented here is based on an ordinary differential equation model of Sommer et al. [240]. By using time delays in the model, the size of the differential equation model is reduced, and hence a more concise mathematical description of the process is obtained.

Eventually, a model is proposed for the voting behavior of the viewers of the German TV singing competition "Unser Star für Baku" aired in 2012. In this TV show, the viewers could vote for their favorite candidates by phone calls or SMS, and they were permanently able to see – in a so-called *livescore* – the percentages of votes that the candidates had received so far. Hence, the viewers could make their voting behavior dependent on the intermediate results displayed in the livescore. In this thesis, a differential equation model is developed for the voting behavior of the TV viewers that incorporates both switches and a time delay. The switches are thereby motivated by the assumption that the viewers vote differently for candidates that are currently winning (i.e. they would be allowed to return in the next episode of the show) and other candidates that are currently losing (i.e. they would have to leave the competition). The time delay accounts, inter alia, for encryption and decryption processes in digital broadcasting and for the time that viewers need to dial the number of their favorite candidate.

Numerical Results

Only very few research works provide reference solutions and/or reference sensitivities of IVPs in differential equations with time delays (Paul [202], ZivariPiran [271]). Reference solutions of IVPs in differential equations with both time delays and switches (or with both time delays and impulses) are also rarely available (to the knowledge of the author, only in Corwin, Thompson, and White [71]), and reference sensitivities for these classes of differential equations are not available at all. In this thesis, several challenging IVPs are formulated, and accurate reference values are provided for the solution and for the sensitivities.

The performance of the numerical methods implemented in Colsol-DDE is assessed. In particular, the convergence of numerically computed solutions and sensitivities to the corresponding reference values is investigated in the limit of small relative tolerances. Further, the performance of the methods on a stiff IVP is investigated.

It is shown in this thesis how discontinuity location works with the modified standard approach. In particular, it is demonstrated on a practical example that the use of extrapolations beyond past discontinuities is beneficial for an efficient localization of discontinuity points.

This thesis further presents numerical investigations that compare the newly developed Internal Numerical Differentiation approach for differential equations with time delays to two classical approaches for sensitivity computation. First, a comparison to finite difference sensitivity computation (so-called "External Numerical Differentiation") reveals that Internal Numerical Differentiation can provide more accurate sensitivities at only 20% of the computation time. Second, a comparison to the numerical solution of the combined nominal and variational IVP yields the result that Internal Numerical Differentiation provides the same accuracy at only 1% of the computation time.

With regard to parameter estimation problems in differential equations with time delays, an important issue is the possible non-smoothness of the considered optimization problems, see Baker and Paul [13]. A non-smooth dependence of the objective function on the unknown parameters raises the suspicion that derivative-based optimization methods – such as the Gauss-Newton method – may show a very poor convergence behavior. More precisely, the local contraction theorem of Bock [39] guarantees convergence only if there are no discontinuities or non-differentiabilities within a ball in parameter space that contains both the initial guesses for the parameters and the solution of the optimization problem. However, in this thesis, a practical non-smooth problem is considered, and it is observed that convergence is obtained in the numerical practice even in situations where convergence is not guaranteed by the convergence theory.

Last but not least, parameter estimation results are presented for two applications for which real-world experimental data are available, namely for the crosstalk of the signaling pathways of IL-6 and GM-CSF, and for the voting behavior of the viewers of the TV show "Unser Star für Baku". For the first application, the new model with time delays yields a better fit to experimental data than the ordinary differential equation model by Sommer et al. [240], despite the fact that the

new model is "simpler" in the sense that it consists of a smaller number of differential states. For the second application, a very good agreement is found between the model and the data from the TV show. Further analysis reveals that the time delay is crucial for explaning the voting behavior of the TV viewers qualitatively, and that the size of the fan-bases of the candidates, i.e. their popularity, has a smaller influence on the outcome of the voting procedure than the time delay.

Outline of The Thesis

Part I introduces basic problems, definitions, and concepts, and is subdivided into three chapters.

In Chapter 1, a general class of differential equations is introduced that comprises the three features time delays, switches, impulses. This new class of differential equations is called *impulsive hybrid discrete-continuous delay differential equations* (IHDDEs), whose study is the subject of this thesis. Furthermore, Chapter 1 introduces an initial value problem in IHDDEs (shortly: IHDDE-IVP), and introduces a consistent terminology for various classes of differential equations that are identified as subclasses of IHDDEs.

Chapter 2 is mainly devoted to the issue of defining a concept of a "solution" of an IHDDE-IVP. In particular, the concept of solution used in this thesis is motivated and put into relation to alternative concepts. Chapter 2 further introduces a classification for the various sources of discontinuities in IHDDE-IVP solutions.

Chapter 3 presents new models for three applications. The considered real-world problems are: the spread of an epidemic within a population; the crosstalk of the signaling pathways of IL-6 and GM-CSF; and the voting behavior of the viewers of the TV show "Unser Star für Baku".

Part II deals with IHDDE-IVP solutions and is subdivided into three chapters.

Chapter 4 presents the existence and uniqueness theory for IVP solutions. For differential equations with constant delays and explicitly known time points of discontinuity in the right-hand-side function, a theorem on existence and uniqueness is formulated and formally proven by relying on the so-called method of steps. It is furthermore discussed how this result can be generalized to all IHDDE-IVPs in which the switching functions and the delay functions do not depend on the unknown IVP solution itself. For IVPs where the switching functions and delay function depend on the unknown solution, sufficient conditions are given under which a given IVP solution can be shown to be unique.

Chapter 5 introduces the modified standard approach as a new concept for solving IVPs in differential equations with time delays. The modified standard approach makes use of extrapolations beyond past discontinuity points. The properties of continuous Runge-Kutta methods realized in the framework of the modified standard approach are investigated. In particular, well-posedness of the numerical method and convergence to the exact solution are shown.

Chapter 6 presents the numerical methods that are employed in the new IHDDE-IVP solver Colsol-DDE. In particular, it is discussed how implicit continuous Runge-Kutta methods of collocation type are combined with an implicit uniform correction procedure and with an implicit quadrature rule in order to obtain error-controlled discrete and continuous approximations of the IVP solution. Special emphasis is put on the numerical treatment of discontinuities in Colsol-DDE.

Part III deals with the sensitivity of IHDDE-IVP solutions with respect to parameters. It is subdivided into three chapters.

Chapter 7 presents the theory of differentiability of IVP solutions with respect to parameters. A theorem on the differentiability of IVP solutions in differential equations with constant time delays is given, and it is discussed how this result can be extended to all IHDDEs in which the switching functions and the delay functions do not depend on the state vector. Furthermore, for problems where the switching functions and/or delay functions depend on the state vector, sufficient conditions are presented under which a given IVP solution is differentiable.

Chapter 8 deals with the numerical computation of sensitivities. This chapter introduces an extension of the concept of Internal Numerical Differentiation for differential equations with time delays. In the context of continuous Runge-Kutta methods, Internal Numerical Differentiation methods for the computation of forward and adjoint sensitivities are presented. Since the sensitivities are, in general, only piecewise smooth functions in time, the expressions for the jumps in the sensitivities are discussed in detail.

Chapter 9 discusses the numerical methods for sensitivity computation that are implemented in the new IHDDE-IVP solver Colsol-DDE. In particular, the extended principle of Internal Numerical Differentiation for differential equations with time delays is applied to the methods that have been

presented in Chapter 6 before. Special attention is paid to the subtleties of the implementation of a discrete adjoint scheme for sensitivity computation.

Part IV is concerned with the estimation of parameters from experimental data and is subdivided into four chapters.

Chapter 10 recalls the concept of maximum likelihood parameter estimation. For the important special case of normally distributed measurements with known covariance, a maximum likelihood estimate is given by the solution of a nonlinear least-squares minimization problem. The chapter also contains the fundamentals of optimization theory.

Chapter 11 recalls the Generalized Gauss-Newton method for solving nonlinear constrained least-squares problems. A fundamental convergence result for this method is given, and it is discussed why the Generalized Gauss-Newton method is well-suited for solving least-squares parameter estimation problems, in particular in view of the statistical stability of the obtained solution. The chapter further discusses an extension that makes the Generalized Gauss-Newton method applicable to ill-posed problems, and a stepsize selection strategy that aims at globalizing the convergence of the method.

Chapter 12 deals with the randomness of the result of maximum likelihood estimation as a consequence of the randomness in the data. The focus of the chapter lies on recalling techniques for the statistical analysis of estimated parameters.

Chapter 13 discusses the special challenges that are associated with the estimation of parameters in differential equations, and, in particular, in IHDDEs. IHDDE-constrained parameter estimation problems are identified as infinite-dimensional optimization problems, and a finite-dimensional parameterization is presented. Theoretical and numerical issues in non-smooth optimization problems are discussed. Eventually, a practical algorithm for parameter estimation in IHDDEs is presented, and the implementation of this algorithm in the software ParamEDE is discussed.

Part V contains the results of numerical investigations and is subdivided into three chapters.

Chapter 14 presents numerical results related to the solution of IVPs. Reference values for the solutions of several challenging IVPs in differential equations with time delays are given, partially in combination with both switches and impulses. The convergence of the methods realized in Colsol-DDE in the limit of small relative tolerances is investigated. It is demonstrated how localization of discontinuity points works with the modified standard approach, and a comparison to the use of the standard approach is made. Further, a simulation study is presented that allows an assessment of the influence of the parameters in the voting behavior of the viewers of the TV singing competition "Unser Star für Baku".

Chapter 15 contains numerical results for the computation of sensitivities of IVP solutions with respect to parameters. Reference values for the sensitivities are given for several IVPs in differential equations with time delays, switches, and impulses. Further, the convergence of the methods for sensitivity computation realized in Colsol-DDE is investigated in the limit of small relative tolerances. The newly developed Internal Numerical Differentiation approach for differential equations with time delays is compared to External Numerical Differentiation and to a combined solution of nominal and variational IVP. Furthermore, different realizations of Internal Numerical Differentiation are compared to each other.

Chapter 16 presents a numerical investigation of the convergence behavior of the Gauss-Newton method applied to a non-smooth least-squares optimization problem in the context of a differential equation with a time delay and with switches. Furthermore, parameter estimation results for two applications with real-world experimental data are given: the crosstalk of the signaling pathways of IL-6 and GM-CSF, and the voting behavior of the viewers of "Unser Star für Baku".

In the final chapter "Summary & Outlook" of this thesis, the main contributions of this thesis are summarized, and some suggestions for future research are made.

It should be noted that the author sets value on a comprehensive presentation of the treated topics. For this reason, textbook knowledge is incorporated into this thesis when appropriate. In particular, this is the case in the Chapters 5 and 6 (recap of standard methods for numerical solution of IVPs in ordinary differential equation and in delay differential equations), and in the Chapters 10-12 (recap of basic parameter estimation theory, of tailored numerical methods for parameter estimation, and for the statistical analysis of solutions).

Fonts

In this thesis, *italic letters* are used to introduce new terms that are used throughout the thesis. Alternative terms that have been used in the literature, which are not adopted here, are indicated by "quotation marks". For emphasis, underlined expressions are used. These conventions have already been used in this introduction.

The nomenclature in this thesis is as follows: Boldface letters $\mathbf{A}, \mathbf{B}$, etc., are used for matrices and matrix-valued functions. Calligraphic letters, e.g. $\mathcal{D}$ and $\mathcal{T}$, are used to represent sets and intervals. Script letters are used to denote function spaces, e.g. $\mathscr{C}$ represents the space of continuous functions. In addition, this font is used for the Landau symbol $\mathscr{O}$, and for the normal distribution, which is denoted by $\mathscr{N}$.

Blackboard bold, e.g. $\mathbb{N}$, $\mathbb{R}$, is used – as customary – to denote the sets of all integer and real numbers. In addition, this font is used to denote the expectation (symbol $\mathbb{E}$) and the variance (symbol $\mathbb{V}$) of random numbers. Finally, gothic-type letters (e.g. $\mathfrak{y}$, $\mathfrak{W}$) are used for in the definition of an IVP if it is desirable to reserve the symbols y and $\mathbf{W}$ exclusively for the solution of the IVP. In addition, gothic-type letters $\mathfrak{e}$ and $\mathfrak{h}$ are used as symbols for random variables.

Part I.

Impulsive Hybrid Discrete-Continuous Delay Differential Equations

1. Considered Problem Class

> *Despite this very satisfactory state of affairs as far as differential equations are concerned, we are nevertheless forced to turn to the study of more complex equations.*
>
> Bellman and Cooke, in the introduction to their book "Differential-Difference Equations" [28]

As discussed in the introduction, real-world dynamic processes with time-delayed reactions or abrupt changes are often appropriately modeled by differential equations that feature time delays, switches, or impulses. In several research works, differential equation models have been proposed that combine time delays with either switches or impulses, or even combine them with both switches and impulses, see e.g. Ballinger and Liu [15] Corwin, Thompson, and White [71], Li, Ma, and Feng [175], Yang and Zhu [266], and Liu, Liu, and Xie [179].

However, to the best knowledge of the author, this thesis is the first research work that allows all of the following properties: multiple delays that depend on the time and on the current state; past states that occur as arguments not only in the right-hand-side function but also in the switching and impulse functions; switches in the differential equation and impulses occur at time points that are implicitly determined as functions of the state itself; and general nonlinear dependencies of all model functions on their arguments.

The topic of this chapter is the formulation of this new and very general class of differential equations, as well as of the corresponding *initial value problem* (IVP).

Organization of This Chapter

The definition of the new class of differential equations, together with the introduction of the necessary notation, is the subject of Section 1.1. Section 1.2 discusses subclasses of the general problem, reviews existing terminology for these subclasses and introduces the terminology of this thesis. Sections 1.3 and 1.4 define the terminology for special properties of the switching and delay functions.

1.1. Problem Definition

In the following, the properties of differential equation models as discussed in the introduction – switches, impulses, and time delays – are combined in a class of equations that has not been formulated or studied previously.

As a starting point, equation (0.3) is recalled:

$$\dot{y}(t) = f(t, y(t), \zeta(t)). \tag{1.1}$$

Herein, $t \in \mathbb{R}$ is the time, $y(t) \in \mathbb{R}^{n_y}$ is the *state* of the system at the time t, and the time derivative $\dot{y}(t) = dy(t)/dt$ represents the rate of change of the state. Further, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$ represents the signs of *switching functions* σ_i, i.e. $\zeta_i(t) = \operatorname{sign}(\sigma_i(t, y(t)))$ for $1 \leqslant i \leqslant n_\sigma$. The purpose of the switching functions is that changes in their sign – i.e. time points of discontinuity in $\zeta_i(t)$ – characterize discontinuous changes of the *right-hand-side function* f. Accordingly, for any given vector $\zeta(t)$, the right-hand-side function f is assumed to be a continuous (or even smooth) function of its other two arguments, and also the switching functions σ_i are supposed to be at least continuous.

Throughout the thesis, functions $y(t)$ will be considered as "solutions" of a differential equation if the switching functions become zero only at isolated time points (see the formal definition of a solution concept in Chapter 2). Having this in mind, the differential equation (1.1) is rewritten as

follows:

$$\dot{y}(t) = f(t, y(t), \zeta(t)) \quad \text{if} \quad \zeta_i(t) \neq 0 \quad \text{for all} \quad i \in \{1, \ldots, n_\sigma\}. \tag{1.2}$$

However, this equation does not impose any condition on $y(t)$ at times t where $\zeta_i(t)$ for at least one i. In particular, not even continuity of $y(t)$ is imposed. Therefore, in order to express that the equation is non-impulsive, it is appropriate to reformulate the differential equation as

$$\dot{y}(t) = f(t, y(t), \zeta(t)) \quad \text{if} \quad \zeta_i(t) \neq 0 \quad \text{for all} \quad i \in \{1, \ldots, n_\sigma\} \tag{1.3a}$$
$$y(t) = y^+(t) = y^-(t) \quad \text{else.} \tag{1.3b}$$

Herein, $y^+(t)$ and $y^-(t)$ are the *right-sided limit* and the *left-sided limit* of the state y at time t:

$$y^+(t) := \lim_{\epsilon \to 0^+} y(t + \epsilon) \tag{1.4a}$$
$$y^-(t) := \lim_{\epsilon \to 0^+} y(t - \epsilon). \tag{1.4b}$$

In the next step, impulses shall be included into the problem formulation. In equation (0.4), the impulse at a time point s_i was formulated as

$$y^+(s_i) = y^-(s_i) + \omega(s_i, y^-(s_i)). \tag{1.5}$$

Herein, ω is the *impulse function*. In general, impulses may occur at time points that are implicitly determined by the time evolution of $y(t)$ itself, which suggests to characterize their location by zeros of switching functions, i.e. in the same way as the discontinuities of the right-hand-side function were determined in equation (1.3a). It is then also natural that the impulse depends on the signs of the switching functions:

$$y^+(s_i) = y^-(s_i) + \omega(s_i, y^-(s_i), \zeta(s_i)). \tag{1.6}$$

Herein, $\zeta(s_i)$ should – of course – represent the signs of the switching functions before the impulse is applied. This is necessary to note because the impulse applied to the state will generally also cause an immediate change in the signs of the switching functions.

The formal definition of a "solution", which follows in Chapter 2, will use right-continuituous functions $y(t)$, i.e.:

$$y(s_i) := y^+(s_i). \tag{1.7}$$

Therefore, in order to express that $\zeta(s_i)$ in equation (1.6) represents, as intended, the switching function signs to the left of s_i, the functions $\zeta_i(t)$, $1 \leqslant i \leqslant n_\sigma$, are defined as

$$\zeta_i(t) := \operatorname{sign}(\sigma_i(t, y^-(t))). \tag{1.8}$$

This makes it possible to define differential equations that account for both switches and impulses as follows:

$$\dot{y}(t) = f(t, y(t), \zeta(t)) \quad \text{if} \quad \zeta_i(t) \neq 0 \quad \text{for all} \quad i \in \{1, \ldots, n_\sigma\} \tag{1.9a}$$
$$y(t) := y^+(t) = y^-(t) + \omega(t, y^-(t), \zeta(t)) \quad \text{else.} \tag{1.9b}$$

It remains to include time delays into the problem formulation. In view of equation (0.2), the differential equation (1.9a) is modified in the following way:

$$\dot{y}(t) = f(t, y(t), \{y(t - \tau_i(t, y(t)))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \text{if} \quad \zeta_i(t) \neq 0 \quad \text{for all} \quad i \in \{1, \ldots, n_\sigma\}. \tag{1.10}$$

Herein, $\tau_i(t, y(t))$ for $1 \leqslant i \leqslant n_\tau$ are called the *delay functions*, $\alpha_i(t, y(t)) := t - \tau_i(t, y(t))$ are called the *deviating arguments*, and $y(t - \tau_i(t, y(t)))$ are called the *past states*.

If time delays are included in the problem formulation by letting the right-hand-side function depend on past states, it is only natural to allow that also the switching and impulse functions depend on past states. However, while t is approaching a zero s of a switching function, it is possible that one or several deviating arguments approach a time point of discontinuity s_{past} in

the past. Therefore, σ_i and ω should be evaluated by using one-sided limits of the state at past time points (in analogy to the one-sided limit $y^-(t)$ of the current state used in equations (1.8) and (1.9b)). For the past states, however, it is reasonable to take the left-sided or the right-sided limit of a past state at a discontinuity point depending on the behavior of $t - \tau_i(t, y(t))$, $1 \leqslant i \leqslant n_\tau$, for t in a left neighborhood of s. More precisely, if the deviating argument approaches s_{past} from the left (from the right) while t approaches s, then the left-sided limit (the right-sided limit) should be taken.

In order to express this in terms of equations, the following is defined:

$$l_i^\alpha(t) = \lim_{t' \to t^-} \operatorname{sign}\big(\alpha_i(t', y(t')) - \alpha_i(t, y^-(t))\big). \tag{1.11}$$

It is thereby assumed that α is a continuous function of its arguments in order to guarantee that the limit exists.

If $\alpha_i(t', y(t'))$ is smaller (greater) than $\alpha_i(t, y^-(t))$ for t' in a left neighborhood of t, then the sign is negative (positive) for t' to the left of t, and thus $l_i^\alpha(t) = -1$ ($l_i^\alpha(t) = +1$). Further, if $\alpha(t', y^-(t'))$ is constant in a left neighborhood of t, then $l_i^\alpha(t) = 0$.

The quantity $l_i^\alpha(t)$ enters the following definition:

$$y^\bullet(\alpha_i(t, y^-(t))) := \begin{cases} y^-(\alpha_i(t, y^-(t))) & \text{if} \quad l_i^\alpha(t) = -1 \\ y^+(\alpha_i(t, y^-(t))) & \text{if} \quad l_i^\alpha(t) = 0 \quad \text{or} \quad l_i^\alpha(t) = +1. \end{cases} \tag{1.12}$$

The fact that the right-sided limit $y^+(\alpha_i(t, y^-(t)))$ is used for $l_i^\alpha(t) = 0$ is motivated by the fact that right-continuous functions $y(t)$ are considered as solutions (see Chapter 2).

Having introduced the notation $y^\bullet$, the generalizations of the equations (1.8) and (1.9b) for arbitrary continuous delay functions τ_i become

$$\zeta_i(t) := \operatorname{sign}(\sigma_i(t, y^-(t), \{y^\bullet(t - \tau_i(t, y(t)))\}_{i=1}^{n_\tau})) \tag{1.13}$$

and

$$\begin{aligned} y(t) := y^+(t) = y^-(t) + \omega(t, y^-(t), \{y^\bullet(t - \tau_i(t, y(t)))\}_{i=1}^{n_\tau}, \zeta(t)) \\ \text{if} \quad \zeta_i(t) = 0 \quad \text{for at least one} \quad i \in \{1, \dots, n_\sigma\}. \end{aligned} \tag{1.14}$$

The equations (1.10), (1.13), and (1.14) form a new class of differential equations, whose study is the subject of this thesis. For a more compact notation, it is suitable to define the set of all possible values of $\zeta(t)$ as

$$\mathcal{I}^\zeta := \{-1, 0, 1\}^{n_\sigma}. \tag{1.15}$$

Further, let

$$\mathcal{I}_0^\zeta := \{\zeta \in \mathcal{I}^\zeta \mid \zeta_j = 0 \text{ for at least one } j \in \{1, \dots, n_\sigma\}\} \tag{1.16a}$$

$$\mathcal{I}_1^\zeta := \{\zeta \in \mathcal{I}^\zeta \mid \zeta_j \neq 0 \ \forall\ j \in \{1, \dots, n_\sigma\}\}. \tag{1.16b}$$

It holds that $\mathcal{I}_0^\zeta \cup \mathcal{I}_1^\zeta = \mathcal{I}^\zeta$ and $\mathcal{I}_0^\zeta \cap \mathcal{I}_1^\zeta = \varnothing$. These definitions allow to introduce, for a given function $y(t)$, $t \in \mathcal{T}$ with $\mathcal{T}$ being some interval, the following sets:

$$\mathcal{D}_0^t(\mathcal{T}) := \{t \in \mathcal{T} \mid \zeta(t) \in \mathcal{I}_0^\zeta\} \tag{1.17a}$$

$$\mathcal{D}_1^t(\mathcal{T}) := \{t \in \mathcal{T} \mid \zeta(t) \in \mathcal{I}_1^\zeta\}. \tag{1.17b}$$

Evidently, $\mathcal{D}_0^t(\mathcal{T}) \cup \mathcal{D}_1^t(\mathcal{T}) = \mathcal{T}$ and $\mathcal{D}_0^t(\mathcal{T}) \cap \mathcal{D}_1^t(\mathcal{T}) = \varnothing$.

With these notations, it is possible to express the equations (1.10), (1.14) in the following, more compact form:

$$\dot{y}(t) = f(t, y(t), \{y(t - \tau_i(t, y(t)))\}_{i=1}^{n_\tau}, \zeta(t)) \qquad \text{for} \quad t \in \mathcal{D}_1^t(\mathbb{R}) \tag{1.18a}$$

$$y(t) = y^+(t) = y^-(t) + \omega(t, y^-(t), \{y^\bullet(t - \tau_i(t, y^-(t)))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_0^t(\mathbb{R}). \tag{1.18b}$$

Please note that the two formulations (1.10), (1.14) and (1.18a), (1.18b) are equivalent. Both

notations will be used in this thesis. The former one is employed when applications are considered in Chapter 3 and in Part V, whereas the latter is used in the other parts of this work.

Equations of the form (1.18a), (1.18b), (1.13) are called *impulsive hybrid discrete-continuous delay differential equations*, and the study of such equations is the subject of this work. However, mathematical models of real-world processes often contain *parameters* that are unknown or only vaguely determined. If their influence on the behavior of the state should be investigated, the first step is to make these unknown parameters "visible" in the equations. Therefore, the unknown parameters are denoted by c and are included into the right-hand-side function, into the delay functions, into the switching functions, and into the impulse functions. This leads to the following formal definition.

Definition 1.1 (Impulsive Hybrid Discrete-Continuous Delay Differential Equation (IHDDE))

An Impulsive Hybrid discrete-continuous Delay Differential Equation (IHDDE) *is an equation, in which the state y as a function of the time $t \in \mathbb{R}$ is determined by both a differential equation and by impulses*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_1^t(\mathbb{R}) \tag{1.19a}$$

$$y(t) = y^+(t) = y^-(t) + \omega(t, y^-(t), c, \{y^\bullet(t - \tau_i(t, y^-(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_0^t(\mathbb{R}). \tag{1.19b}$$

The state y is a function $y : \mathbb{R} \to \mathcal{D}^y$, $\mathcal{D}^y \subset \mathbb{R}^{n_y}$, and $c \in \mathcal{D}^c \subset \mathbb{R}^{n_c}$ are parameters. Further, $y^-(t)$ and $y^+(t)$ denote the left-sided limit and the right-sided limit of the state y at a time t. The symbol $y^\bullet$ denotes the left-sided or the right-sided limit of the state y at a past time point $t - \tau_i(t, y^-(t), c)$, $1 \leqslant i \leqslant n_\tau$, depending on the behavior of the corresponding deviating argument as follows:

$$l_i^\alpha(t) = \lim_{t' \to t^-} sign\big(\alpha_i(t', y(t'), c) - \alpha_i(t, y^-(t), c)\big) \tag{1.20a}$$

$$y^\bullet(\alpha_i(t, y^-(t), c)) = \begin{cases} y^-(\alpha_i(t, y^-(t), c)) & \text{if} \quad l_i^\alpha(t) = -1 \\ y^+(\alpha_i(t, y^-(t), c)) & \text{if} \quad l_i^\alpha(t) = 0 \quad \text{or} \quad l_i^\alpha(t) = +1. \end{cases} \tag{1.20b}$$

The delay functions $\tau_i : \mathbb{R} \times \mathcal{D}^y \times \mathcal{D}^c \to \mathbb{R}_0^+$, $1 \leqslant i \leqslant n_\tau$ are time-, state-, and parameter-dependent and have non-negative values. There further are switching functions $\sigma_i : \mathbb{R} \times \mathcal{D}^y \times \mathcal{D}^c \times (\mathcal{D}^y)^{n_\tau} \to \mathbb{R}$, $1 \leqslant i \leqslant n_\sigma$, and their signs $\zeta_i(t)$ are defined by

$$\zeta_i(t) := sign(\sigma_i(t, y^-(t), c, \{y^\bullet(t - \tau_j(t, y^-(t), c))\}_{j=1}^{n_\tau})) \in \{-1, 0, 1\}. \tag{1.21}$$

The signs of all switching functions are denoted by $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))^T$, with $\zeta(t) \in \mathcal{I}^\zeta$, and with $\mathcal{I}^\zeta$ as defined in equation (1.15).

The sets $\mathcal{D}_1^t(\mathbb{R})$ and $\mathcal{D}_0^t(\mathbb{R})$, which determine the domains where the time evolution of $y(t)$ is prescribed by the differential equation (1.19a) and by the impulse relation (1.19b) are defined by equation (1.17), i.e. by the signs of the switching functions and hence implicitly by the unknown function $y(t)$ itself.

The right-hand-side function f is a function $f : \mathbb{R} \times \mathcal{D}^y \times \mathcal{D}^c \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_1^\zeta$, with $\mathcal{I}_1^\zeta$ as in equation (1.16b), i.e. f is defined whenever all switching function signs $\zeta(t)$ are non-zero. Further, there is an impulse function $\omega : \mathbb{R} \times \mathcal{D}^y \times \mathcal{D}^c \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_0^\zeta \to \mathbb{R}^{n_y}$, with $\mathcal{I}_0^\zeta$ as in equation (1.16a), i.e. ω is defined whenever at least one switching function sign is zero.

This class of differential equations allows, as demanded, a dependency of the right-hand-side function on multiple past states, with delays that are themselves functions of the current state. In addition, also the switching functions and the impulse functions depend on the past states. The time points of switches and impulses are implicitly defined by zeros of the switching functions and thus by the state of the system itself. Moreover, all functions are allowed to be nonlinear functions of their arguments.

Due to the above-mentioned characteristics, impulsive hybrid discrete-continuous delay differential equations are a highly challenging problem class. The name of this new problem class is motivated by established terminology for simpler classes of differential equations:

- *delay differential equations* for differential equations of the simpler form $\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau})$, see e.g. Paul [201], Guglielmi and Hairer [122], Bellen and Zennaro [26], and Enright and Hayashi [96],

- *impulsive delay differential equations*, for delay differential equations with impulses, typically at a priori known time points, see e.g. Anokhin, Berezansky, and Braverman [6], Ballinger and Liu [15], Corwin, Thompson, and White [71], and Xu and Yang [262],
- and *hybrid discrete-continuous dynamic systems* for differential equations with additional switching conditions, possibly also including impulses, see e.g. Galán, Feehery, and Barton [111], Mao and Petzold [184], and Schlegl, Buss, and Schmidt [221].

In many practical situations, differential equations as models for real-world processes are not considered for all $t \in \mathbb{R}$ but only on some finite time interval $\mathcal{T} = [t^{ini}, t^{fin}]$, where t^{ini} is called the *initial time* and t^{fin} is called the final time. Moreover, a solution (in a sense that is yet to be defined for IHDDEs) $y(t)$ is sought for specific initial conditions. For differential equations without time delays, it is thereby sufficient to specify $y(t^{ini})$, i.e. the state at the initial time. Contrariwise, when time delays are incorporated in the problem formulation, it is often necessary to define $y(t)$ also for times $t < t^{ini}$. This leads to the following definition of *initial value problems in IHDDEs*, which allows for initial times and final times that depend on the parameters c.

Definition 1.2 (Initial Value Problem in IHDDEs (IHDDE-IVP), Model Functions of an IHDDE-IVP)

Let $\mathcal{T}(c) := [t^{ini}(c), t^{fin}(c)] \subset \mathbb{R}$ be some time interval, where $t^{ini} : \mathcal{D}^c \to \mathbb{R}$ and $t^{fin} : \mathcal{D}^c \to \mathbb{R}$ are functions of the parameters with $-\infty < t^{ini}(c) < t^{fin}(c) < \infty$. An Initial Value Problem in IHDDEs (IHDDE-IVP) *for the state $y : (-\infty, t^{fin}(c)] \to \mathcal{D}^y$ is defined by associating the IHDDE of Definition 1.1 with initial conditions, i.e.*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{1.22a}$$

$$\begin{aligned} y(t) &= y^+(t) \\ &= y^-(t) + \omega(t, y^-(t), c, \{y^\bullet(t - \tau_i(t, y^-(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \end{aligned} \tag{1.22b}$$

$$y(t^{ini}(c)) = y^{ini}(c) \tag{1.22c}$$

$$y(t) = \phi(t, c) \quad \text{for} \quad t < t^{ini}(c). \tag{1.22d}$$

The domains and co-domains of $f, \{\tau_i\}_{i=1}^{n_\tau}, \{\sigma_i\}_{i=1}^{n_\sigma}$ and ω are the same as in Definition 1.1. Further, also ζ, $y^\bullet$, $\mathcal{D}_0^t(\mathcal{T}(c))$ and $\mathcal{D}_1^t(\mathcal{T}(c))$ are defined as before.

As the initial time t^{ini} and the final time t^{fin}, also the initial state *$y^{ini} : \mathcal{D}^c \to \mathcal{D}^y$ is a function of the parameters. The function $\phi : (-\infty, t^{fin}(c)] \times \mathcal{D}^c \to \mathcal{D}^y$ is called the* initial function. *Together, the functions $f, \{\tau_i\}_{i=1}^{n_\tau}, \{\sigma_i\}_{i=1}^{n_\sigma}, \omega, t^{ini}, y^{ini}$, ϕ, and t^{fin} are called the* model functions of the IHDDE-IVP.

For a shorter notation of the relevant intervals, the following definitions are used throughout the thesis:

$$\mathcal{T}^\phi(c) := (-\infty, t^{ini}(c)) \quad \text{and} \quad \mathcal{T}^f(c) := (-\infty, t^{fin}(c)], \tag{1.23}$$

i.e. $\mathcal{T}^\phi(c)$ denotes the parameter-dependent interval in which the state is given by the initial function ϕ, and $\mathcal{T}^f(c)$ denotes the "full" time interval that comprises both $\mathcal{T}^\phi(c)$ and $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$.

The initial function in the IHDDE-IVP and the consideration of an extension of $\mathcal{T}(c)$ to the left is dispensable if $t - \tau_i(t, y(t), c) \geqslant t^{ini}(c)$ for $1 \leqslant i \leqslant n_\tau$ and for all $t \in \mathcal{T}(c)$, but necessary otherwise. In many cases it is sufficient to consider a finite extension to the left, e.g. in the case of constant delays $\tau_i(t, y(t), c) \equiv \tau_i(c)$, but in order to treat the case where the delay is state-dependent and the lower bound for $t - \tau_i(t, y(t), c)$ is a priori unknown, the interval $\mathcal{T}^f(c)$ is defined with left-sided endpoint $-\infty$.

The initial function is also defined for times $t \geqslant t^{ini}(c)$, even though it is needed in the problem formulation only for times $t < t^{ini}(c)$ according to equation (1.22d). This extension becomes relevant, e.g., for the differentiability theory presented in Chapter 7 of this thesis. As right border of this extension any time to the right of $t^{ini}(c)$ would be sufficient, but for simplicity of notation the final time $t^{fin}(c)$ is used.

It is remarked that the terminology of calling problem (1.22) an "initial value problem" is sloppy, as not only the initial value but also the initial function determine the evolution of the state. However, the term is widely accepted in the literature of delay differential equations, see

Bellman and Cooke [28], Bellen and Zennaro [26], Guglielmi and Hairer [122], and Bocharov and Romanyukha [33]. Therefore, this term is therefore also used in this thesis.

Having formulated a new class of differential equations and the corresponding IVP in Definitions 1.1 and 1.2, the next important step is to agree on a notion of a "solution" and to define the function space of these "solutions". Before taking these steps in Chapter 2, some subclasses of IHDDEs are introduced, the terminology for these subclasses is defined, and special types of switching functions and delay functions are discussed.

1.2. Subclasses

IHDDEs and the associated IHDDE-IVPs constitute a new and very general class of differential equations and IVPs, respectively. They comprise several "simpler" differential equations and IVPs as subclasses. Some of these "simpler" differential equations have already been encountered in the derivation of IHDDEs in Section 1.1. The formal definition of these subclasses and the introduction of a straightforward terminology in this section is useful in subsequent chapters, e.g., for the categorization of the applications that are presented in Chapter 3. In addition, some alternative terminologies that have been used in the literature are mentioned in this section.

For all definitions in this section, it is assumed that $c \in \mathcal{D}^c$ is an arbitrary but fixed vector of parameter values and that the considered interval is denoted by $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$. The sets $\mathcal{D}_1^t(\mathcal{T}(c))$ and $\mathcal{D}_0^t(\mathcal{T}(c))$ and the domains and co-domains of the model functions as well as of the state y are the same as in the Definitions 1.1 and 1.2, with obvious modifications where necessary (e.g. if a function lacks some of its arguments compared to IHDDE(-IVP)s). The continuity assumptions made in Section 1.1 on f, σ_i and τ_i are assumed to be fulfilled. Furthermore, it is pointed out that most of the IVPs defined in the following require a generalized concept of a solution.

1.2.1. Ordinary Differential Equation

An elementary class of differential equations are *ordinary differential equations*, as defined in the following.

Definition 1.3 (Ordinary Differential Equation (ODE))

An Ordinary Differential Equation (ODE) *is an equation, in which the state y as a function of the time t is characterized by the differential equation*

$$\dot{y}(t) = f(t, y(t), c) \quad \textit{for} \quad t \in \mathbb{R}. \tag{1.24}$$

Definition 1.4 (Initial Value Problem in ODEs)

An Initial Value Problem in ODEs (ODE-IVP) *associates the differential equation (1.24) with an interval $\mathcal{T}(c)$ and with an initial value:*

$$\dot{y}(t) = f(t, y(t), c) \quad \textit{for} \quad t \in \mathcal{T}(c). \tag{1.25a}$$

$$y(t^{ini}(c)) = y^{ini}(c). \tag{1.25b}$$

ODEs and ODE-IVPs are comprised in the Definitions 1.1 and 1.2 of IHDDEs and IHDDE-IVPs by setting $n_\sigma = 0$ and $n_\tau = 0$, i.e. such that there are no switching functions and no past states.

1.2.2. Hybrid Discrete-Continuous Ordinary Differential Equations

If switching functions are included into the problem formulations for ODEs, and if the right-hand-side function f is a function of the switching function signs, then the resulting equation is in this thesis called a *hybrid discrete-continuous ordinary differential equation.*

Definition 1.5 (Hybrid Discrete-Continuous Ordinary Differential Equation (HODE))

A Hybrid discrete-continuous Ordinary Differential Equation (HODE) *is an equation, in which the state y as a function of the time t is characterized by a differential equation and a continuity*

condition as follows:

$$\dot{y}(t) = f(t, y(t), c, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathbb{R}) \tag{1.26a}$$
$$y(t) := y^+(t) = y^-(t) \quad \textit{for} \quad t \in \mathcal{D}_0^t(\mathbb{R}). \tag{1.26b}$$

Herein, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$ *and* $\zeta_i(t) := sign(\sigma_i(t, y^-(t), c))$, $1 \leqslant i \leqslant n_\sigma$, *are the signs of the switching functions* σ_i.

Definition 1.6 (Initial Value Problems in HODEs (HODE-IVP))

An Initial Value Problem in HODEs (HODE-IVP) *associates the differential equation (1.26a) and the continuity condition (1.26b) with an interval* $\mathcal{T}(c)$ *and with an initial value:*

$$\dot{y}(t) = f(t, y(t), c, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{1.27a}$$
$$y(t) := y^+(t) = y^-(t) \quad \textit{for} \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \tag{1.27b}$$
$$y(t^{ini}(c)) = y^{ini}(c). \tag{1.27c}$$

HODEs and HODE-IVPs are comprised in the Definitions 1.1 and 1.2 of IHDDEs and IHDDE-IVPs by setting $n_\tau = 0$ and setting the impulse functions identically zero, i.e.: $\omega(t, y^-(t), c, \zeta) \equiv 0$.

Equations of the form (1.26) have frequently been given names including the word "switched" or "switching", e.g. "switched system", "switched-mode dynamical system", "switched differential equations", "differential equations with switching conditions", etc. The attribute "hybrid discrete-continuous" that is used in this thesis is motivated by the wish to emphasize that the right-hand-side function f of the differential equation is a function of both discrete variables (i.e. ζ) and continuous variables (i.e. t and y).

1.2.3. Impulsive Ordinary Differential Equations

If switching functions are included into the problem formulation of an ODE in order to chracterize the location of discontinuities in the state $y(t)$, then the resulting equation is called an *impulsive ordinary differential equation.*

Definition 1.7 (Impulsive Ordinary Differential Equation (IODE))

An Impulsive Ordinary Differential Equation (IODE) *is an equation, in which the state* y *as a function of the time* t *is characterized by both a differential equation and by impulses as follows:*

$$\dot{y}(t) = f(t, y(t), c) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathbb{R}) \tag{1.28a}$$
$$y(t) := y^+(t) = y^-(t) + \omega(t, y^-(t), c, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_0^t(\mathbb{R}). \tag{1.28b}$$

Herein, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$ *and* $\zeta_i(t) = sign(\sigma_i(t, y^-(t), c))$, $1 \leqslant i \leqslant n_\sigma$, *are the signs of the switching functions* σ_i.

Definition 1.8 (Initial Value Problem for IODEs (IODE-IVP))

An Initial Value Problem in IODEs (IODE-IVP) *associates the differential equation (1.28a) and the impulse condition (1.28b) with an interval* $\mathcal{T}(c)$ *and with an initial value:*

$$\dot{y}(t) = f(t, y(t), c) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{1.29a}$$
$$y(t) := y^+(t) = y^-(t) + \omega(t, y^-(t), c, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \tag{1.29b}$$
$$y(t^{ini}(c)) = y^{ini}(c). \tag{1.29c}$$

IODEs and IODE-IVPs are comprised in the Definitions 1.1 and 1.2 of IHDDEs and IHDDE-IVPs by setting $n_\tau = 0$ and eliminating the dependence of the right-hand-side function f on the signs of the switching functions. Hence, the switching functions σ_i are needed in IODEs merely to indicate the time points where impulses have to be applied; their signs $\zeta_i(t)$ are further used for the definition of the impulse.

In the literature, an equation of the form (1.28) is often simply called "impulsive differential equation". In this thesis, it is stressed that equation (1.28) is otherwise "ordinary" in the sense that it does not include time delays.

1.2.4. Impulsive Hybrid Discrete-Continuous Ordinary Differential Equations

If a zero of the switching functions may trigger both a switch in the right-hand-side function and an impulse, then the equation is called an *impulsive hybrid discrete-continuous ordinary differential equation*.

Definition 1.9 (Impulsive Hybrid Discrete-Continuous Ordinary Differential Equation (IHODE))

An Impulsive Hybrid discrete-continuous Ordinary Differential Equation (IHODE) *is an equation, in which the state y as a function of the time t is characterized by both a differential equation and by impulses as follows:*

$$\dot{y}(t) = f(t, y(t), c, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_1^t(\mathbb{R}) \tag{1.30a}$$

$$y(t) := y^+(t) = y^-(t) + \omega(t, y^-(t), c, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_0^t(\mathbb{R}). \tag{1.30b}$$

Herein, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$ and $\zeta_i(t) = sign(\sigma_i(t, y^-(t), c))$, $1 \leqslant i \leqslant n_\sigma$, are the signs of the switching functions σ_i.

Definition 1.10 (Initial Value Problems in IHODEs (IHODE-IVP))

An Initial Value Problem in IHODEs (IHODE-IVP) *associates the differential equation (1.30a) and the impulse condition (1.30b) with an interval $\mathcal{T}(c)$ and with an initial value:*

$$\dot{y}(t) = f(t, y(t), c, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{1.31a}$$

$$y(t) := y^+(t) = y^-(t) + \omega(t, y^-(t), c, \zeta(t)) \quad \text{for} \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \tag{1.31b}$$

$$y(t^{ini}(c)) = y^{ini}(c). \tag{1.31c}$$

IHODEs and IHODE-IVPs are comprised in the Definitions 1.1 and 1.2 of IHDDEs and IHDDE-IVPs by setting $n_\tau = 0$.

Equations of the form (1.30) have also been simply called "hybrid discrete/continuous systems", without stressing the presence of impulses. The viewpoint behind this terminology is that the impulse is regarded as a "discrete event" in the sense that the evolution of the system is not described by the continuous dynamics (i.e., by the ODE).

1.2.5. Delay Differential Equations

The differential equations considered above, i.e. HODEs, IODEs, and IHODEs, are subclasses of IHDDEs with $n_\tau = 0$ and $n_\sigma > 0$. On the other hand, if an IHDDE without switching functions but with time delays is considered, then the result is a so-called *delay differential equation*.

Definition 1.11 (Delay Differential Equation (DDE))

A Delay Differential Equation (DDE) *is an equation, in which the state y as a function of the time t is characterized by the differential equation*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}) \quad \text{for} \quad t \in \mathbb{R} \tag{1.32}$$

Definition 1.12 (Initial Value Problem in DDEs (DDE-IVP))

An Initial Value Problem in DDEs (DDE-IVP) *associates the differential equation (1.32) with an interval $\mathcal{T}(c)$, with an initial value, and with an initial function:*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}) \quad \text{for} \quad t \in \mathcal{T}(c) \tag{1.33a}$$

$$y(t^{ini}(c)) = y^{ini}(c) \tag{1.33b}$$

$$y(t) = \phi(t, c) \quad \text{for} \quad t < t^{ini}(c). \tag{1.33c}$$

DDEs and DDE-IVPs are comprised in the Definitions 1.1 and 1.2 of IHDDEs and IHDDE-IVPs by setting $n_\sigma = 0$.

It should be mentioned that there is a rich variety of alternative terms for equations of the form (1.32). Other popular names are, e.g., "differential-difference equations", "retarded ordinary differential equations", "differential equations with deviating argument", and "differential delay

equations". Sometimes, equations of the form (1.32) are treated as a special type of "functional differential equations".

1.2.6. Hybrid discrete-continuous Delay Differential Equations

Equations that feature both time delays and switches, but no impulses, are in this thesis called *hybrid discrete-continuous delay differential equations*.

Definition 1.13 (Hybrid discrete-continuous Delay Differential Equation (HDDE))

A Hybrid Discrete-Continuous Delay Differential Equation (HDDE) *is an equation, in which the state y as a function of the time t is characterized by a differential equation and a continuity condition as follows:*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_0^t(\mathbb{R}) \tag{1.34a}$$

$$y(t) := y^+(t) = y^-(t) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathbb{R}). \tag{1.34b}$$

Herein, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$, and further $\zeta_i(t) := sign(\sigma_i(t, y(t), c, \{y(t - \tau_j(t, y(t), c))\}_{j=1}^{n_\tau}))$ for $1 \leqslant i \leqslant n_\sigma$ are the signs of the switching functions σ_i

Definition 1.14 (Initial Value Problem in HDDEs (HDDE-IVP))

An Initial Value Problem in HDDEs (HDDE-IVP) *associates the differential equation (1.34a) and the continuity condition (1.34b) with an interval $\mathcal{T}(c)$, with an initial value, and with an initial function:*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{1.35a}$$

$$y(t) = y^+(t) = y^-(t) \quad \textit{for} \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \tag{1.35b}$$

$$y(t^{ini}(c)) = y^{ini}(c) \tag{1.35c}$$

$$y(t) = \phi(t, c) \quad \textit{for} \quad t < t^{ini}(c). \tag{1.35d}$$

Herein, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$ and $\zeta_i(t) := sign(\sigma_i(t, y^-(t), c, \{y^\bullet(t - \tau_j(t, y^-(t), c))\}_{j=1}^{n_\tau}))$ for $1 \leqslant i \leqslant n_\sigma$ are the signs of the switching functions σ_i, which are defined by using the left-sided limit of the current state and the left-sided limit or the right-sided limit of the past state depending on the value of

$$l_i^\alpha = \lim_{t' \to t^-} sign\big(\alpha_i(t', y(t'), c) - (\alpha_i(t, y^-(t), c))\big) \tag{1.36}$$

as follows:

$$y^\bullet(\alpha_i(t, y^-(t), c)) = \begin{cases} y^-(\alpha_i(t, y^-(t), c)) & \textit{if} \quad l_i^\alpha = -1 \\ y^+(\alpha_i(t, y^-(t), c)) & \textit{if} \quad l_i^\alpha = 0 \quad \textit{or} \quad l_i^\alpha = +1. \end{cases} \tag{1.37}$$

Herein, the symbol $y^\bullet$ is used in the definition of the switching function signs because the initial function $\phi(t, c)$ is allowed to be discontinuous.

HDDEs and HDDE-IVPs are comprised in the Definitions 1.1 and 1.2 of IHDDEs and IHDDE-IVPs by setting $\omega(t, y^-(t), c, \{y^\bullet(t - \tau_i(t, y^-(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \equiv 0$.

Alternative names for equation (1.34) found in the literature are "switching systems with delay", "switching time-delay systems", "delay differential equations with switches", and "switched system with time delay".

1.2.7. Impulsive Delay Differential Equations

Definition 1.15 (Impulsive Delay Differential Equation (IDDE))

An Impulsive Delay Differential Equation (IDDE) *is an equation, in which the state y as a function of the time t is characterized by both a differential equation and by impulses as follows:*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathbb{R}) \tag{1.38a}$$

$$y(t) := y^+(t) = y^-(t) + \omega(t, y^-(t), c, \{y^\bullet(t - \tau_i(t, y^-(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_0^t(\mathbb{R}). \tag{1.38b}$$

Herein, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$ *and* $\zeta_i(t) = sign(\sigma_i(t, y^-(t), c), \{y^\bullet(t - \tau_j(t, y^-(t), c))\}_{j=1}^{n_\tau})$, $1 \leqslant i \leqslant n_\sigma$, *are the signs of the switching functions* σ_i, *which are defined by using the left-sided of the current state, and the left-sided or right-sided limit of the past state according to equations (1.36), (1.37).*

Definition 1.16 (Initial Value Problem in IDDEs (IDDE-IVP))

An Initial Value Problem in IDDEs (IDDE-IVP) *associates the differential equation (1.38a) and the impulse condition (1.38b) with an interval* $\mathcal{T}(c)$, *with an initial value, and with an initial function:*

$$\dot{y}(t) = f(t, y(t), c, \{y(t - \tau_i(t, y(t), c))\}_{i=1}^{n_\tau}) \qquad for \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{1.39a}$$

$$\begin{aligned} y(t) &:= y^+(t) \\ &= y^-(t) + \omega(t, y^-(t), c, \{y^\bullet(t - \tau_i(t, y^-(t), c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad for \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \end{aligned} \tag{1.39b}$$

$$y(t^{ini}(c)) = y^{ini}(c) \tag{1.39c}$$

$$y(t) = \phi(t, c) \quad for \quad t < t^{ini}(c). \tag{1.39d}$$

IDDEs and IDDE-IVPs are comprised in the Definitions 1.1 and 1.2 of IHDDEs and IHDDE-IVPs by eliminating the dependence of the right-hand-side function f on the signs of the switching functions. Similar as for IODEs and IODE-IVPs, the switching functions are used here only to characterize the time points where impulses have to be applied and to define the choice of the impulse functions.

1.3. Switching Function Characterization

In IHDDE-IVPs, the zeros of $\zeta_i(t) = \text{sign}(\sigma_i(t, y^-(t), c, \{y^\bullet(t - \tau_j(t, y(t), c))\}_{j=1}^{n_\tau}))$ characterize the set $\mathcal{D}_0^t(\mathcal{T}(c))$ and hence those times where impulses have to be applied or where the right-hand-side function f may change discontinuously. Due to the dependency of the switching functions on current and past states, the location of the zeros of $\zeta_i(t)$ is in general a priori unknown. However, in some cases, namely if a switching function takes the form

$$\sigma_i(t, y^-(t), c, \{y^\bullet(t - \tau_j(t, y(t), c))\}_{j=1}^{n_\tau}) \equiv \tilde{\sigma}_i(t, c). \tag{1.40}$$

the set where it becomes zero can be computed a priori. Even more restrictive is the form

$$\sigma_i(t, y^-(t), c, \{y^\bullet(t - \tau_j(t, y(t), c))\}_{j=1}^{n_\tau}) \equiv t - \tilde{\sigma}_i(c). \tag{1.41}$$

These special cases are termed as follows.

Definition 1.17 (Time-Dependent Switching Function)

If a switching function σ_i *is of the special form (1.40), then it is called a* time-dependent switching function.

Definition 1.18 (Simple Time-Dependent Switching Function)

If a switching function σ_i *is of the special form (1.41), then it is called a* simple time-dependent switching function.

The general case, i.e. a switching function that depends on current and/or past states, is consequently called a *state-dependent switching function.*

Definition 1.19 (State-Dependent Switching Function)

If a switching function σ_i *is not a time-dependent switching function, then it is called a* state-dependent switching function.

In the case that all switching functions of an IHDDE-IVP are time-dependent or simple time-dependent, it is possible to compute all times $t \in \mathcal{D}_0^t(\mathcal{T}(c))$ a priori, which leads to a significant simplification of the problem. In order to be able to express such a property, the following is defined.

Definition 1.20 (IHDDE with (Simple) Time-Dependent Switching Functions, IHDDE with State-Dependent Switching Functions)

If all switching function σ_i, $1 \leqslant i \leqslant n_\sigma$ in an IHDDE are (simple) time-dependent, then it is called an IHDDE with (simple) time-dependent switching functions*. Otherwise it is called an* IHDDE with state-dependent switching functions*. According terminology applies to the associated IHDDE-IVP, and also to the subclasses defined in Section 1.2, i.e. HODE(-IVP), IODE(-IVP), IHODE(-IVP), HDDE(-IVP), and IDDE(-IVP).*

1.4. Delay Characterization

Similar to the switching functions, also the delay functions can be further characterized. The simplest case of delay functions are *constant delays*.

Definition 1.21 (Constant Delays)

A delay function $\tau_i(t, y(t), c)$ is called a constant delay*, if it holds that $\tau_i(t, y(t), c) \equiv \tilde{\tau}_i(c)$.*

Further, it is distinguished between delay functions that depend on the state $y(t)$ and those that do not depend on the state as follows:

Definition 1.22 (Time-Dependent Delays)

A delay function $\tau_i(t, y(t), c)$ is called a time-dependent delay*, if it holds that $\tau_i(t, y(t), c) \equiv \tilde{\tau}_i(t, c)$.*

Definition 1.23 (State-Dependent Delays)

Delay functions $\tau_i(t, y(t), c)$ that are neither constant nor time-dependent are referred to as state-dependent delays*.*

Apart from their dependencies, delay functions are also classified according to the values that they attain. Of particular theoretical and numerical relevance is the question whether delays are *vanishing* for some times on the considered interval.

Definition 1.24 (Vanishing Delay at a Time t)

A delay $\tau_i(t, y(t), c)$ is called a vanishing delay at time t*, if for $c \in \mathcal{D}^c$ and a function $y : \mathcal{T}(c) \to \mathcal{D}^y$, it holds that*

$$\tau_i(t, y(t), c) = 0. \tag{1.42}$$

Whether a delay vanishes at a time t or not is in general not only a property of the delay function. Of course, it depends also on the state y and the parameters c for which the delay function is evaluated.

2. Basic Solution Theory

> *The treatment of such equations requires from the very start a generalization of the concept of solution.*
>
> Filippov, in the preface to his book "Differential Equations with Discontinuous Right Hand Side" [105].

In Chapter 1 a very general class of differential equations termed impulsive hybrid discrete-continuous delay differential equations (IHDDEs) was introduced. The formulation of this class of differential equations allowed two sources of discontinuities: On the one hand, there are time points where at least one switching function becomes zero, which leads to a discontinuous change of one of the arguments of the right-hand-side function, and which triggers an impulsive change of the state. On the other hand, the deviating arguments may cross a time point of discontinuity in the past, which may also lead to a discontinuous change in one argument (or in several arguments) of the right-hand-side function.

It was mentioned several times in Chapter 1 that the presence of discontinuities in the arguments of the right-hand-side function of a differential equation makes it necessary to define what properties a candidate function $y(t)$ has to fulfill in order to be called a "solution" of an initial value problem (IVP). For "discontinuous ordinary differential equations", a class of equations that is closely related to hybrid discrete-continuous ordinary differential equations (HODEs), this fact has led to the statement by Filippov quoted above. However, this issue has also occured in the context of other types of differential equations, see e.g. Bellen and Guglielmi [24] for the class of so-called "delay differential equations of neutral type" and Ballinger and Liu [15] for equations that in the terminology of this thesis would be called "impulsive delay differential equations with simple time-dependent switching function".

A large number of different concepts for "solutions" has been proposed in the context of HODE-IVPs, see e.g. the early article by Hájek [129], the book by Filippov [105], and the survey by Cortés [69]. The solution concepts presented therein could also be used as a basis for developing a variety of solution concepts for IHDDE-IVPs. The content and purpose of this chapter is to select, motivate, and formally define one specific solution concept that is used for the remainder of this thesis.

Organization of This Chapter

Section 2.1 deals with the impulse condition of an IHDDE-IVP. It is demonstrated that this condition leads, in a natural way, to an elementary requirement that a function has to fulfill in order to be called an IHDDE-IVP solution. Furthermore, the consequences of this requirement for problems with switches but without impulses are discussed. Section 2.2 addresses the problem that the arguments of the right-hand-side function exhibit, due to the presence of impulses in the past, discontinuities away from the zero sets of the switching functions. The discussion of this issue leads to the formulation of a second requirement for IHDDE-IVP solutions.

Section 2.3 takes up the findings of the previous sections and condenses them into the formal definition of an IHDDE-IVP solution. Section 2.4 discusses the consequences of the presence of parameters in the model functions. The chapter is concluded by Section 2.5, which establishes the terminology for the various kinds of discontinuities that may occur in IHDDE-IVP solutions.

2.1. The Impulse Condition

The main goal of this chapter is to define what a solution of an IHDDE-IVP is. To this end, regard at first the impulse condition (1.19b), which allows the state to be discontinuous at those time points where at least one of the switching functions is zero. The same condition also requires that both the left-hand-side limit and the right-hand-side limit of the state y exist at this time point.

Correspondingly, there should be some minimal distance between two successive impulses, because otherwise the limits y^+ and y^- may not be defined. Formally, this is accounted for by the *first requirement* for IHDDE-IVP solutions as follows:

(R1) A solution $y(t)$ should be such that the set $\mathcal{D}_0^t(\mathcal{T}(c))$, which contains all time points where at least one switching function is zero, contains only a finite number of times.

Remark that (R1) is equivalent to demanding that two time points $t_1 \in \mathcal{D}_0^t(\mathcal{T}(c))$ and $t_2 \in \mathcal{D}_0^t(\mathcal{T}(c))$ are separated by some minimum distance $\underline{\Delta t}$.

Requirement (R1) is generally necessary in order to ensure that the limits y^+ and y^- are defined at all times $t \in \mathcal{D}_0^t(\mathcal{T}(c))$ for the case that the applied impulses are non-zero. However, (R1) is also imposed if the impulse functions evaluate to zero. The two following simple examples, both of them belonging to the class of HODE-IVPs, illustrate the effect that the requirement (R1) has in this case.

Example 2.1

Consider, on the interval $\mathcal{T} = [0, 2]$, the HODE-IVP

$$\dot{y}(t) = \begin{cases} +1 & \text{for} \quad \zeta(t) = -1 \\ +0.5 & \text{for} \quad \zeta(t) = +1 \end{cases} \tag{2.1a}$$

$$y^+(t) = y^-(t) \quad \text{for} \quad \zeta(t) = 0 \tag{2.1b}$$

$$y(0) = 0, \tag{2.1c}$$

with a switching function $\sigma(t, y(t)) := y(t) - 1$ and $\zeta(t)$ defined by

$$\zeta(t) = \text{sign}(y^-(t) - 1). \tag{2.2}$$

It is noted that the sign $\zeta(t)$ is, in this example, defined with the left-sided limit $y^-(t)$ only for reasons of notational consistency with the general IHDDE-IVP case, see Definition 1.2. Since the IVP is non-impulsive, $\zeta(t)$ could also have been defined by $\text{sign}(y(t) - 1)$.

Consider, as a proposition for a solution $y(t)$ of the HODE-IVP (2.1), the following function:

$$y(t) = \begin{cases} t & \text{for} \quad t \in [0, 1) \\ 0.5 + 0.5t & \text{for} \quad t \in [1, 2]. \end{cases} \tag{2.3}$$

For this function $y(t)$, the switching function is negative at the initial time ($\sigma(t^{ini}, y(t^{ini})) = -1$) so that the sign $\zeta(t)$ is initially -1. For this value of $\zeta(t)$, the right-hand-side function is equal to $+1$, which is fulfilled by the function $y(t)$ for $t \in [0, 1)$. After the zero of the switching function at $t = 1$, the time derivative of the function $y(t)$ as defined by equation (2.3) is equal to 0.5, and thus equal to the value of the right-hand-side function for $\zeta(t) = +1$. In addition, $y(t)$ is continuous at $t = 1$ as imposed by equation (2.1b), so that $y(t)$ can indeed be regarded as a solution of the HODE-IVP (2.1).

There are, however, other functions which fulfill all equations (2.1), e.g.

$$y(t) = \begin{cases} t & \text{for} \quad t \in [0, 1) \\ 1 & \text{for} \quad t \in [1, 2]. \end{cases} \tag{2.4}$$

This function "stalls" at the value $y(t) = 1$ after $t = 1$, so that the sign of the switching function remains $\zeta(t) = 0$. However, for all $t \in [1, 2]$, the continuity condition (2.1b) holds, so that $y(t)$ "solves" the equations (2.1). Observe that it is also possible to create infinitely many other functions $y(t)$ that fulfill all equations in (2.1) by leaving the value 1 at any time $t \in (1, 2)$.

It is clear that all functions $y(t)$ that are identical to 1 for some finite time interval violate requirement (R1). Hence, by imposing (R1), only the function defined by equation (2.3) remains as a solution. In many practical examples, the function that obeys requirement (R1) is also the only (physically) reasonable solution of the IVP.

Sometimes, however, it turns out that staying in the domain where the switching function is zero is the only way to fulfill all equations in an IVP. This case is illustrated by the following example.

Example 2.2

Consider, on the interval $\mathcal{T} = [0, 2]$, the HODE-IVP

$$\dot{y}(t) = \begin{cases} +1 & \text{for} \quad \zeta(t) = -1 \\ -1 & \text{for} \quad \zeta(t) = +1 \end{cases} \tag{2.5a}$$

$$y^+(t) = y^-(t) \quad \text{for} \quad \zeta(t) = 0 \tag{2.5b}$$

$$y(0) = 0, \tag{2.5c}$$

with a switching function $\sigma(t) = y(t) - 1$, the sign of which is defined by

$$\zeta(t) := \text{sign}(y^-(t) - 1). \tag{2.6}$$

Example 2.2 differs from Example 2.1 only by the value of the differential equation for $\zeta(t) = +1$, compare equations (2.1a) and (2.5a). This time, there is no function $y(t)$ that fulfills all equations in (2.5) and additionally obeys the requirement (R1), because no matter what differential equation is chosen for $t > 1$, the sign $\zeta(t)$ of the switching function contradicts the choice of the differential equation.

There exist practical applications where, like in Example 2.2, the vector fields of both right-hand-side functions point toward the regime where the switching function is zero, and the option to remain in this regime is indeed a physically reasonable solution, see e.g. Lenz [171]. This is the main reason why more general solution concepts have been developed in the context of HODE-IVPs. Of particular popularity is the use of so-called *Filippov Solutions*, in reference to Filippov [104, 105], where ODEs with discontinuous right-hand-side functions are studied extensively.

For the benefit of elegantly avoiding accumulation points of impulses it is acceptable for the remainder of this thesis to restrict the attention to those "almost classical solutions" that obey requirement (R1), i.e. solutions where the switching functions become zero only at isolated time points. In particular, this is sufficient for the applications under investigation (see Chapter 3 and Part V of this thesis).

It should be mentioned that for a time-dependent switching function $\sigma_i(t, c)$ the requirement (R1) is easily identified as a condition on $\sigma_i(t, c)$ itself, because there can obviously be no solution that obeys (R1) if $\sigma_i(t, c)$ has countably or uncountably infinite zeros on the finite time interval $\mathcal{T}(c)$.

By imposing the requirement (R1), a solution $y(t)$ is piecewise continuous with only finitely many discontinuities in the interval $\mathcal{T}(c)$, if it is assumed, as obvious, that the solution is continuous in the set $\mathcal{D}_1^t(\mathcal{T}(c))$. Hence, the limits y^+ and y^- are defined on $\mathcal{T}(c)$. In addition, however, the left-sided or right-sided limit also need to be defined at the time points in the past (see the equations (1.19b), (1.21)), so it is reasonable to request that the number of discontinuities in the initial function is also finite. In summary, this motivates that the state $y : \mathcal{T}^f(c) \to \mathcal{D}^y$ should be at least a piecewise continuous function with at most finitely many discontinuities on the full time interval.

2.2. The Differential Equation

According to the definition of IHDDE-IVPs, see Definition 1.2, the function $y(t)$ should fulfill the differential equation (1.19a). However, due to the presence of discontinuities at times $t \in \mathcal{D}_0^t(\mathcal{T}(c))$, and possibly also in the initial function or at the initial time, the past states $y(t - \tau_i(t, y(t), c))$ are generally discontinuous, even for $t \in \mathcal{D}_1^t(\mathcal{T}(c))$. Correspondingly, it can in general not be expected that there exists a function $y(t)$ that has, for all times $t \in \mathcal{D}_1^t(\mathcal{T}(c))$, a classical two-sided derivative that fulfills the differential equation (1.19a). The classical definition of solution is therefore generally insufficient for problems with time delays, see e.g. El'sgol'ts and Norkin [92], page 43ff, and Bellen and Zennaro [26], page 4f.

This issue can, in general, be approached in the same way as in the HODE-IVP case, i.e. by the use of more general solution concepts such as those proposed in Hájek [129], Filippov [105], and Cortés [69]. In particular, Filippov solutions could be used. An alternative is the solution concept employed by Ballinger and Liu [15] for a class of functional differential equations that covers, e.g., IDDEs with time-dependent delays and impulses at a sequence of a priori known time points. Therein, so-called *Carathéodory solutions* are used, i.e. solutions have to satisfy the differential

equation only in an "almost everywhere" sense on the set $\mathcal{D}_1^t(\mathcal{T}(c))$ (everywhere except for a set of Lebesgue measure zero). This allows, in particular, that the deviating arguments oscillate rapidly around a past discontinuity point, as it is the case in the following simple DDE example.

Example 2.3

Consider, on the interval $\mathcal{T} = [0, 2]$, the DDE-IVP

$$\dot{y}(t) = y(t - \tau(t)) \tag{2.7a}$$
$$y(0) = 0 \tag{2.7b}$$
$$y(t) = \begin{cases} 1 & \text{for} \quad t \in (-\infty, -2) \\ 0 & \text{for} \quad t \in [-2, 0) \end{cases} \tag{2.7c}$$

with a delay function that is defined by

$$\tau(t) = \begin{cases} t + 2 - (t-1)^2 \cdot \sin\left(\frac{1}{t-1}\right) & \text{for} \quad t \neq 1 \\ t + 2 & \text{for} \quad t = 1. \end{cases} \tag{2.8}$$

In this example, the initial function has a discontinuity at $t = -2$, and the deviating argument $\alpha(t) = t - \tau(t)$ is such that it oscillates rapidly around -2 when $t \to 1$. More precisely, the function $\alpha(t) + 2$ has infinitely many zeros in any open interval containing $t = 1$, and each zero results in a discontinuity in the right-hand-side function f. Note that this behavior occurs despite the delay function being a continuously differentiable function of time.

In order to avoid situations like those encountered in Example 2.3, the following *second requirement* is imposed for IHDDE-IVP solutions:

(R2) A solution $y(t)$ should be such that for any time s where the state y is discontinuous, the function $\operatorname{sign}\big(\alpha_i(t, y^-(t), c) - s\big)$ is piecewise constant with only finitely many discontinuities for $t \in \mathcal{T}(c)$ and for all $i = 1, \ldots, n_\tau$.

If the delay function is only time-dependent, i.e. $\tau_i(t, c)$, then requirement (R2) can be considered as an explicit condition on $\tau_i(t, c)$. If a time-dependent delay function violates the requirement (R2) – as it is the case in Example 2.3 – there clearly cannot be any solution $y(t)$ such that (R2) is fulfilled. This is analogous to the case of time-dependent switching functions, which have to obey requirement (R1).

Even though it is fairly easy to construct artificial problems like Example 2.3, the requirement (R2) is, with regard to practical problems, only a mild additional condition compared to the solution concept of Ballinger and Liu [15]. At the same time, imposing (R2) significantly simplifies theoretical investigations regarding existence and uniqueness of solutions. Requirement (R2) is also almost mandatory if the solutions should be found numerically, as general-purpose numerical methods for solving DDE-IVPs with an infinite number of discontinuities in the right-hand-side function are very hard or impossible to develop and currently unavailable.

2.3. Formal Definition of IHDDE-IVP Solutions

Observe that both requirements (R1) and (R2) together ensure that a solution has the following properties. First, there are only finitely many times where impulses occur or where the differential equation does not need to be fulfilled because at least one switching function is zero. Second, there are only finitely many times where the differential equation cannot be fulfilled due to discontinuities in the past states (if continuity of the model functions in all real-valued arguments is assumed). This leads to the definition of the following fuction space as a space for solutions of IHDDE-IVPs:

Definition 2.4 (Piecewise (Right-)Continuously Differentiable Function)
Let $-\infty < t_a < t_b < \infty$, and let $\mathcal{D}^y \subset \mathbb{R}^{n_y}$. Then define the function spaces

$$\begin{aligned}
\mathscr{PD}([t_a,t_b],\mathcal{D}^y) := \{y : [t_a,t_b] \to \mathcal{D}^y \mid\ & \\
& y^+(t) = y(t) \text{ and } \dot{y}^+(t) \text{ exist } \forall t \in [t_a,t_b) \quad \wedge \\
& y^-(t) \text{ exists } \forall t \in (t_a,t_b] \quad \wedge \\
& \text{and } y^-(t) = y^+(t),\ \dot{y}^-(t) \text{ exists and } \dot{y}^-(t) = \dot{y}^+(t) \\
& \text{for all but at most a finite number of points } t \in (t_a,t_b]\} \qquad (2.9a)\\
\mathscr{PD}([t_a,t_b),\mathcal{D}^y) := \{y : [t_a,t_b) \to \mathcal{D}^y \mid\ & \\
& y^+(t) = y(t) \text{ and } \dot{y}^+(t) \text{ exist } \forall t \in [t_a,t_b) \quad \wedge \\
& y^-(t) \text{ exists } \forall t \in (t_a,t_b) \quad \wedge \\
& \text{and } y^-(t) = y^+(t),\ \dot{y}^-(t) \text{ exists and } \dot{y}^-(t) = \dot{y}^+(t) \\
& \text{for all but at most a finite number of points } t \in (t_a,t_b)\} \qquad (2.9b)\\
\mathscr{PD}((-\infty,t_b],\mathcal{D}^y) := \{y : (-\infty,t_b] \to \mathcal{D}^y \mid\ & \forall t_c < t_b,\ y|_{[t_c,t_b]} \in \mathscr{PD}([t_c,t_b],\mathcal{D}^y)\} \qquad (2.9c)\\
\mathscr{PD}((-\infty,t_b),\mathcal{D}^y) := \{y : (-\infty,t_b) \to \mathcal{D}^y \mid\ & \forall t_c < t_b,\ y|_{[t_c,t_b)} \in \mathscr{PD}([t_c,t_b),\mathcal{D}^y)\} \qquad (2.9d)
\end{aligned}$$

By convention, define $\dot{y}(t) := \dot{y}^+(t)$. Then, the function spaces defined by equations (2.9) are called the spaces of piecewise (right-)continuously differentiable functions *on the intervals $[t_a,t_b]$, $[t_a,t_b)$, $(-\infty,t_b]$, and $(-\infty,t_b)$.*

With the help of these function spaces, the solution of an IHDDE-IVP is defined as follows:

Definition 2.5 (Solution of an IHDDE-IVP, IHDDE-IVP Solution)
Let $y \in \mathscr{PD}(\mathcal{T}^f(c),\mathcal{D}^y)$ be such that

- *the requirements (R1) and (R2) are fulfilled,*
- *$y(t)$ is continuous and $\dot{y}^+(t)$ fulfills the differential equation (1.22a) for $t \in \mathcal{D}_1^t(\mathcal{T}(c))$,*
- *$y^+(t)$ fulfills the impulse condition (1.22b) for $t \in \mathcal{D}_0^t(\mathcal{T}(c))$,*
- *$y(t)$ fulfills the initial conditions (1.22c), (1.22d) for $t \leqslant t^{ini}(c)$.*

Then the function $y(t)$ is called a solution of the IHDDE-IVP *or, alternatively,* an IHDDE-IVP solution.

This solution concept can be regarded as a special kind of Carathéodory solutions that ensures that the exceptional set where the differential equation is not fulfilled contains only a finite number of time points.

Please note that the use of right-continuity at the initial time (see equations (1.22c) and (1.22d)), at times $t \in \mathcal{D}_0^t(\mathcal{T}(c))$ (equation (1.22b)), as well as right-continuous differentiability for all $t \in \mathcal{D}_1^t(\mathcal{T}(c))$, is conventional. IHDDEs could, alternatively, also be analyzed by using left-sided continuity and differentiability.

Some necessary assumptions on the model functions of the IHDDE-IVP for the existence of the solutions are apparent: For example, since $y \in \mathscr{PD}(\mathcal{T}^f(c),\mathcal{D}^y)$, it is clear that the initial function ϕ has to be in $\mathscr{PD}(\mathcal{T}^\phi(c),\mathcal{D}^y)$. Obvious conditions on time-dependent delay functions and time-dependent switching functions have already been mentioned in Sections 2.1 and 2.2. The development of an existence and uniqueness theory for IHDDE-IVP solutions (in the sense of Definition 2.5), including a complete set of sufficient conditions, is given in Chapter 4.

It is remarked that the requirements (R1) and (R2) were used to motivate why $\mathscr{PD}(\mathcal{T}^f(c),\mathcal{D}^y)$ is an appropriate function space for IHDDE-IVP solutions, but that the fact that y is in this space implies neither (R1) nor (R2) (e.g. the function defined in equation (2.4) is piecewise continuously differentiable but violates (R1)). Therefore, these conditions are stated separately in Defintion 2.5.

In a similar fashion, $y \in \mathscr{PD}(\mathcal{T}^f(c),\mathcal{D}^y)$ only states that there are finitely many discontinuities in y, so it would be possible to insert (physically unreasonable) discontinuities at times $t \in \mathcal{D}_1^t(\mathcal{T}(c))$ that are unmotivated by the problem formulation. In order to avoid this, it has been explicitly demanded that $y(t)$ is continuous for $t \in \mathcal{D}_1^t(\mathcal{T}(c))$, so that discontinuities may only be present at times $t \in \mathcal{D}_0^t(\mathcal{T}(c))$, i.e. in the zeros of the switching functions.

The remainder of this chapter is devoted to discussing properties of IHDDE-IVP solutions.

2.4. Dependence on Parameters

All model functions of the IHDDE-IVP may depend on the parameters c. As a consequence, also solutions $y(t)$ of the IHDDE-IVP (1.22) generally depend on the parameters c, which motivates to use the notation $y(t;c)$. However, for the remainder of this part of the thesis, as well as for Part II, the dependency on parameters is neglected and the notation $y(t)$ is used until the parameter dependence of the solution becomes relevant in Part III.

2.5. Discontinuities

2.5.1. Propagation of Discontinuities and Discontinuity Order

A crucial property of solutions of differential equations with time delays is that a discontinuity in the state y, e.g. at the initial time, may lead to further discontinuities in the right-hand-side function f and hence in the time derivative $\dot{y}$. In turn, a discontinuity in $\dot{y}$ may lead to discontinuities in $\ddot{y}$, and so on. This effect is known as *propagation of discontinuities*. The special instance of discontinuities in $\dot{y}$ arising from those in y was already discussed in the context of requirement (R2) in order to ensure that the number of discontinuities in $\dot{y}$ is finite.

The following definition is useful for discussing the discontinuity propagation in IHDDE-IVP solutions.

Definition 2.6 (Discontinuity Order)

For $t' \in \mathcal{T}^f(c)$ let $k \geqslant 1$ be the largest integer number such that the time derivatives of $y(t)$ up to order $k-1$, $y^{(k')}(t')$, $k' = 0, \ldots, k-1$, exist in t' and such that $y^{(k-1)}$ is Lipschitz continuous in t'. Then k is called the discontinuity order *of y in t'. Further, the time t' is called a* (time) point of discontinuity of order k. *In addition, a time $t' \in \mathcal{T}^f(c)$ where $y(t)$ is discontinuous, is called a* time point of discontinuity of order 0.

It is remarked that this definition of the discontinuity order is in line with the definition in Paul [201], i.e. the discontinuity order k denotes the lowest time derivative that does not exist. The definition is, however, different from the one used in Bellen and Zennaro [26], where the discontinuity order is equal to the highest time derivative that does exist.

It is further useful to establish a consistent notation for the sets that contain the time points of discontinuity of an IHDDE-IVP solution $y(t)$ up to a given order on some open or closed interval $\hat{\mathcal{T}}$:

$$\mathcal{D}^t_{\star,k}(\hat{\mathcal{T}}) := \{t \in \hat{\mathcal{T}} \mid \text{t is a discontinuity of order } k \text{ of } y(t)\}. \tag{2.10}$$

2.5.2. Sources of Discontinuity of Orders 0 and 1

By Definition 2.5, solutions of IHDDE-IVPs are such that $\mathcal{D}^t_{\star,0}(\mathcal{T}^f(c))$ and $\mathcal{D}^t_{\star,1}(\mathcal{T}^f(c))$ contain at most a finite number of distinct time points. Because of the requirement (R1), the same holds true for the set $\mathcal{D}^t_0(\mathcal{T}(c))$, which contains all times $t \in \mathcal{T}(c)$ where at least one switching function is zero. The following lemma clarifies the relation of these sets to each other.

Lemma 2.7 (Sources of Discontinuity of Orders 0 and 1)

From $t \in \mathcal{D}^t_{\star,0}(\mathcal{T}(c))$ it follows that $t \in \mathcal{D}^t_0(\mathcal{T}(c))$. The reverse is not necessarily true.

Let $t \in \mathcal{D}^t_{\star,1}(\mathcal{T}(c))$ and let the right-hand-side function f, the switching functions σ_i, $1 \leqslant i \leqslant n_\sigma$, and the delay functions τ_i, $1 \leqslant i \leqslant n_\tau$ of the IHDDE-IVP be continuous in all their arguments. Then either $t \in \mathcal{D}^t_0(\mathcal{T}(c))$, or $t - \tau_i(t, y(t), c) \in \mathcal{D}^t_{\star,0}(\mathcal{T}^f(c))$ for at least one delay function τ_i. The reverse is not necessarily true.

Proof

IHDDE-IVP solutions are continuous in $\mathcal{D}^t_1(\mathcal{T}(c))$ by definition, hence discontinuities of order 0 on the interval $\mathcal{T}(c)$ may only occur $\mathcal{D}^t_0(\mathcal{T}(c))$ (i.e. in the zeros of the switching functions).

The reverse does in general not hold, because the impulse function that is applied at a time $t \in \mathcal{D}^t_0(\mathcal{T}(c))$ might evaluate to zero.

For discontinuities of order 1, the proof is obtained by contradiction. Consider $t \in \mathcal{D}^t_1(\mathcal{T}(c))$ such that $t - \tau_i(t, y(t), c) \notin \mathcal{D}^t_{\star,0}(\mathcal{T}^f(c))$ for $1 \leqslant i \leqslant n_\tau$ and observe that in this case both the current state argument and the past state arguments of the switching functions σ_i are continuous at the

time t. Since the switching functions are continuous and non-zero in $t \in \mathcal{D}_1^t(\mathcal{T}(c))$, it holds that the signs of the switching functions are constant (and non-zero) in $[t-\epsilon, t+\epsilon] \cap \mathcal{T}(c)$ for some $\epsilon > 0$. In summary, all arguments of f are continuous in time at t, and since f itself is continuous, $\dot{y}$ is continuous as well, which gives a contradiction.

The reverse statement does, in general, not hold because zeros of switching functions might lead only to discontinuities of order higher than 1 in y. ■

2.5.3. Sources of Discontinuities of Arbitrary Order

In general, three main sources for discontinuities of various orders in IHDDE-IVP solutions can be identified. The first kind of discontinuities are those that are present in the initial function, which are referred to as *initial discontinuities*.

Definition 2.8 (Initial Discontinuity)

Let $s \in \mathcal{T}^{\phi}(c)$ be a time point of discontinuity (of some order $k < \infty$) in the initial function ϕ. Then ϕ has an initial discontinuity *(of order k) at the time point s.*

Note that the number of initial discontinuities of order 0 or 1 is finite by choosing $\mathscr{PD}(\mathcal{T}^f(c), \mathcal{D}^y)$ as solution space, whereas there is no restriction on the number of initial discontinuities of higher order.

The second source of discontinuities are the zeros of the switching functions, which are called *root discontinuities*.

Definition 2.9 (Root Discontinuity)

Let $s \in \mathcal{T}(c)$ be such that

$$\sigma_i(s, y^-(s), c, \{y^{\bullet}(s - \tau_k(s, y^-(s), c))\}_{k=1}^{n_\tau}) = 0 \quad \textit{for at least one} \quad i \in \{1, \ldots, n_\sigma\}. \tag{2.11}$$

Then y has a root discontinuity *at the time point s.*

The total number of root discontinuities is always finite for IHDDE-IVP solutions because of requirement (R1), regardless of the discontinuity order.

The term *root discontinuity* is motivated by the fact that s is the root of a switching function. It should be mentioned, however, that in the literature on ODEs or differential-algebraic equations with switching functions, root discontinuities are often called "switching time" or "switching point". In the case of IVPs with non-zero impulses at the switching times, also the term "impulse time" is used.

The third source of discontinuities is the propagation of discontinuities as discussed in Subsection 2.5.1, so it is inituitive to call them *propagated discontinuities*. The definition of their location is, however, more subtle than that of root discontinuities. For example, let $s \in \mathcal{T}^f(c)$ be a discontinuity of order 0, then the location of a propagated discontinuity is not described by the zeros of the function $\alpha_i(t, y^-(t), c) - s$. The reason is that right-continuity of the state y has been imposed. Hence, a propagated discontinuity of order 1 may occur when the function $\alpha_i(t, y^-(t), c) - s$ changes its sign from -1 to 0 or vice versa – but it may remain in s for some finite time or change its sign from 0 to $+1$ or vice versa without causing a discontinuity in the past state.

These observations are formalized in the following definitions:

Definition 2.10 (Propagation Switching Function, Signs of Propagation Switching Functions)

Let $s \in \mathcal{T}^f(c)$ be a time point of discontinuity of some order $k < \infty$. Then the function

$$\sigma_{i,s}^{\alpha}(t, y(t), c) := \alpha_i(t, y(t), c) - s \tag{2.12}$$

is called the propagation switching function*. Further,*

$$\zeta_{i,s}^{\alpha,+}(t) := sign^+(\sigma_{i,s}^{\alpha}(t, y^-(t), c)) \tag{2.13}$$

is called the sign of the propagation switching function*, where* $sign^+(x)$ *for any* $x \in \mathbb{R}$ *is a "simplified sign function" defined as follows:*

$$sign^+(x) := \begin{cases} +1 & for \quad x \geqslant 0 \\ -1 & for \quad x < 0. \end{cases} \tag{2.14}$$

The superscript "+" in $\zeta_{i,s}^{\alpha,+}(t)$ is thereby used to recall that the simplified sign function (2.14) has been used for its definition.

Definition 2.11 (Propagated Discontinuity, Parent Discontinuity, Child Discontinuity)
Let $s \in \mathcal{T}^f(c)$ *be a time point of discontinuity of some order* $k < \infty$*, and let* $\zeta_{i,s}^{\alpha,+}(t)$*,* $1 \leqslant i \leqslant n_\tau$*, be the associated signs of the propagation switching functions. Let further* t *be a point where the sign* $\zeta_{i,s}^{\alpha,+}(t)$ *is discontinuous. Then* y *has a* propagated discontinuity *at the time point* t*.*

If it is necessary to express that the discontinuity at t *originates from the discontinuity in* s*, then the discontinuity at* t *is called the* child (discontinuity) *of the discontinuity in* s*. Furthermore, the discontinuity in* s *is called the* parent (discontinuity) *of the discontinuity in* t*.*

The total number of propagated discontinuities of order 1 is finite for any IHDDE-IVP solution because of requirement (R2). However, (R2) imposes no restriction on the number of propagated discontinuities of higher order.

Having defined both root and propagated discontinuities, it is instructive to regard again the requirements (R1) and (R2). Both requirements aim at limiting the number of low order discontinuities (orders 0 or 1) to some finite number, but nevertheless they are stated in slightly different ways.

More precisely, root discontinuities occur at the zeros of the switching functions, which are assumed to form a set of finitely many points (R1). In contrast, propagated discontinuities occur when the propagation switching functions enters or leave the zero from/to negative values. This is ensured if, as specified by requirement (R2), $\operatorname{sign}\big(\alpha_i(t, y^-(t), c) - s\big)$ undergoes only finitely many discontinuities. In particular, this implies finitely many discontinuities in the simplified sign function $\operatorname{sign}^+\big(\alpha_i(t, y^-(t), c) - s\big)$, and hence finitely many propagated discontinuities of order 1 in y.

For the sake of completeness it is mentioned that there may also be "additional discontinuities" in the IHDDE-IVP solution y that do not fall into any of the above-mentioned categories, e.g. if the right-hand-side function has discontinuities that are not characterized by the zero of a switching function. Such "additional discontinuities" are excluded throughout the thesis, whenever necessary, by making smoothness assumptions for the model functions.

3. Applications

Den Roman doch nicht vergessen!

German TV entertainer and musician Stefan Raab, 36 seconds before the end of the voting time in the first show of "Unser Star für Baku", asking the TV viewers to vote for his favorite candidate Roman Lob.

Novel Models Presented in This Chapter

This chapter presents three novel differential equation models with time delays. Two of the three model feature, in addition, discontinuities or non-differentiabilities of the right-hand-side function, and one model contains an impulse.

The first model that is introduced describes the spread of an epidemic within a population. It is an extension of a model by Cooke and van den Driessche [68] and shows exemplarily how impulses and discontinuous right-hand-side function can be used in mathematical epidemiology. The effects that are considered here are: (a) the arrival of an infected population in a previously healthy population and (b) the development of a drug and a vaccine a certain time after the number of casualities due to the disease has reached a given threshold.

The second model describes cytokine signaling in cells. A variant of an ordinary differential equation (ODE) model by Sommer et al. [240] is developed that makes use of time delays. The resulting delay differential equation (DDE) model has, compared to the original model, two differential states less while keeping the number of parameters constant. Hence, it provides a more concise description of the signaling process.

The third model presented in this chapter is related to the German TV show "Unser Star für Baku", a singing competition that was held in 2012 in order to find the German representative in the Eurovision Song Contest. The special feature of the voting procedure in this TV show was the so-called *livescore*, which displayed the current percentage of votes for each of the candidates. The voting results in one episode of the show are studied as a function of time. Based on the observations, a differential equation model is presented that involves both time delays and switching functions.

Organization of This Chapter

The chapter is subdivided into three sections, where each of the sections introduces one of the models. Section 3.1 presents the epidemiological model, Section 3.2 presents the model for cytokine signaling, and eventually Section 3.3 introduces the model for the voting behavior of the TV viewers of the show "Unser Star für Baku".

3.1. Epidemiology

3.1.1. An SEIRS Model with Two Delays

Differential equation models with time delays are popular as mathematical models for the spread of epidemics. Some references are Hethcote, Lewis, and van den Driessche [145], Genik and van den Driessche [115], Takeuchi, Ma, and Beretta [244], Taylor and Carr [245], Röst, Huang, and Székely [218]. Furthermore, DDEs have also been suggested as a description for the spread of malicious objects in computer networks, see e.g. Mishra and Saini [190].

In this section, the following model by Cooke and van den Driessche [68] is used a starting point for several model extensions and refinements:

$$\dot{y}_1(t) = bY(t) - \lambda\frac{y_1(t)y_3(t)}{Y(t)} + \gamma y_3(t-\tau_1)\exp(-d\tau_1) - dy_1(t) \tag{3.1a}$$

$$\dot{y}_2(t) = \lambda \frac{y_1(t)y_3(t)}{Y(t)} - \lambda \frac{y_1(t-\tau_2)y_3(t-\tau_2)}{Y(t-\tau_2)} \exp(-d\tau_2) - dy_2(t) \tag{3.1b}$$

$$\dot{y}_3(t) = \lambda \frac{y_1(t-\tau_2)y_3(t-\tau_2)}{Y(t-\tau_2)} \exp(-d\tau_2) - (\epsilon + \gamma + d)y_3(t) \tag{3.1c}$$

$$\dot{y}_4(t) = \gamma y_3(t) - \gamma y_3(t-\tau_1) \exp(-d\tau_1) - dy_4(t), \tag{3.1d}$$

where $Y(t) := y_1(t) + y_2(t) + y_3(t) + y_4(t)$. This model falls into the class of DDEs with constant delays.

The first component of the state vector, $y_1(t)$, represents the number of individuals (people, animals, or computer nodes, depending on the concrete situation) in a population that are "susceptible" to the epidemic, i.e. those individuals that are "healthy" but who may become infected. The second state vector component $y_2(t)$ stands for "exposed" individuals, i.e. those who are infected but who are not yet infectious to others. The third state vector component $y_3(t)$ represents the "infected" individuals, i.e. those that are infected and who may infect others. Eventually $y_4(t)$ represents the "removed" or "recovered" class of individuals, i.e. those who had the epidemic but have recovered. The sum over all components of the state vector, denoted by $Y(t)$, stands for the total size of the population.

Because of the names of the classes of individuals, i.e. "susceptible", "exposed", "infected", and "recovered", the differential equation system (3.1) is frequently called an SEIRS model. Thereby, the last "S" indicates that recovered individuals may become susceptible again.

The motivation for the differential equation system is as follows. The first term in equation (3.1a), $bY(t)$, represents new born individuals. The term $\lambda y_1(t)y_3(t)/Y(t)$, which appears in the equations (3.1a) and (3.1b), describes the rate with which susceptible individuals become exposed. After some latency time τ_2, exposed individuals become infected, which is represented by the term $\lambda y_1(t-\tau_2)y_3(t-\tau_2)\exp(-d\tau_2)/Y(t-\tau_2)$. Infected individuals die from the disease at a rate $\epsilon y_3(t)$, and recover from the disease with a rate $\gamma y_3(t)$. Recovered individuals are immune to the epidemic for a certain immunization time τ_1, after which they become susceptible again, see the term $\gamma y_3(t-\tau_1)\exp(-d\tau_1)$. Eventually, there is a disease-independent death rate in each class of individuals, which is represented by the terms $dy_1(t)$, $dy_2(t)$, $dy_3(t)$, and $dy_4(t)$.

3.1.2. Model Modifications and Extensions

The DDE model (3.1) for the spread of epidemics is modified and extended as follows.

As a first modification, it is assumed that infected individuals do not reproduce, therefore the birth term in equation (3.1a) is altered to $b\tilde{Y}(t)$ with $\tilde{Y}(t) = y_1(t) + y_2(t) + y_4(t)$.

As a second extension, it is assumed that a group of infected individuals invades a previously healthy population at time s. For example, s may denote the time when a group of tourists, infected with an exotic disease, returns to their home country. This motivates to introduce a simple time-dependent switching function

$$\sigma_1(t) := t - s, \tag{3.2}$$

and at the zero of this switching function, an impulse is applied:

$$y(s) = y^-(s) + \begin{pmatrix} 0 \\ 0 \\ \nu \\ 0 \end{pmatrix}. \tag{3.3}$$

The symbol ν denotes the size of the newly arrived infected population.

For $t > s$, the epidemic will spread within the population. In order to count the total number of deaths within the infected class, a new differential equation is introduced:

$$\dot{y}_6(t) = (\epsilon + d)y_3(t). \tag{3.4}$$

The symbol $y_5(t)$ is reserved for an additional class that is introduced later.

It is assumed that the spread of the epidemic leads to an increased research activity to fight the disease once the number of casualties has reached a certain threshold φ. The research effort leads, after some time τ_3, to the development of a new drug. The new drug reduces the death rate due

to the epidemic to some value $\tilde{\epsilon} < \epsilon$ and increases the recovery rate to some value $\tilde{\gamma} > \gamma$.

In order to model such a behavior, a state-dependent switching function is defined:

$$\sigma_2(y(t-\tau_3)) = y_6(t-\tau_3) - \varphi. \tag{3.5}$$

Note that $y_6(t)$ counts all deaths in the infected class rather than only those caused by the disease.

Having defined the switching function σ_2, the differential equation (3.1c) for the infected class becomes a function of $\zeta_2(t) := \text{sign}(\sigma_2(y^-(t-\tau_3)))$:

$$\dot{y}_3(t) = \begin{cases} \lambda \frac{y_1(t-\tau_2)y_3(t-\tau_2)}{Y(t-\tau_2)} \exp(-d\tau_2) - (\epsilon+\gamma+d)y_3(t) & \text{for} \quad \zeta_2(t) = -1 \\ \lambda \frac{y_1(t-\tau_2)y_3(t-\tau_2)}{Y(t-\tau_2)} \exp(-d\tau_2) - (\tilde{\epsilon}+\tilde{\gamma}+d)y_3(t) & \text{for} \quad \zeta_2(t) = +1 \end{cases}. \tag{3.6}$$

The changes in the recovery rate and in the death rate due to the disease also need to be taken into account in the differential equations for $y_4(t)$, $y_6(t)$, and $y_1(t)$. In particular, in order to keep the equations consistent, a third switching function is introduced:

$$\sigma_3(y(t-\tau_4)) = y_6(t-\tau_4) - \varphi, \tag{3.7}$$

with $\tau_4 = \tau_1 + \tau_3$. The differential equation for the recovered class then becomes

$$\dot{y}_4(t) = \begin{cases} \gamma y_3(t) - \gamma y_3(t-\tau_1)\exp(-d\tau_1) - dy_4(t) & \text{for} \quad \zeta_2(t) = -1 \\ \tilde{\gamma} y_3(t) - \gamma y_3(t-\tau_1)\exp(-d\tau_1) - dy_4(t) & \text{for} \quad \zeta_2(t) = +1 \quad \text{and} \quad \zeta_3(t) = -1 \\ \tilde{\gamma} y_3(t) - \tilde{\gamma} y_3(t-\tau_1)\exp(-d\tau_1) - dy_4(t) & \text{for} \quad \zeta_2(t) = +1 \quad \text{and} \quad \zeta_3(t) = +1 \end{cases} \tag{3.8}$$

with $\zeta_3(t) := \text{sign}(\sigma_3(y_6^-(t-\tau_4)))$. Correspondingly, the differential equation for the susceptible class reads

$$\dot{y}_1(t) = \begin{cases} b\tilde{Y}(t) - \lambda \frac{y_1(t)y_3(t)}{Y(t)} + \gamma y_3(t-\tau_1)\exp(-d\tau_1) - dy_1(t) & \text{for} \quad \zeta_3(t) = -1 \\ b\tilde{Y}(t) - \lambda \frac{y_1(t)y_3(t)}{Y(t)} + \tilde{\gamma} y_3(t-\tau_1)\exp(-d\tau_1) - dy_1(t) & \text{for} \quad \zeta_3(t) = +1, \end{cases} \tag{3.9}$$

and the differential equation for $y_6(t)$ is modified to

$$\dot{y}_6(t) = \begin{cases} (\epsilon+d)y_3(t) & \text{for} \quad \zeta_2(t) = -1 \\ (\tilde{\epsilon}+d)y_3(t) & \text{for} \quad \zeta_2(t) = +1. \end{cases} \tag{3.10}$$

The increased research on the epidemic may further lead, after some time τ_4, to the development of a vaccine, which makes individuals permanently immune. In mathematical terms, a fourth switching function

$$\sigma_4(y(t-\tau_5)) = y_6(t-\tau_5) - \varphi \tag{3.11}$$

is defined. The differential equation for the susceptible class is then modified once again, such that it depends on the sign $\zeta_4(t) := \text{sign}(\sigma_4^-(y(t-\tau_5)))$:

$$\dot{y}_1(t) = \begin{cases} b\tilde{Y}(t) - \lambda \frac{y_1(t)y_3(t)}{Y(t)} + \gamma y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) & \text{for } \zeta_3(t) = -1 \text{ and } \zeta_4(t) = -1 \\ b\tilde{Y}(t) - \lambda \frac{y_1(t)y_3(t)}{Y(t)} + \tilde{\gamma} y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) & \text{for } \zeta_3(t) = +1 \text{ and } \zeta_4(t) = -1 \\ b\tilde{Y}(t) - \lambda \frac{y_1(t)y_3(t)}{Y(t)} + \gamma y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) - \rho y_1(t) & \text{for } \zeta_3(t) = -1 \text{ and } \zeta_4(t) = +1 \\ b\tilde{Y}(t) - \lambda \frac{y_1(t)y_3(t)}{Y(t)} + \tilde{\gamma} y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) - \rho y_1(t) & \text{for } \zeta_3(t) = +1 \text{ and } \zeta_4(t) = +1. \end{cases} \tag{3.12}$$

Eventually, the size of the new vaccinated class is denoted by $y_5(t)$, and its time evolution is

described by the differential equation

$$\dot{y}_5(t) = \begin{cases} 0 & \text{for} \quad \zeta_4(t) = -1 \\ \rho y_1(t) & \text{for} \quad \zeta_4(t) = +1. \end{cases} \tag{3.13}$$

In addition, $\tilde{Y}(t)$ is redefined as

$$\tilde{Y}(t) = y_1(t) + y_2(t) + y_3(t) + y_4(t) + y_5(t). \tag{3.14}$$

The symbol ρ in the equations (3.12) and (3.13) is the rate with which susceptibles are vaccinated. No vaccination terms are added to the differential equations for $y_2(t)$, $y_3(t)$, and $y_4(t)$. The underlying assumption for this is that exposed and infected individuals would not react to the vaccine and that the recovered individuals are still too weak to receive the vaccine.

Summarizing, the time evolution of the state $y(t) = (y_1(t), y_2(t), \ldots, y_6(t))^T$ is described by an IHDDE of the form (1.19) with constant delays and with both simple time-dependent and state-dependent switching functions. In order to see the relationship to the general expression for the impulse functions (1.19b), define

$$\omega(t, y^-(t), \{y^\bullet(t-\tau_i)\}_{i=1}^5, \zeta(t)) \equiv \omega(\zeta(t)) = \begin{cases} \begin{pmatrix} 0 & 0 & \nu & 0 \end{pmatrix}^T & \text{for} \quad \zeta(t) = (0, \pm 1, \pm 1, \pm 1) \\ \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}^T & \text{else} \end{cases}, \tag{3.15}$$

i.e. an impulse is applied only the root of the first switching function, and the impulse is independent of the time and independent of the state vector.

3.2. Systems Biology

3.2.1. An ODE Model for the Crosstalk of Two Cytokines

Biological cells react to their environment. For example, they react if a cytokine is detected in the vicinity of the cell. This is done, for example, by so-called receptors on the outer membrane of the cell. Once the receptors bind to the cytokine, they may initiate a signaling cascade within the cell that leads to cell growth, cell differentiation, cell death, or other things.

Two particular cytokines are Interleukin-6 (IL-6) and the granulocyte-macrophage colony-stimulating factor (GM-CSF). Sommer et al. [240] have proposed the following ODE model for the interaction ("crosstalk") of the signaling pathways of these two cytokines:

$$\dot{y}_1(t) = +p - \alpha_r \cdot y_1(t) \cdot y_2(t) \tag{3.16a}$$
$$\dot{y}_2(t) = -\alpha_r \cdot \left[u_{GMCSF} + y_1(t)\right] \cdot y_2(t) \tag{3.16b}$$
$$\dot{y}_3(t) = +\alpha_r \cdot \left[u_{GMCSF} + y_1(t)\right] \cdot y_2(t) - b \cdot y_3(t) \cdot u_{GMCSF} - \delta_r \cdot y_3(t) \cdot y_{14}(t) \tag{3.16c}$$
$$\dot{y}_4(t) = +b \cdot y_3(t) \cdot u_{GMCSF} \tag{3.16d}$$
$$\dot{y}_5(t) = +p - \alpha_r \cdot y_5(t) \cdot y_6(t) \tag{3.16e}$$
$$\dot{y}_6(t) = -\alpha_r \cdot \left[u_{IL6} + y_5(t)\right] \cdot y_6(t) \tag{3.16f}$$
$$\dot{y}_7(t) = +\alpha_r \cdot \left[u_{IL6} + y_5(t)\right] \cdot y_6(t) - \delta_r \cdot y_7(t) \cdot y_{14}(t) \tag{3.16g}$$
$$\dot{y}_8(t) = -\alpha_{STAT3+} \cdot y_7(t) \cdot y_{12}(t) \cdot y_8(t) - \alpha_{STAT3} \cdot y_7(t) \cdot y_8(t) + \mu_1 \cdot y_{10}(t) + \mu_2 \cdot y_{10}(t) \tag{3.16h}$$
$$\dot{y}_9(t) = +\alpha_{STAT3+} \cdot y_7(t) \cdot y_{12}(t) \cdot y_8(t) + \alpha_{STAT3} \cdot y_7(t) \cdot y_8(t) - \nu \cdot y_9(t) \tag{3.16i}$$
$$\dot{y}_{10}(t) = +\nu \cdot y_9(t) - \mu_1 \cdot y_{10}(t) - \mu_2 \cdot y_{10}(t) \tag{3.16j}$$
$$\dot{y}_{11}(t) = -\alpha_{SK} \cdot y_3(t) \cdot y_{11}(t) + \alpha_{STAT3+} \cdot y_7(t) \cdot y_{12}(t) \cdot y_8(t) \tag{3.16k}$$
$$\dot{y}_{12}(t) = +\alpha_{SK} \cdot y_3(t) \cdot y_{11}(t) - \alpha_{STAT3+} \cdot y_7(t) \cdot y_{12}(t) \cdot y_8(t) \tag{3.16l}$$
$$\dot{y}_{13}(t) = +\mu_2 \cdot y_{10}(t) - \gamma \cdot y_{13}(t) \tag{3.16m}$$
$$\dot{y}_{14}(t) = -\delta_r \cdot y_7(t) \cdot y_{14}(t) - \delta_r \cdot y_3(t) \cdot y_{14}(t) + 10 \cdot \gamma \cdot y_{13}(t) - \delta_{SOCS3} \cdot y_{14}(t) \tag{3.16n}$$
$$\dot{y}_{15}(t) = +\delta_r \cdot y_7(t) \cdot y_{14}(t) \tag{3.16o}$$
$$\dot{y}_{16}(t) = +\delta_r \cdot y_3(t) \cdot y_{14}(t) \tag{3.16p}$$

In the following, the model is explained.

The state vector component $y_1(t)$ represents the concentration of GM-CSF that is produced by the cell itself at a production rate p. Furthermore, an external stimulus of GM-CSF with concentration u_{GMCSF} can be applied experimentally. The state vector component $y_2(t)$ represents the concentration of the GM-CSF receptor complex. Together, GM-CSF and the GM-CSF receptor complex form the active GM-CSF receptor complex at an activation rate α_r, whose concentration is given by $y_3(t)$. The active GM-CSF receptor complex may be blocked due to overstimulation with GM-CSF. This blocking happens at a rate b, and the concentration of the blocked receptor complex is given by the state vector component $y_4(t)$.

The state vector component $y_5(t)$ represents the concentration of IL-6 that is produced by the cell itself at a production rate p. Furthermore, an external stimulus of IL-6 with concentration u_{IL6} can be applied experimentally. The state vector component $y_6(t)$ represents the concentration of the IL-6 receptor complex. Together, IL-6 and the IL-6 receptor complex form the active IL-6 receptor complex at an activation rate α_r. The concentration of the active receptor complex is given by $y_7(t)$.

The state vector component $y_8(t)$ stands for the concentration of the signal transducer and activator of transcription 3 (STAT-3) in the cytoplasm of the cell. STAT-3 can be phosphorylated ("activated") by binding to the active IL-6 receptor complex with activation rate α_{STAT3}. The concentration of the resulting phosphorylated STAT-3 (pSTAT-3) is given by $y_9(t)$. Then, pSTAT-3 is transported into the nucleus with rate ν (dimerization of pSTAT-3 is neglected in the model), and the concentration of pSTAT-3 in the nucleus is given by $y_{10}(t)$.

The state vector component $y_{11}(t)$ stands for the concentration of a supporting kinase (SK). SK is phosphorylated ("activated") in the presence of the active GM-CSF receptor complex, and the concentration of the resulting activated SK (aSK) is given by $y_{12}(t)$. The phosphorylation of STAT-3 by the active IL-6 receptor complex is enhanced in the presence of aSK, and the additional phosphorylation rate is given by α_{STAT3+}.

Nuclear pSTAT-3 is exported out of the nucleus as unphosphorylated STAT-3 by two different processes, which have rate constants μ_1 and μ_2. Only the second process is associated with an export of SOCS-3 mRNA. The concentration of SOCS-3 mRNA is given by the state vector component $y_{13}(t)$. SOCS-3 mRNA is translated in SOCS-3 at a rate γ, and each mRNA produces 10 SOCS-3. The SOCS-3 concentration is given by $y_{14}(t)$. Eventually, SOCS-3 may deactivate the active IL-6 receptor complex and the active GM-CSF receptor complex at a rate δ_r. The concentrations of the deactivated IL-6 and GM-CSF receptor complexes are given by $y_{15}(t)$ and $y_{16}(t)$, respectively. SOCS-3 may further degrade at a rate δ_{SOCS3} without deactivating an active receptor complex.

3.2.2. A Modified Model with Two Delays

As an alternative to the ODE model (3.16), the crosstalk of the IL-6 signaling pathway and the GM-CSF signaling pathway can also be modeled by using DDEs. Here, a model with two delays is considered, and only the differential equations for those components are given that change compared to equation (3.16):

$$\dot{y}_8(t) = -\alpha_{STAT3+} \cdot y_7(t) \cdot y_{12}(t) \cdot y_8(t) - \alpha_{STAT3} \cdot y_7(t) \cdot y_8(t) + \nu \cdot y_9(t - \tau_1) \tag{3.17a}$$

$$\dot{y}_{14}(t) = -\delta_r \cdot y_7(t) \cdot y_{14}(t) - \delta_r \cdot y_3(t) \cdot y_{14}(t) + \kappa \cdot y_9(t - \tau_2) - \delta_{SOCS3} \cdot y_{14}(t). \tag{3.17b}$$

In this DDE model, the pSTAT-3 molecules remain in the nucleus for some time τ_1. After this time, they are exported as unphosphorylated STAT-3 into the cytoplasm (this is represented by the term $\nu y_9(t - \tau_1)$). Furthermore, SOCS-3 production is given by $\kappa y_9(t - \tau_2)$, i.e. it is directly dependent on the amount of pSTAT-3 at the past time point $t - \tau_2$. This means that τ_2 represents the time that passes between import of pSTAT-3 in the nucleus and export of SOCS-3 mRNA plus the time needed for translation of the mRNA.

The differential equations for $y_{10}(t)$ and for $y_{13}(t)$ can be deleted, because nuclear pSTAT-3 and SOCS-3 mRNA are not needed in the DDE model. Hence, the size of the differential equation system is reduced by 2 compared to the ODE model.

The number of parameters in the system is the same as in the ODE model. The reaction rates for the export of nuclear pSTAT-3 μ_1 and μ_2 as well as the translation rate of SOCS-3 mRNA γ are present only in the ODE model. Contrariwise, the two time delays τ_1 and τ_2 as well as κ are

needed only in the DDE model. Thereby, the quotient κ/ν can be interpreted as the number of SOCS-3 molecules that are produced per pSTAT-3 molecule that is imported into the nucleus.

It should be noted that the ODE model was based on the assumption that each SOCS-3 mRNA is translated into 10 SOCS-3 molecules. The developed DDE model avoids making this assumption.

3.3. "Unser Star für Baku"

3.3.1. Background

The *Eurovision Song Contest* is an annual singing competition that has been held for the first time in 1956. Each country that is an active member of the *European Broadcasting Union* may participate in the competition by sending a contestant to the competition. Typically, around 40 countries take part in the Eurovision Song Contest, and the show takes place in the country of last years' winner. In 2011, Azerbaijain had won the show, and therefore the Eurovision Song Contest 2012 took place in the capital of Azerbaijan, Baku.

In order to find the German representative for the Eurovision Song Contest 2012, the German TV channels *ARD* and *ProSieben* organized a series of eight TV shows called *Unser Star für Baku* (english: "Our Star for Baku"). Twenty young talented singers entered this national selection, and it was the TV viewers who decided, by telephone calls or by sending SMS, which of the candidates would proceed to the next round.

3.3.2. Voting Procedure and Observations

The very special feature of the voting procedure in "Unser Star für Baku" was the so-called *livescore*. The livescore was displayed during the entire show and showed the percentage of votes that each of the singers had received so far.

In the first show of "Unser Star für Baku", five out of ten candidates would be declared as winners and would be allowed to return in the next round, whereas the other five candidates would have to leave the competition. A few minutes before the end of the voting time, it could be observed in the livescore that only six of the ten candidates had a realistic chance to make it to the next round, while the remaining four candidates were lagging far behind.

The voting results for three candidates in the leading group during the last 120 seconds of voting time are displayed in more detail in Figure 3.1 (t = 120 in the figure represents the end of the voting time). It can be observed that the ranking order of the candidates is rapidly changing. For example, the candidate Kai rises from rank 6 to rank 1 within 50 seconds, but then he drops down again to rank 6 within the next 35 seconds and is eventually voted out of the show (see Figure 3.1e).

The candidate Roman, whose voting results are displayed in the Figures 3.1c and 3.1d, was a clear favorite of the show. Of all the ten candidates in the first show, he was the only one to get standing ovations from the studio audience. Furthermore, he received an excellent feedback from music experts in the show. Nevertheless, 30 seconds before the end of the voting time, he drops down to rank 5, being only 0.1% ahead of candidate Leonie. Romans tensed smile – as a reaction to this situation – can be seen in Figure 3.2.

Another interesting observation is made for the candidate Shelly, see the voting results in the Figures 3.1a and 3.1b. At $t = 35s$, i.e. $85s$ before the end of the voting time, she drops down to rank 6, meaning that she would have to leave the show. At that time, Thomas D., a German musician and member of the expert jury of "Unser Star für Baku", intervened by asking the TV viewers to vote for Shelly ("Lasst mir bloß die Shelly drin!"). Nevertheless, for the following 20 seconds, she remains on rank 6 with a constant percentage of votes. Then, however, she receives suffiently many votes that allow her to climb up to rank 1 within $25s$, and a rapid increase in the percentage of votes by about 1% is observed during a time interval of about 40 seconds.

3.3.3. Building a Model

In the following, the goal is to develop a differential equation model that explains the voting behavior of the TV viewers. Therefore, for a show with n candidates, let $z_i(t)$, $1 \leqslant i \leqslant n$, denote the number of votes that candidate i has received until the time t. Further, let $z_{n+1}(t) = \sum_{i=1}^{n} z_i(t)$ be the total number of votes for all candidates that have been received until t. The percentage of

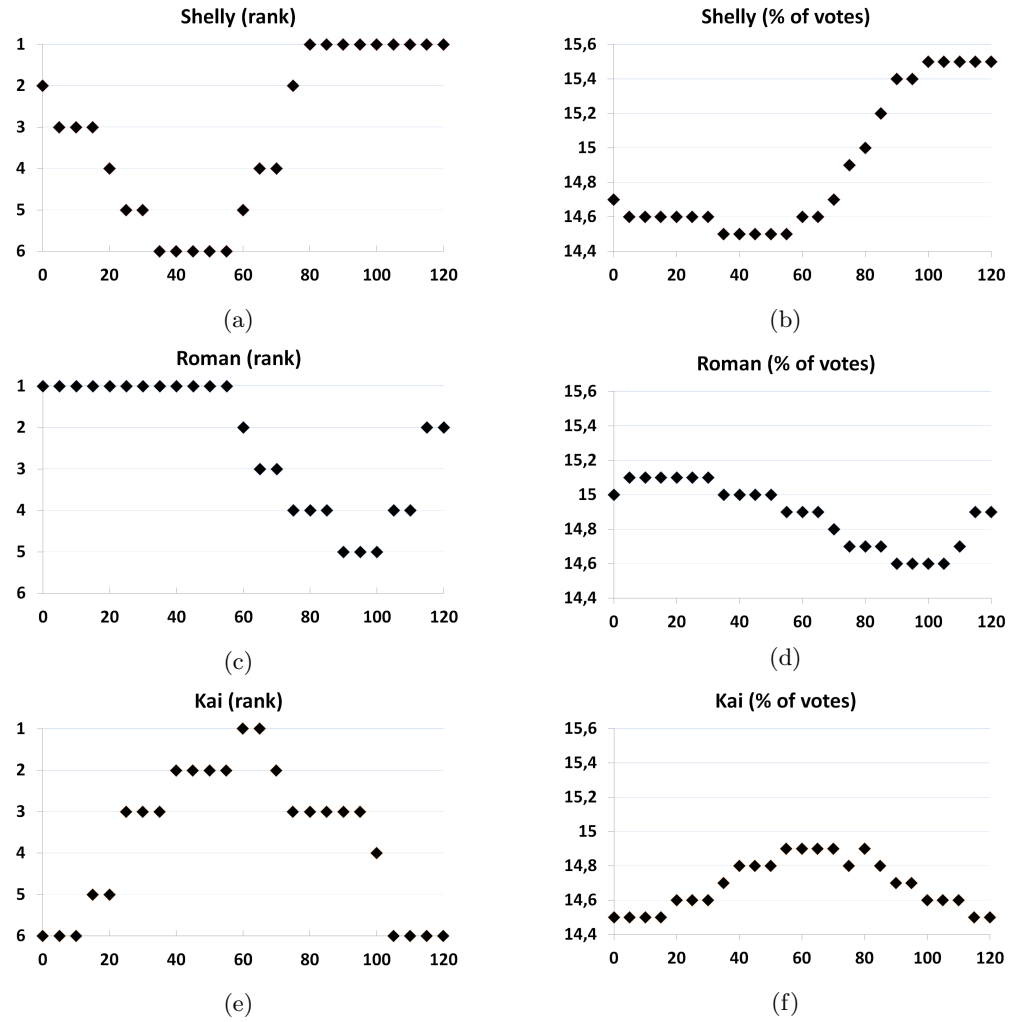


Figure 3.1.: Voting results for three selected candidates of "Unser Star für Baku" during the last 120 seconds of voting time in the first episode of the show. The horizontal axis represents the time in seconds, with $t = 120$ denoting the end of the voting time.

votes that is visible in the livescore is then given by $y_i(t) = z_i(t)/z_{n+1}(t)$ for $1 \leqslant i \leqslant n$. Let further $y_{n+1}(t) := z_{n+1}(t)$, then it follows from elementary differentiation rules that

$$\dot{y}_i(t) = 100 \cdot \frac{\dot{z}_i(t)z_4(t) - \dot{z}_4(t)z_i(t)}{(z_4(t))^2} \quad \text{for} \quad 1 \leqslant i \leqslant n \tag{3.18a}$$

$$\dot{y}_{n+1}(t) = \sum_{i=1}^{n} \dot{z}_i(t). \tag{3.18b}$$

Modeling $\dot{z}_i(t)$ as a function of the percentage values $y_i(t)$ that are displayed in the livescore is the topic of the remainder of this section.

Laziness

As a first step for the construction of the model, the issue is addressed that an apparent favorite such as Roman is very much in danger to be voted out of the show shortly before the end. It is suggested that this can be explained by a certain "laziness" of the TV viewers. Having available the information on the current percentage of votes, the TV viewers are likely to concentrate on those candidates that are on the edge between winning and losing. In contrast to that, TV viewers are unlikely to vote for candidates that are seemingly safe at the top of the table, and they are also unlikely to vote for those that are at the bottom of the table without any reasonable chance of a comeback. Such a laziness is only natural, and also rational, because any vote by phone call

Figure 3.2.: "Unser Star für Baku" candidate Roman Lob, tensely observing the livescore, in which he has dropped to rank 5 half a minute before the end of the voting time. The livescore is visible in the left part of the screenshot, and the remaining voting time is displayed as a countdown in the bottom of the screenshot.
Screenshot reprinted with kind permission of ©Brainpool TV GmbH.

or SMS was associated with costs of 0.50€.

Here, the following differential equation is proposed as a basic model that accounts for laziness:

$$\dot{z}_i(t) = \begin{cases} k_i \cdot \frac{y_i(t)}{\nu(t)} & \text{for} \quad y_i(t) < \nu(t) \\ k_i \cdot \exp(-\lambda(y_i(t) - \nu(t))) & \text{for} \quad y_i(t) > \nu(t). \end{cases} \tag{3.19}$$

Herein, $\nu(t)$ denotes the mean percentage value between the last winner and the first loser, which is formally defined below. Further, k_i is a parameter that represents the quality of the singing performance of candidate i, and λ represents the "laziness" of the TV viewers.

The number of votes that candidate i receives per time unit, i.e. the *voting activity* $\dot{z}_i(t)$, thus attains the maximum value k_i if a candidate is exactly at the threshold. If the candidate falls below the threshold, the value k_i is multiplied by $y_i(t)/\nu(t)$, i.e. with a linear function of the percentage value $y_i(t)$. If the candidate is above the threshold, the voting activity decreases exponentially with the difference to the threshold multiplied by the laziness λ.

It remains to define $\nu(t)$ formally. For this purpose, let $r(i)$ be the function that attributes, to each candidate i, his or her current position in the ranking. Further, let $r^{-1}(j)$ be the inverse function, that yields for a rank j the candidate index i that is presently on this rank. If five candidates are allowed to proceed to the next round, the threshold is given by

$$\nu(t) := \frac{y_{r^{-1}(5)}(t) + y_{r^{-1}(6)}(t)}{2}. \tag{3.20}$$

This means that the threshold $\nu(t)$ is, in this case, defined by the mean percentage value between the candidates on the ranks 5 and 6.

Time Delay

The second issue that is addressed is that a time delay was observed between the intervention of the jury member Thomas D. in favor of the candidate Shelly, and the reception of a significant number of additional votes for Shelly (see the discussion in Subsection 3.3.2).

The occurence of a time delay in the voting procedure is plausible for several reasons. For example, the signal of the TV show has to be broadcast by the TV channel, the viewers need some time to react upon the current intermediate result displayed in the livescore, they need to dial the number of the candidate, the telecommunication company has to establish the telephone connection, and eventually the incoming votes need to be processed and displayed in the livescore. In addition, it needs to be taken into account that the show was aired rather recently in 2012, when many German households had already changed their receiving installations to digital broadcasting.

The encryption and decryption processes in digital broadcasting alone may add up to a time delay of several seconds.

It is thus reasonable to introduce a time delay into the differential equation model (3.19), which leads to the following equation:

$$\dot{z}_i(t) = \begin{cases} k_i \cdot \frac{y_i(t-\tau)}{\nu(t-\tau)} & \text{for} \quad y_i(t-\tau) < \nu(t-\tau) \\ k_i \cdot \exp(-\lambda(y_i(t-\tau) - \nu(t-\tau))) & \text{for} \quad y_i(t-\tau) > \nu(t-\tau). \end{cases} \tag{3.21}$$

This differential equation is to be interpreted as follows: The voting activity that is observed now (at time t) at the TV station depends on the information $y_i(t-\tau)$, $\nu(t-\tau)$ that was send out at the past time $t-\tau$.

Panic

Towards the end of the voting time in "Unser Star für Baku", a countdown was displayed, see the bottom of Figure 3.2. This increased the drama in the TV show, and might have encouraged more TV viewers to vote and/or might have made them vote more frequently. In order to reflect this behavior in the model, a *panic function* is introduced:

$$g^{panic}(t) = \begin{cases} 1 & \text{for} \quad t < t^{fin} - \delta \\ 1 + \frac{t-(t^{fin}-\delta)}{\delta}\rho & \text{for} \quad t > t^{fin} - \delta \end{cases}. \tag{3.22}$$

The panic function is thus constantly equal to 1 until the time $t^{fin} - \delta$ is reached, where t^{fin} represents the end of the voting time. From time $t^{fin} - \delta$ on, the panic function increases linearly, until it reaches the value $1 + \rho$ at t^{fin}. Accordingly, ρ can be interpreted as a "panic factor", and δ represents the duration of the panic.

The panic function enters the differential equation model as a multiplicative factor, i.e.

$$\dot{z}_i(t) = \begin{cases} k_i \cdot g^{panic}(t) \cdot \frac{y_i(t-\tau)}{\nu(t-\tau)} & \text{for} \quad y_i(t-\tau) < \nu(t-\tau) \\ k_i \cdot g^{panic}(t) \cdot \exp(-\lambda(y_i(t-\tau) - \nu(t-\tau))) & \text{for} \quad y_i(t-\tau) > \nu(t-\tau). \end{cases} \tag{3.23}$$

This means that the voting activity increases, from $t^{fin} - \delta$ on, linearly until the end of the voting time t^{fin}.

Summary

The differential equation model for the voting behavior of the TV viewers of "Unser Star für Baku", consists of the equations (3.18), (3.23), (3.22), and (3.20). In order to bring these equations into the standard form of IHDDE-IVPs, the following simple time-dependent switching function is introduced:

$$\sigma_1(t) = t - t^{fin} - \delta. \tag{3.24}$$

This switching function characterizes the beginning of the panic. With the sign $\zeta_1(t) = \text{sign}(\sigma_1(t))$, the linear panic function can be expressed as

$$g^{panic}(t, \zeta_1(t)) = \begin{cases} 1 & \text{if} \quad \zeta_1(t) = -1 \\ 1 + \frac{t-(t^{fin}-\delta)}{\delta}\rho & \text{if} \quad \zeta_1(t) = +1 \end{cases}. \tag{3.25}$$

In addition, state-dependent switching functions are needed in order to define the threshold. For notational simplicity, the case is considered that two out of three candidates are selected. The following three state-dependent switching functions are defined:

$$\sigma_2(y(t-\tau)) = y_1(t-\tau) - y_2(t-\tau) \tag{3.26a}$$
$$\sigma_3(y(t-\tau)) = y_1(t-\tau) - y_3(t-\tau) \tag{3.26b}$$
$$\sigma_4(y(t-\tau)) = y_2(t-\tau) - y_3(t-\tau). \tag{3.26c}$$

If the signs of the switching functions are denoted by $\zeta_i(t) := \text{sign}(\sigma_i(y(t-\tau)))$ for $2 \leqslant i \leqslant 4$, then the threshold at the relevant time point $t-\tau$ in the past can be defined as

$$\nu(t-\tau) = \begin{cases} (y_1(t-\tau) + y_2(t-\tau)/2 & \text{if} \quad (\zeta_2(t), \zeta_3(t), \zeta_4(t)) = (\pm 1, -1, -1) \\ (y_1(t-\tau) + y_3(t-\tau)/2 & \text{if} \quad (\zeta_2(t), \zeta_3(t), \zeta_4(t)) = (-1, \pm 1, +1) \\ (y_2(t-\tau) + y_3(t-\tau)/2 & \text{if} \quad (\zeta_2(t), \zeta_3(t), \zeta_4(t)) = (+1, +1, \pm 1) \end{cases} . \tag{3.27}$$

Moreover, the differential equations for $y_i(t)$ can be formulated as follows:

$$\dot{y}_1(t) = 100 \cdot \frac{k_1 \cdot g^{panic}(t, \zeta(t)) \cdot \beta_1(t, y(t-\tau), \zeta(t)) \cdot y_4(t) - \dot{y}_4(t) y_1(t) y_4(t)/100}{(y_4(t))^2} \tag{3.28a}$$

$$\dot{y}_2(t) = 100 \cdot \frac{k_2 \cdot g^{panic}(t, \zeta(t)) \cdot \beta_2(t, y(t-\tau), \zeta(t)) \cdot y_4(t) - \dot{y}_4(t) y_2(t) y_4(t)/100}{(y_4(t))^2} \tag{3.28b}$$

$$\dot{y}_3(t) = 100 \cdot \frac{k_3 \cdot g^{panic}(t, \zeta(t)) \cdot \beta_3(t, y(t-\tau), \zeta(t)) \cdot y_4(t) - \dot{y}_4(t) y_3(t) y_4(t)/100}{(y_4(t))^2} \tag{3.28c}$$

$$\begin{aligned} \dot{y}_4(t) = & k_1 \cdot g^{panic}(t, \zeta(t)) \cdot \beta_1(t, y(t-\tau), \zeta(t)) + k_2 \cdot g^{panic}(t, \zeta(t)) \cdot \beta_2(t, y(t-\tau), \zeta(t)) \\ & + k_3 \cdot g^{panic}(t, \zeta(t)) \cdot \beta_3(t, y(t-\tau), \zeta(t)). \end{aligned} \tag{3.28d}$$

Therein, $g^{panic}(t, \zeta(t))$ is given by equation (3.25), and the functions $\beta_i(t, y(t-\tau), \zeta(t))$ for $1 \leqslant i \leqslant 3$ are given as

$$\beta_1(t, y(t-\tau), \zeta(t)) = \begin{cases} \frac{y_1(t-\tau)}{\frac{1}{2}[y_1(t-\tau)+y_2(t-\tau)]} & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (-1, -1, -1) \\ \frac{y_1(t-\tau)}{\frac{1}{2}[y_1(t-\tau)+y_3(t-\tau)]} & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (-1, -1, +1) \\ \exp\left(-\lambda(y_1(t-\tau) - \frac{1}{2}\left[y_1(t-\tau) + y_2(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (+1, -1, -1) \\ \exp\left(-\lambda(y_1(t-\tau) - \frac{1}{2}\left[y_1(t-\tau) + y_3(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (-1, +1, +1) \\ \exp\left(-\lambda(y_1(t-\tau) - \frac{1}{2}\left[y_2(t-\tau) + y_3(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (+1, +1, \pm 1) \end{cases} , \tag{3.29}$$

as

$$\beta_2(t, y(t-\tau), \zeta(t)) = \begin{cases} \frac{y_2(t-\tau)}{\frac{1}{2}[y_2(t-\tau)+y_1(t-\tau)]} & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (+1, -1, -1) \\ \frac{y_2(t-\tau)}{\frac{1}{2}[y_2(t-\tau)+y_3(t-\tau)]} & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (+1, +1, -1) \\ \exp\left(-\lambda(y_2(t-\tau) - \frac{1}{2}\left[y_2(t-\tau) + y_1(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (-1, -1, -1) \\ \exp\left(-\lambda(y_2(t-\tau) - \frac{1}{2}\left[y_2(t-\tau) + y_3(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (+1, +1, +1) \\ \exp\left(-\lambda(y_2(t-\tau) - \frac{1}{2}\left[y_1(t-\tau) + y_3(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (-1, \pm 1, +1) \end{cases} , \tag{3.30}$$

and as

$$\beta_3(t, y(t-\tau), \zeta(t)) = \begin{cases} \frac{y_3(t-\tau)}{\frac{1}{2}[y_3(t-\tau)+y_1(t-\tau)]} & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (-1, +1, +1) \\ \frac{y_3(t-\tau)}{\frac{1}{2}[y_3(t-\tau)+y_2(t-\tau)]} & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (+1, +1, +1) \\ \exp\left(-\lambda(y_3(t-\tau) - \frac{1}{2}\left[y_3(t-\tau) + y_1(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (-1, -1, +1) \\ \exp\left(-\lambda(y_3(t-\tau) - \frac{1}{2}\left[y_3(t-\tau) + y_2(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (+1, +1, -1) \\ \exp\left(-\lambda(y_3(t-\tau) - \frac{1}{2}\left[y_1(t-\tau) + y_2(t-\tau)\right])\right) & \text{for} \quad (\zeta_2, \zeta_3, \zeta_4) = (\pm 1, -1, -1) \end{cases} . \tag{3.31}$$

The argument t of the switching function signs has been surpressed in these expressions.

The impulse function is identically zero $\omega \equiv 0$. The developed model falls into the category of HDDEs with one constant delay and with both state-dependent and simple time-dependent switching functions.

For the general case of n candidates, the definition of $n \cdot (n-1)/2$ switching functions is sufficient to conclude, from the values of the sign functions, the current ranks of all candidates and thus also the threshold $\nu(t-\tau)$.

Part II.

Solutions of IHDDE-IVPs

4. Existence and Uniqueness Theory

> *This method allows the opportunity to determine the solution x(t) on several finite intervals and simultaneously to prove the existence of a solution [...].*
>
> El'sgol'ts and Norkin, on the method of steps, Chapter I.2 in their book "Introduction to the Theory and Application of Differential Equations with Deviating Arguments" [92].

Two elementary questions are immediate for any mathematical problem: Does there exist a solution? And, if yes, is the solution unique? The purpose of this chapter is to find answers to these questions for initial value problems in impulsive hybrid discrete-continuous delay differential equations (IHDDE-IVPs) that were defined in Chapter 1, under the definition of IHDDE-IVP solutions that was established in Chapter 2.

Evidently, it is reasonable to approach this issue by having a look at the established existence and uniqueness results for subclasses of IHDDE-IVPs and for related classes of IVPs.

Literature Survey

For initial value problems in hybrid discrete-continuous ordinary differential equations (HODE-IVPs) it is clear that classical solutions generally exist only away from the zero sets of the switching functions, because the right-hand-side function of the HODE may change discontinuously at the zero sets. For the treatment of those points where the right-hand-side function is discontinuous, a variety of more general solution concepts were proposed. Two examples for generalized solution concepts that were mentioned in Chapter 2. On the one hand, Carathéodory solutions, which fulfill the differential equation almost everywhere. And, on the other hand, Filippov solutions, which allow to "slide" on the zero set of a switching function if the vector field to either side of the zero set points toward the zero set. Existence and uniqueness results for HODE-IVPs using these and other solution concepts can be found in the early article by Hájek [129], in the book by Filippov [105], in the survey by Cortés [69], and references therein.

For initial value problems in delay differential equations (DDE-IVPs) the problem of discontinuities (of order 1) in the solution may arise as a consequence of discontinuities (of order 0) in the initial functions or at the initial time. Accordingly, the use of the classical solution concept is generally insufficient for problems with time delays, which motivates to use the generalized solution concepts that are known from the theory of HODE-IVPs. Somewhat surprisingly, however, the available literature on the existence and uniqueness theory for solutions of DDE-IVPs is predominantly restricted to classical solutions; the theorems thus come along with the requirement of a continuous initial function that links continuously to the initial value. In particular, such a continuity assumption occurs in the presentation of theoretical results in the well-known textbooks by Driver [82], page 290ff, Hale and Verduyn Lunel [130], page 38ff, Kuang [167], page 18ff, Bellen and Zennaro [26], page 32f, and Smith [239], page 25ff. For a fairly general class of functional differential equations, Hale and Verduyn Lunel, page 58f, and Kolmanovskii and Myshkis [163], page 100, regard existence and uniqueness of Carathéodory solutions, but their theorems do not cover the case of IVPs with state-dependent delays and discontinuous initial functions.

Consider next the literature on initial value problems in either hybrid discrete-continuous delay differential equations (HDDE-IVPs) or in impulsive delay differential equations (IDDE-IVPs). In this context, the works by Krishna and Anokhin [166] and Ballinger and Liu [15] are mentioned, and also the chapter on generalized solutions in the book by Kolmanovskii and Myshkis [163], page 126ff. However, the definitions of solutions used in these works differ from the one used in this thesis, and hence also the conditions for existence and uniqueness are substantially different from the ones given in this chapter.

It is, instead, a paper by Bellen and Guglielmi [24], whose point of view and set of conditions is possibly the closest relative to the theory presented in this chapter, even though this publication deals with initial value problems in so-called "delay differential equations of neutral type".

Novel Results Presented in This Chapter

This chapter of the thesis contributes to the available existence and uniqueness theory in two ways.

For the first contribution, the idea of the method of steps is picked up, which is a frequently used technique for showing existence and uniqueness of classical solutions, see Bellman and Cooke [28], El'sgol'ts and Norkin [92], Bellen and Zennaro [26], and Smith [239]. Here, this method is used in the non-standard setting to prove existence and uniqueness of the Carathéodory-type solutions defined in Chapter 2 on given time intervals. A theorem is formulated and proven in detail for the case of HDDE-IVPs with constant delays and simple time-dependent switching functions. Furthermore, it is discussed that the ideas of the proof carry over to all IHDDE-IVPs provided that state-dependencies of the switching and delay functions are excluded.

State dependencies in switching and delay functions in IHDDE-IVPs (or in subclasses of IVPs) are the subject of the second contribution of this chapter. More precisely, for the case of impulsive hybrid discrete-continuous ordinary differential equations (IHODE-IVPs), the notion of consistent switching function signs is established. This allows to concisely formulate a necessary condition for uniqueness of a given Carathéodory-type solution in the sense of Chapter 2. In a similar way, a uniqueness result for solutions of DDE-IVPs and IHDDE-IVPs is established. The concise formulation of the necessary condition for the DDE-case requires the introduction of some additional concepts, which are also used in subsequent chapters of the thesis.

Organization of This Chapter

Section 4.1 briefly recalls standard existence and uniqueness results for ODE-IVPs. Existence and uniqueness of HDDE-IVPs with constant delays and simple time-dependent switching functions are the subject of Section 4.2. The extensions to IHDDE-IVPs with time-dependent delays, general time-dependent switching functions, and impulses are considered in Section 4.3.

For problems with state dependencies in the switching functions or in the delay functions, the discussion is restricted to the problem of showing uniqueness for a given IVP solution. A uniqueness theorem for IHODE-IVPs is given in Section 4.4, which is – after introducing some suitable definitions – transferred to DDE-IVPs in Section 4.5 and to IHDDE-IVPs in Section 4.6.

Notation

Throughout the chapter, it turns out to be helpful to distinguish notationally between the statement of an IVP and its solution. Therefore, in this chapter, the symbol $\mathfrak{y}$ is used to formulate IVPs, whereas y is used either as a symbol for a vector in $\mathbb{R}^{n_y}$ or for the solution of an IVP.

4.1. Preliminaries: ODEs

A helpful technique for proving existence and uniqueness of solutions of IHDDE-IVPs, or subclasses thereof, is to reduce the original problem to an equivalent sequence of ODE-IVPs and applying existence and uniqueness results for these ODE-IVPs. Therefore, two well-known results of ODE theory are recalled first.

Theorem 4.1 (Peano's Theorem on Existence of ODE-IVP Solutions [207])

Let $c \in \mathcal{D}^c$ be a vector of arbitrary but fixed parameter values and let $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$ be the corresponding time interval for which the ODE-IVP of Definition 1.4 is considered. Further, let $\Delta y > 0$ and define the set $\mathcal{D}^y := \{y \mid \|y - y^{ini}(c)\|_\infty \leqslant \Delta y\}$. Let the following assumptions be fulfilled:

(C) Continuity*: The right-hand-side function $f(t, y, c)$ is continuous with respect to t and y for $(t, y) \in \mathcal{T}(c) \times \mathcal{D}^y$.*

(B) Boundedness*: The right-hand-side function f is bounded by*

$$\|f(t,y,c)\|_\infty < M_f \tag{4.1a}$$

$$M_f < \frac{\Delta y}{t^{fin}(c) - t^{ini}(c)} \tag{4.1b}$$

for $(t,y) \in \mathcal{T}(c) \times \mathcal{D}^y$.

Then there exists a solution $y(t)$ of the ODE-IVP on the interval $\mathcal{T}(c)$.

Proof
See Hartman [132]. ■

Theorem 4.2 (The Picard-Lindelöf Theorem on Uniqueness of ODE-IVP Solutions [178, 209])

Consider an ODE-IVP as in Definition 1.4, and let c, $\mathcal{T}(c)$, $\mathcal{D}^y$ be as in Theorem 4.1. Assume that (C), (B) hold and that, in addition, the following assumption is fulfilled:

(L) Lipschitz-Continuity*: The right-hand-side function $f(t,y,c)$ is uniformly Lipschitz continuous with respect to y for $(t,y) \in \mathcal{T}(c) \times \mathcal{D}^y$.*

Then there exists a unique solution $y(t)$ of the ODE-IVP on the interval $\mathcal{T}(c)$.

Proof
See Hartman [132]. ■

4.2. HDDEs with Constant Delays and Simple Time-Dependent Switching Functions

As a subclass of IHDDE-IVPs, the simpler case of HDDE-IVPs with constant delays and simple time-dependent switching functions is considered first:

Definition 4.3 (Initial Value Problem in HDDEs with Simple Time-Dependent Switching Functions and Constant Delays)

An Initial Value Problem in HDDEs with simple time-dependent switching functions and constant delays *for the state $\boldsymbol{\mathfrak{y}} : \mathcal{T}^f(c) \to \mathcal{D}^y$ is given by*

$$\dot{\boldsymbol{\mathfrak{y}}}(t) = f(t, \boldsymbol{\mathfrak{y}}(t), c, \{\boldsymbol{\mathfrak{y}}(t-\tau_i(c))\}_{i=1}^{n_\tau}, \zeta(t)) \quad \textit{for} \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{4.2a}$$

$$\boldsymbol{\mathfrak{y}}(t) = \boldsymbol{\mathfrak{y}}^+(t) = \boldsymbol{\mathfrak{y}}^-(t) \qquad \textit{for} \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \tag{4.2b}$$

$$\boldsymbol{\mathfrak{y}}(t^{ini}(c)) = y^{ini}(c) \tag{4.2c}$$

$$\boldsymbol{\mathfrak{y}}(t) = \phi(t,c) \quad \textit{for} \quad t < t^{ini}(c). \tag{4.2d}$$

All definitions of functions, intervals, and sets carry over from the Definitions 1.1 and 1.2, with the exception that $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))$ here denotes the signs of simple time-dependent switching functions

$$\zeta_i(t) = sign(\sigma_i(t,c)) \tag{4.3a}$$

$$\sigma_i(t,c) = t - \tilde{\sigma}_i(c). \tag{4.3b}$$

Herein, $\tilde{\sigma}_i : \mathcal{D}^c \to \mathbb{R}$, and the delay functions $\tau_i : \mathcal{D}^c \to \mathbb{R}^+$ assume, for fixed parameters c, fixed values (i.e. the delays are independent of t and $y(t)$ and thus constant delays in the sense of Definition 1.21).

The parameter-dependent (but constant) delays are positive here, whereas in Definitions 1.1, 1.2 of IHDDE(-IVP)s non-negative delay functions were used. This is no restriction, because the right-hand-side function f depends on the current state anyway.

The following theorem guarantees the existence of a solution (in the sense of Definition 2.5) of HDDE-IVPs with simple time-dependent switching functions and constant delays.

Theorem 4.4 (Global Existence of Solutions of HDDE-IVPs with Simple Time-Dependent Switching Functions and Constant Delays)

Let $c \in \mathcal{D}^c$ be a vector of fixed parameter values and let $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$ be the corresponding time interval for which the HDDE-IVP of Definition 4.3 is considered. Let $\Delta y > 0$ and define the set $\mathcal{D}^y := \{y \mid \|y - y^{ini}(c)\|_\infty \leqslant \Delta y\}$. The following assumptions should be fulfilled:

(C) Continuity*: The right-hand-side function $f(t, y, c, \{v_i\}_{i=1}^{n_\tau}, \zeta)$ is continuous with respect to t, y, and $\{v_i\}_{i=1}^{n_\tau}$ for $(t, y, \{v_i\}_{i=1}^{n_\tau}, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_1^\zeta$. Further, for the initial function it holds that $\phi(\cdot, c) \in \mathscr{PD}((-\infty, t^{ini}(c)), \mathcal{D}^y)$.*

(B) Boundedness*: The right-hand-side function f is bounded by*

$$\|f(t, y, c, \{v_i\}_{i=1}^{n_\tau}, \zeta)\|_\infty < M_f \tag{4.4a}$$

$$M_f < \frac{\Delta y}{t^{fin}(c) - t^{ini}(c)} \tag{4.4b}$$

for $(t, y, \{v_i\}_{i=1}^{n_\tau}, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_1^\zeta$.

Then there exists a solution $y(t)$ of the HDDE-IVP (Definition 4.3) on the interval $\mathcal{T}(c)$.

Before coming to the proof of Theorem 4.4, it is first remarked that it is called a global existence theorem because it guarantees the existence of the solution on the full time interval $\mathcal{T}^f(c)$, and not just locally in the neighborhood of the initial time. Finding such global assertions regarding existence and uniqueness is typical for this part of the thesis; similarly, in Part III, assertions on the differentiability of solutions with respect to parameters on the full time interval $\mathcal{T}^f(c)$ are of interest. The reason for this is that several problems regarding non-existence, non-uniqueness or non-differentiability arise at the time points where (root or propagated) discontinuities occur, and hence in general not locally in the neighborhood of the initial time.

The proof of Theorem 4.4 relies on applying the *method of steps*. The fundamental idea of the method of steps is described in the following remark.

Remark 4.5 (Fundamental Idea of the Method of Steps)

Consider a DDE with a single constant delay τ, i.e. $\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), \mathfrak{y}(t-\tau))$, with initial condition $\mathfrak{y}(t) = \phi(t)$ for $t \leqslant t^{ini}$, and let both f and ϕ be continuous in their arguments. Then the DDE-IVP is equivalent to a sequence of ODE-IVPs, e.g. on $[t^{ini}, t^{ini} + \tau)$:

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), \phi(t-\tau)) \tag{4.5a}$$

$$\mathfrak{y}(t^{ini}) = \phi(t^{ini}) \tag{4.5b}$$

Further, if the solution of the ODE-IVP (4.5) on $[t^{ini}, t^{ini} + \tau)$ is denoted by $y_1(t)$, then the DDE-IVP on $[t^{ini} + \tau, t^{ini} + 2\tau)$ is equivalent to the ODE-IVP

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), y_1(t-\tau)) \tag{4.6a}$$

$$\mathfrak{y}(t^{ini} + \tau) = y_1(t^{ini} + \tau). \tag{4.6b}$$

In the original paper by Bellman [27] the method was introduced as a computational method for solving DDEs. But the fact that DDE-IVP solutions are, under certain conditions, equivalent to the solution of a sequence of ODE-IVPs has also been used for proving existence and uniqueness of DDE-IVP solutions, e.g. in the textbooks by El'sgol'ts and Norkin [92] and Smith [239]. Here, the equivalence between DDE-IVPs and ODE-IVPs also helps for the proof of Theorem 4.4.

Proof (of Theorem 4.4)

The idea of the proof is to decompose the interval $\mathcal{T}(c)$ into a finite number of subintervals such that on each subinterval the HDDE-IVP is equivalent to an ODE-IVP for which the assumption of Peano's existence theorem are fulfilled.

As a first step, observe that there exists a finite number of discontinuities of order 0 in $y(t)$ for $t \leqslant t^{ini}(c)$ because $\phi(\cdot, c) \in \mathscr{PD}((-\infty, t^{ini}(c)), \mathcal{D}^y)$. Let $\mathcal{M}$ be a set that contains the time points of these discontinuities.

Set $\tilde{t} = t^{ini}(c)$ and choose an increment Δt by

$$\Delta t = \min(\Delta t_1, \Delta t_2, \Delta t_3), \tag{4.7}$$

where

$$\Delta t_1 := \min(\{t^{fin}(c) - \tilde{t}\} \cup \{\tau_i(c) \quad | \quad 1 \leqslant i \leqslant n_\tau\}) \tag{4.8a}$$

$$\Delta t_2 := \min(\{t^{fin}(c) - \tilde{t}\} \cup \{t' + \tau_i(c) - \tilde{t} \quad | \quad 1 \leqslant i \leqslant n_\tau,\ t' \in \mathcal{M},\ t' > \tilde{t} - \tau_i(c)\}) \tag{4.8b}$$

$$\Delta t_3 := \min(\{t^{fin}(c) - \tilde{t}\} \cup \{\tilde{\sigma}_i(c) - \tilde{t} \quad | \quad 1 \leqslant i \leqslant n_\sigma,\ \tilde{\sigma}_i(c) > \tilde{t}\ \}) \tag{4.8c}$$

(the element $t^{fin}(c) - \tilde{t}$ is included in the argument list of the minimum function in order to avoid the occurence of minima of empty sets).

Consider the HDDE-IVP on the interval $[\tilde{t}, \tilde{t} + \Delta t) \subset \mathcal{T}(c)$. Because of Δt_1, all deviating arguments to non-zero delays remain to the left of $\tilde{t}$, and hence in a time domain where the state $y(t)$ is known; in particular, if $\tilde{t} = t^{ini}(c)$ then the past states are given by evaluations of the initial function. Further, due to the choice of Δt_3, no zeros of switching functions are present. Hence, the HDDE-IVP is, on this interval, equivalent to an ODE-IVP with right-hand-side function

$$f_{ODE}(t, \mathfrak{y}(t), c) \equiv f(t, \mathfrak{y}(t), c, \{y(t - \tau_i(c))\}_{i=1}^{n_\tau}, \tilde{\zeta}), \tag{4.9}$$

where $\tilde{\zeta}$ are the non-zero signs of the switching functions on the interval $(\tilde{t}, \tilde{t} + \Delta t)$. Particularly for the case $\tilde{t} = t^{ini}(c)$, the ODE right-hand-side function is given by

$$f_{ODE}(t, \mathfrak{y}(t), c) \equiv f(t, \mathfrak{y}(t), c, \{\phi(t - \tau_i(c))\}_{i=1}^{n_\tau}, \tilde{\zeta}). \tag{4.10}$$

Furthermore, because of the choice of Δt_2, no propagation of initial discontinuities of order 0 occurs. This, together with the continuity assumption (C) of Theorem 4.4, ensures that the function f_{ODE} fulfills the continuity assumption (C) of Theorem 4.1. Similarly, assumption (B) of Theorem 4.4 ensures that the function f_{ODE} fulfills assumption (B) of Theorem 4.1. Hence, there exists a solution $y_1(t)$ of the ODE-IVP, and the same function $y_1(t)$ is also a solution of the HDDE-IVP on the considered interval because the two problems are equivalent.

In the *breaking point* $\hat{t} = \tilde{t} + \Delta t$, the state y is chosen to be continuous: $y(\hat{t}) = y^-(\hat{t})$. This is the unique possible choice regardless of whether $\hat{t} \in \mathcal{D}_1^t(\mathcal{T}(c))$, because solutions need to be continuous in $\mathcal{D}_1^t(\mathcal{T}(c))$, or whether $\hat{t} \in \mathcal{D}_0^t(\mathcal{T}(c))$, in which case continuity follows from equation (4.2b) of the HDDE-IVP.

The arguments can now be iterated: Set $\tilde{t} = \hat{t}$ and define a new increment Δt according to equations (4.7) and (4.8). The procedure can then be continued on each successively defined subinterval $[\tilde{t}, \tilde{t} + \Delta t)$, i.e. the existence of the HDDE-IVP solution is obtained from the existence of the solution of an equivalent ODE-IVP. The applicability of the arguments is not compromised if two or all Δt_i, $i \in \{1, 2, 3\}$, are identical, e.g. because the time point of a propagated discontinuity is identical to the time point of a root discontinuity.

The total number of subintervals is finite, because there is a finite number of initial discontinuities and root discontinuities, and the first expression in equation (4.7) is bounded from below by the smallest delay. Hence, after finitely many subintervals, the increment Δt is chosen such that the end point $\hat{t} = \tilde{t} + \Delta t$ is equal to the final time $t^{fin}(c)$.

The function $y(t)$ constructed piecewise on the subintervals and linked continuously at the breaking points is in $\mathscr{PD}(\mathcal{T}^f(c), \mathcal{D}^y)$, is continuous and the right-sided time derivative fulfills the differential equation in $t \in \mathcal{D}_1^t(\mathcal{T}(c))$, fulfills the continuity condition for $t \in \mathcal{D}_0^t(\mathcal{T}(c))$ and fulfills the initial conditions. It is thus a solution of the HDDE-IVP. ■

For uniqueness of solutions of HDDE-IVPs in the form of Definition 4.3, consider the following theorem.

Theorem 4.6 (Global Uniqueness of Solutions of HDDE-IVPs with Explicit Switching Functions and Constant Delays)

Consider a HDDE-IVP as in Definition 4.3, and let c, $\mathcal{T}(c)$, $\mathcal{D}^y$ be as in Theorem 4.4. Assume that (C), (B) hold and that in addition the following assumption is fulfilled:

(L) Lipschitz-Continuity*: The right-hand-side function $f(t, y, c, \{v_i\}_{i=1}^{n_\tau}, \zeta)$ is uniformly Lipschitz continuous with respect to y for $(t, y, \{v_i\}_{i=1}^{n_\tau}, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_1^\zeta$.*

Then there exists a unique solution $y(t)$ to the HDDE-IVP on the interval $\mathcal{T}(c)$.

Proof
The proof is analogously to Theorem 4.4, with the only modification that instead of the Theorem by Peano the Theorem by Picard and Lindelöf is applied to the ODE-IVPs on the subintervals, which gives the uniqueness of the solution on each subinterval. Moreover, note that in each breaking point the solution is uniquely determined by the continuity assumptions in $\mathcal{D}_0^t(\mathcal{T}(c))$ (by equation (4.2b)) and in $\mathcal{D}_1^t(\mathcal{T}(c))$ (by Definition 2.5 of IHDDE-IVP solutions). ■

It is noted that in case of a zero delay, Lipschitz continuity of the right-hand-side function f in assumption (L) is obviously also needed with respect to the corresponding argument v_i.

4.3. More General Existence and Uniqueness Results

The Theorems 4.4 and 4.6 on the existence and uniqueness of solutions of HDDE-IVPs with explicit switching functions and constant delays can be generalized, as discussed in the following.

4.3.1. Non-Vanishing Time-Dependent Delays

One possible extension is to replace the constant delays $\tau_i(c)$ in the HDDE-IVP (Definition 4.3) by time-dependent delays $\tau_i(t, c)$. The existence and uniqueness assertions carry over to this case as long as the time-dependent delays fulfill some assumptions.

At first, it is obvious that the delay function should be such that the time points where initial discontinuities of order 0 occur are crossed only a finite number of times, i.e. requirement (R2) has to be fulfilled. This is the case, e.g., if all deviating arguments $\alpha_i(t, c) = t - \tau_i(t, c)$ are strictly increasing functions. A second condition is that the delays should not vanish on $\mathcal{T}(c)$ so that the analogy to a sequence of ODE-IVPs holds. And finally, the delay functions $\tau_i(t, c)$ have to be continuous in time so that the right-hand-side functions of the equivalent ODE-IVPs fulfill the continuity assumption (C) of Theorems 4.1 and 4.2.

If these assumptions are fulfilled by the delay functions, then it is still possible to find some finite number of breaking points such that between the breaking points the signs of the switching functions are constant and the past state is known and continuous. Hence, an equivalent ODE-IVP can be formulated. If the right-hand-side function f and the initial function ϕ of the HDDE-IVP are such that the assumptions (C) and (B) of Theorem 4.4 (and the assumption (L) of Theorem 4.6) are fulfilled, then the equivalent ODE-IVP is such that assumptions of Theorem 4.1 (and of Theorem 4.2 are fulfilled). From the proof of Theorem 4.4 (and from the proof of Theorem 4.6) it is then clear that a solution of the HDDE-IVP exists (and is unique). Note that it is uncritical in this context if a propagated discontinuity and a root discontinuity occur at the same breaking point.

4.3.2. General Time-Dependent Switching Functions

In a similar fashion, the existence and uniqueness theorems can also be extended to the case of general time-dependent switching functions $\sigma_i(t, c)$ that are continuous in t. In contrast to simple time-dependent switching functions that were used in Definition 4.3, general time-dependent switching functions may have several zeros or zeros of higher multiplicity. However, as long as the total number of their root discontinuities is still finite (i.e. requirement (R1) holds), it is once again possible to split up $\mathcal{T}(c)$ into finitely many subintervals on which the HDDE-IVP is equivalent to an ODE-IVP for which the assumptions of Theorem 4.1 (or Theorem 4.2) are fulfilled. Hence, existence and uniqueness of the HDDE-IVP solution follows from the equivalence of the HDDE-IVP to a sequence of ODE-IVPs, the basic ODE theory of Section 4.1, and the fact that HDDE-IVP solutions have to be continuous in the breaking points.

4.3.3. Impulses

There is also the possibility to extend the existence and uniqueness theorems to problems with impulses, which leads to additional discontinuities of order 0 in y in the interval $\mathcal{T}(c)$. The time points of these discontinuities may be crossed by time-dependent deviating arguments only finitely many times, so that requirement (R2) is fulfilled. Additionally, it has to be ensured that the state $y(t)$ does not leave the set $\mathcal{D}^y$, i.e. for the proper formulation of existence and uniqueness theorems

it is necessary to choose a combination of a sufficiently large set $\mathcal{D}^y$ and a sufficiently small upper bound on the impulse functions.

4.3.4. Intermediate Summary

The combination of the extensions discussed so far in this section answers questions regarding the existence and uniqueness of solutions of IHDDE-IVPs as long as neither the delay functions nor the switching functions depend on the state y. For IHDDE-IVPs, or subclasses thereof, where the delay functions or the switching functions are state-dependent, it becomes very hard to give a priori conditions that ensure global existence or uniqueness of solutions. A main issue is that very restrictive conditions on the delay functions and switching functions need to be imposed such that the requirements (R1) and (R2) hold regardless of the time evolution of the unknown solution $y(t)$.

The theory developed in the following sections is therefore concerned with a different question: Given a solution $y(t)$ of an IHDDE-IVP with state-dependent switching and/or delay functions, what are the conditions so that uniqueness of this solution can be guaranteed?

In the following, uniqueness theorems are formulated for initial value problems a) in systems with state-dependent switching functions but without delays, i.e. IHODEs, b) in systems with state-dependent delay functions but without switching functions, i.e. DDEs, and finally c) in IHDDEs with both state-dependent switching and delay functions.

4.4. IHODEs with State-Dependent Switching Functions

As an introductory example, consider first an IHODE-IVP with a single state-dependent switching function and a (possibly zero) impulse:

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c, \zeta_1(t)) \qquad \text{for} \quad t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{4.11a}$$

$$\mathfrak{y}(t) := \mathfrak{y}^+(t) = \mathfrak{y}^-(t) + \omega(t, \mathfrak{y}^-(t), c, \zeta_1(t)) \quad \text{for} \quad t \in \mathcal{D}_0^t(\mathcal{T}(c)) \tag{4.11b}$$

$$\mathfrak{y}(t^{ini}(c)) = y^{ini}(c). \tag{4.11c}$$

By Definition 2.5, a solution $y(t)$ has to be such that $\mathcal{D}_0^t(\mathcal{T}(c))$ contains only a finite number of points where the sign $\zeta_1(t) = \text{sign}(\sigma_1(t, y^-(t), c))$ of the switching function is zero (requirement (R1)). If the switching function σ_1 is continuous in all its arguments, then the solution will be that of an ODE-IVP with a constant, non-zero sign $\tilde{\zeta}_1 = \zeta_1(t)$ between two successive zeros of σ_1.

Assume that there is a solution $y(t)$ that fulfills this requirement, then the idea for showing uniqueness is as follows: Make sure that the right-hand-side function f fulfills the assumptions of the Picard-Lindelöf Theorem on those intervals where the IHODE-IVP is equivalent to an ODE-IVP. In addition, after a root discontinuity has occured, make sure that there is a unique choice for the sign $\zeta_1(t)$. A helpful tool for formulating a condition that ensures the latter is the following definition.

Definition 4.7 (Consistent Choice of Switching Function Signs)

Consider an IHODE-IVP according to Definition 1.10. Let $\tilde{t} \in \mathcal{T}(c)$, $\tilde{y} \in \mathcal{D}^y$. Then $\zeta' \in \mathcal{I}_1^\zeta$ is called a consistent choice of switching function signs*, if there exists $\Delta t \geqslant \underline{\Delta t} > 0$ and a solution $y_{\zeta'}(t)$ of the ODE-IVP*

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c, \zeta') \tag{4.12a}$$

$$\mathfrak{y}(\tilde{t}) = \tilde{y} \tag{4.12b}$$

such that

$$\zeta_i(t) = sign(\sigma_i(t, y_{\zeta'}^-(t), c)) \tag{4.13a}$$

$$\equiv \zeta_i' \tag{4.13b}$$

for all $i = 1, \ldots, n_\sigma$ and for $t \in (\tilde{t}, \tilde{t} + \Delta t)$. Otherwise, ζ' it is called an inconsistent choice of switching function signs*.*

Consistency of the choice of the switching function signs does not imply that the associated ODE-IVP can be solved uniquely; it merely states that there exists at least one solution of the ODE-IVP (4.12) such that equation (4.13) holds.

Consider, for illustration, again the simple case of an IHODE-IVP with only one switching function, and assume that the right-hand-side function and the switching function are continuous in all real-valued arguments. If, for some time $\tilde{t}$ and state $y(\tilde{t})$, the switching function $\sigma(\tilde{t}, y(\tilde{t}), c)$ is non-zero, e.g. positive, then continuity of the switching function and continuity of a solution $y(t)$ for $t \in \mathcal{D}_1^t(\mathcal{T}(c))$ ensure that the choice $\zeta' = -1$ is always inconsistent. Hence, there can be at most one consistent choice, $\zeta' = +1$, of the switching function.

The situation is different at the time point of a root discontinuity. Here, the condition $\zeta_1' \in \mathcal{I}_1^\zeta = \{-1, 1\}$ in Definition 4.7 means that the solution $y(t)$ has to be such that the switching function σ_1 is non-zero in $(\tilde{t}, \tilde{t} + \Delta t)$. If the impulse itself leads to a non-zero value of σ_1, then the previous arguments still apply and there can be at most one consistent choice of the switching function sign. But if there is no impulse or if the impulse does not affect the value of the switching function, then in general both $\zeta_1' = 1$ or $\zeta_1' = -1$ may be possible choices of the switching function signs to the right of the time point of the root discontinuity.

Assume that for both choices the right-hand-side function f in the corresponding ODE-IVPs (4.12) are such that the two ODE-IVPs can be solved uniquely (at least locally), and denote the solutions of the two ODE-IVPs by $y_1(t)$ and $y_{-1}(t)$, respectively. It may then happen that the solution $y_1(t)$ – which corresponds to the choice $\zeta_1' = 1$ – is such that the switching function $\sigma_1(t, y_1(t), c)$ actually becomes positive after the time point of the root discontinuity. If yes, the choice $\zeta_1' = 1$ is called a consistent choice, otherwise, it is called an inconsistent choice.

In general, it may happen that there is no, one, or that there are several consistent choices of the switching function signs after the time point of a root discontinuity. A situation where there is no consistent choice was already encountered in Example 2.3. Clearly, for the existence of a solution, there always has to be at least one consistent choice of the switching function signs, and assuming that a solution exists implies that there is at least one such consistent choice. Accordingly, in order to guarantee that the solution is unique, it needs to be ensured that there is only one consistent choice of the switching function signs after the time points where root discontinuities occur. This is one of the assumptions in the following uniqueness theorem.

Theorem 4.8 (Global Uniqueness of IHODE-IVP Solutions)

Consider an IHODE-IVP as in Definition 1.10 with a vector c of fixed parameter values, and let $\mathcal{D}^y \subset \mathbb{R}^{n_y}$ be some open domain. Let $y : \mathcal{T}(c) \to \mathcal{D}^y$ be a solution of the problem. Assume that the following conditions are fulfilled:

(C) Continuity*: The right-hand-side function $f(t, y, c, \zeta)$ is continuous with respect to t and y for $(t, y, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times \mathcal{I}_1^\zeta$. The switching functions $\sigma_i(t, y, c)$, $1 \leqslant i \leqslant n_\sigma$, are continuous with respect to t and y for $(t, y) \in \mathcal{T}(c) \times \mathcal{D}^y$.*

(L) Lipschitz-Continuity*: The right-hand-side function $f(t, y, c, \zeta)$ is uniformly Lipschitz continuous with respect to y for $(t, y, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times \mathcal{I}_1^\zeta$.*

(B) Boundedness*: The right-hand-side function is bounded by*

$$\|f(t, y, c, \zeta)\|_\infty < M_f < \infty \tag{4.14}$$

for $(t, y, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times \mathcal{I}_1^\zeta$.

(CS) Consistent Signs*: For $(t, y) = (t^{ini}(c), y^{ini}(c))$ and for each $(t, y(t))$ with $t \in \mathcal{D}_0^t(\mathcal{T}(c))$ there exists exactly one consistent choice of the switching function signs.*

Then the IHODE-IVP solution $y(t)$ is unique on the interval $\mathcal{T}(c)$.

Proof

Assume that there is some solution $\bar{y}(t)$ different from $y(t)$. However, their values at $t^{ini}(c)$ are identical, and there is a unique consistent choice of the switching function signs for $t \in (t^{ini}(c), t^{ini}(c) + \Delta t)$ for some $\Delta t > 0$. It is then possible to consider some closed interval $[t^{ini}(c), t^{ini}(c) + \delta t]$ and some closed neighborhood $\overline{\mathcal{D}^y} \subset \mathcal{D}^y$ where the IHODE-IVP is equivalent to an ODE-IVP, and where the model functions of the ODE-IVP fulfill the assumptions of the Picard-Lindelöf theorem (Theorem 4.2). Hence, the solution is locally unique.

By keeping the steps small enough and knowing that the solution $y(t)$ lies in the open domain $\mathcal{D}^y$, the same argument can be applied for a sequence of intervals and closed neighborhoods, such that it follows $\bar{y}(t) = y(t)$ on the interval $[t^{ini}(c), s_1)$. Hereby, s_1 denotes the earliest time point where any of the switching function signs becomes zero.

In s_1, identical impulses are applied so that $\bar{y}^+(s_1) = y^+(s_1)$. For $t \in (s_1, s_1 + \Delta t)$, $\Delta t > 0$, there exists again a unique consistent choice of the switching function signs, and a suitable interval and closed neighborhood of $y^+(s_1)$ can be constructed such that local uniqueness of the ODE-IVP solution follows. The above arguments can then be used to show that $\bar{y}(t) = y(t)$ on all subintervals and at all time points where root discontinuities occur. ■

It is remarked that the conditions (C), (L) and (B) of Theorem 4.8 are all formulated for $y \in \mathcal{D}^y$, but from the proof it is clear that it is sufficient to have continuity, Lipschitz continuity and boundedness of the right-hand-side function in a tubular neighborhood around the considered solution $y(t)$.

In comparison to the uniqueness result for ODE-IVPs, Theorem 4.2, the sole difference is the condition (CS). In order to check whether this additional condition is fulfilled, one option is to simply go through all possible choices of switching function signs and check whether equation (4.13) holds. This idea can also be considered as a blueprint for the development of numerical methods for checking uniqueness; however, if multiple switching functions are zero at the same time point the number of possible choices of switching function signs grows combinatorically and it may become computationally expensive to solve all associated ODE-IVPs.

4.5. DDEs with State-Dependent Delay Functions

The next subclass of IHDDE-IVPs that is investigated with respect to the uniqueness of solutions are DDEs with state-dependent delays. A simple case with only one state-dependent delay function is considered first:

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c, \mathfrak{y}(t - \tau_1(t, \mathfrak{y}(t), c))) \tag{4.15a}$$

$$\mathfrak{y}(t^{ini}(c)) = y^{ini}(c) \tag{4.15b}$$

$$\mathfrak{y}(t) = \phi(t, c) \quad \text{for} \quad t < t^{ini}(c). \tag{4.15c}$$

As usual, the right-hand-side function f and the delay function τ_1 shall be continuous. For simplicity, let the initial function $\phi(t, c)$ and the initial value $y^{ini}(c)$ be such that there is only one discontinuity in the states for $t \leqslant t^{ini}(c)$. The time point of this sole discontinuity is denoted by s_1.

Consider a solution $y(t)$ of the problem, and recall the definition of the propagation switching function σ^{α}_{1,s_1} for the sole deviating argument α_1:

$$\sigma^{\alpha}_{1,s_1}(t, y(t), c) = \alpha_1(t, y(t), c) - s_1. \tag{4.16}$$

By requirement (R2), there are only finitely many discontinuities in the following function[1]:

$$\zeta^{\alpha}_{1,s_1}(t) = \operatorname{sign}(\sigma^{\alpha}_{1,s_1}(t, y^-(t), c)). \tag{4.17}$$

More specifically, this implies that there are only finitely many discontinuities in the simplified sign function (recall equation (2.14))

$$\zeta^{\alpha,+}_{1,s_1}(t) = \operatorname{sign}^+(\sigma^{\alpha}_{1,s_1}(t, y^-(t), c)). \tag{4.18}$$

For every time t for which the past time point $\alpha_1(t, y(t), c)$ is to the left or to the right of the discontinuity point s_1, there is a small neighborhood where the deviating argument remains to the left or to the right of s_1, because α_1 is continuous in all its arguments and a solution $y(t)$ of a DDE-IVP is continuous on $\mathcal{D}^t_1(\mathcal{T}(c)) = \mathcal{T}(c)$ as well. Hence, the past states are locally continuous when the sign function is non-zero.

[1] For consistency with the general IHDDE-IVP case, the sign function is defined with the left-sided limit $y^-(t)$, although for DDE-IVPs the state is continuous for $t \in (t^{ini}(c), t^{fin}(c)]$, so that this specification is unnecessary.

Under the assumption that the delay does not vanish for the considered solution $y(t)$, the past state has a known and given value. Then the DDE-IVP is locally equivalent to an ODE-IVP with a continuous right-hand-side function

$$f_{ODE}(t, \mathfrak{y}(t), c) = f(t, \mathfrak{y}(t), c, y(t - \tau_1(t, \mathfrak{y}(t), c))) \tag{4.19}$$

and with some suitable additional assumptions on Lipschitz continuity of f, ϕ, and τ_1, and boundedness of f, uniqueness of the ODE-IVP solution can be guaranteed.

It remains to deal with non-uniqueness issues at those time points t where the deviating argument $t - \tau_1(t, y(t), c)$ is equal to the discontinuity point s_1. One approach in this context is to find some definition of a "consistent choice of propagation switching function signs", i.e. some analogous version of Definition 4.7 for propagation switching functions. Although it is principally easy to transfer the idea, the concrete formulation is significantly more technical.

One of the reasons for this is that the propagation switching functions are not independent of each other, e.g. for two discontinuity points $s_1 < s_2$ in the initial function and a deviating argument α_i it is clear that for $\zeta_{i,s_1}^{\alpha,+}(t) = -1$ it follows that $\zeta_{i,s_2}^{\alpha,+} = -1$. Consequently, it holds that if the total number of discontinuities in the past is known, then it is sufficient to know for each deviating argument α_i the sum of all associated propagation switching function signs rather than all of them individually, in order to determine the number of discontinuities to the left and to the right of $\alpha_i(t, y(t), c)$. This observation leads to the definition of *discontinuity interval indicators* as the DDE-equivalent of the switching function signs:

Definition 4.9 (Discontinuity Interval, Discontinuity Interval Indicators)

Consider a DDE-IVP according to Definition 1.12. Let $t \in \mathcal{T}(c)$, and let $y \in \mathscr{PD}((-\infty, t], \mathcal{D}^y)$ be some piecewise continuously differentiable function. Denote the total number of discontinuities of order 0 in y by n_s, and denote the discontinuity points by s_i, $1 \leqslant i \leqslant n_s$, which shall be sorted ascendingly $s_1 < s_2 < \cdots < s_{n_s}$. Then the intervals $(-\infty, s_1)$, $[s_1, s_2)$, $\ldots$, $[s_{n_s}, t^{fin}(c)]$ are called discontinuity intervals.

Further, let

$$\zeta_{i,s_j}^{\alpha,+}(t) = sign^+(\alpha_i(t, y^-(t), c) - s_j) \tag{4.20}$$

be the signs of the propagation switching functions for $1 \leqslant i \leqslant n_\tau$ and $1 \leqslant j \leqslant n_s$. Then

$$\xi_i^\alpha(t) = n_s + 1 + \frac{1}{2} \sum_{j=1}^{n_s} (\zeta_{i,s_j}^{\alpha,+}(t) - 1) \tag{4.21}$$

are called the discontinuity interval indicators. *As abbreviation, $\xi^\alpha := (\xi_1^\alpha, \ldots, \xi_{n_\tau}^\alpha)^T$ is used. It holds that $\xi^\alpha \in \mathcal{I}^{\xi^\alpha} := \{1, \ldots, n_s + 1\}^{n_\tau}$.*

The discontinuity interval indicators are defined in such a way that $\xi_i^\alpha(t) = 1$ if $\alpha_i(t, y^-(t), c)$ is to the left of the left-most discontinuity point s_1, and that $\xi_i^\alpha(t) = n_s + 1$ if $\alpha_i(t, y^-(t), c)$ is to the right of the right-most discontinuity point s_{n_s}. For all other values $2 \leqslant k \leqslant n_s$, $\xi_i^\alpha(t) = k$ indicates that the deviating argument $\alpha_i(t, y^-(t), c)$ is located in the discontinuity interval $[s_{k-1}, s_k]$.

Another technical complication in transferring the definition of consistent choice of switching function signs to propagated switching functions is as follows. In Definition 4.7, any possible value of switching function signs is assumed, and then an ODE-IVP is formulated that is completely independent on the signs that the switching functions actually have. In other words, a hypothetical right-hand-side function is used that is decoupled from the values of the switching function signs. In order to do something equivalent for DDE-IVPs, it is necessary to find an equivalent decoupling of the value used in the past state argument of the differential equation from the actual value of the function y at the time point in the past. This is made possible by the following definition.

Definition 4.10 (Deduced Functions)

Let $\tilde{t} \in \mathbb{R}$, $y \in \mathscr{PD}((-\infty, \tilde{t}], \mathcal{D}^y)$ be some continuously differentiable function, and let n_s be the total number of discontinuities of order 0 in y. The discontinuity points are denoted by s_i with $1 \leqslant i \leqslant n_s$, and shall be sorted ascendingly $s_1 < s_2 < \cdots < s_{n_s}$.

Then the functions $z_1, z_2, \ldots, z_{n_s+1}$ *defined by*

$$z_1(t) = \begin{cases} y(t) & \text{for} \quad t < s_1 \\ y^-(s_1) + \dot{y}^-(s_1)(t - s_1) & \text{for} \quad t \geqslant s_1 \end{cases} \tag{4.22a}$$

$$z_i(t) = \begin{cases} y^+(s_{i-1}) + \dot{y}^+(s_{i-1})(t - s_{i-1}) & \text{for} \quad t < s_{i-1} \\ y(t) & \text{for} \quad s_{i-1} \leqslant t < s_i \quad \text{for} \quad 2 \leqslant i \leqslant n_s \\ y^-(s_i) + \dot{y}^-(s_i)(t - s_i) & \text{for} \quad t \geqslant s_i \end{cases} \tag{4.22b}$$

$$z_{n_s+1}(t) = \begin{cases} y^+(s_{n_s}) + \dot{y}^+(s_{n_s})(t - s_{n_s}) & \text{for} \quad t < s_{n_s} \\ y(t) & \text{for} \quad t \geqslant s_{n_s}. \end{cases} \tag{4.22c}$$

are called the deduced functions *of the function* y.

The deduced functions are defined in such a way that they are equal to the function y in parts of the interval $(-\infty, \tilde{t}]$. In particular, $z_1(t)$ is equal to $y(t)$ for $t \in (-\infty, s_1)$, $z_i(t)$ is equal to $y(t)$ for $t \in [s_{i-1}, s_i)$ for $2 \leqslant i \leqslant n_s$, and $z_{n_s+1}(t)$ is equal to $y(t)$ for $t \in [s_{n_s}, \tilde{t}]$. Outside of the set where a deduced function equals $y(t)$ it is continued in such a way that it is continuously differentiable and Lipschitz continuous on $t \in (-\infty, \tilde{t}]$. Any other definition of deduced functions that continues the function $y(t)$ in a different, but continuously differentiable and Lipschitz continuous manner, would also be suitable for the remainder of this chapter.

With the help of Definitions 4.9 and 4.10 it is possible to define a *consistent choice of discontinuity interval indicators*, which transfers the idea behind Definition 4.7 (consistent switching function signs) to the treatment of DDE-IVPs.

Definition 4.11 (Consistent Choice of Discontinuity Interval Indicators)

Consider a DDE-IVP according to Definition 1.12. Let $\tilde{t} \in \mathcal{T}(c)$, *and let* $y \in \mathscr{PD}((-\infty, \tilde{t}], \mathcal{D}^y)$ *be some piecewise continuously-differentiable function. Let the time points of discontinuity of order* 0 *in* y *be denoted by* s_i, *with* $1 \leqslant i \leqslant n_s$, *where* n_s *is the total number, and let them be sorted ascendingly:* $s_1 < s_2 < \cdots < s_{n_s}$. *Further, let* $z_i : (-\infty, \tilde{t}] \to \mathcal{D}^y$, $1 \leqslant i \leqslant n_s + 1$, *be the continuously-differentiable deduced functions of the function* y.

Then $\xi^{\alpha\prime} = (\xi_1^{\alpha\prime}, \ldots, \xi_{n_\tau}^{\alpha\prime}) \in \mathcal{I}^{\xi^\alpha}$ *is called a* consistent choice of discontinuity interval indicators, *if there exists* $\Delta t \geqslant \underline{\Delta t} > 0$ *and a solution* $y_{\xi^{\alpha\prime}}(t)$ *of the ODE-IVP*

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c, \{z_{\xi_i^{\alpha\prime}}(t - \tau_i(t, \mathfrak{y}(t), c))\}_{i=1}^{n_\tau}) \tag{4.23a}$$

$$\mathfrak{y}(\tilde{t}) = y(\tilde{t}) \tag{4.23b}$$

$$\mathfrak{y}(t) = y(t) \quad \text{for} \quad t < \tilde{t} \tag{4.23c}$$

such that

$$\xi_i^\alpha(t) = n_s + 1 + \frac{1}{2} \sum_{j=1}^{n_s} (\zeta_{i,s_j}^{\alpha,+}(t) - 1) \tag{4.24a}$$

$$\equiv \xi_i^{\alpha\prime} \tag{4.24b}$$

with

$$\zeta_{i,s_j}^{\alpha,+}(t) = sign^+(\alpha_i(t, y_{\xi^{\alpha\prime}}^-(t), c) - s_j) \tag{4.25}$$

for $1 \leqslant i \leqslant n_\tau$ *and all* $t \in (\tilde{t}, \tilde{t} + \Delta t)$.

Observe that the deduced functions occur as argument in the differential equation (4.23a) in order to obtain an expression for the past state arguments that a) corresponds to the choice of $\xi^{\alpha\prime}$ and b) is independent of the value of the function y at the past time point given by the deviating argument.

For an interpretation of a consistent choice of discontinuity interval indicators, think again of a DDE-IVP with a single state-dependent delay $\tau_1(t, y(t), c)$ and a single discontinuity point $s_1 \leqslant t^{ini}(c)$. Let $y(t)$ be a solution of this DDE-IVP. Clearly, if for some $\tilde{t} \in \mathcal{T}(c)$ the state $y(\tilde{t})$ is such that the past time point is not equal to s_1, $\alpha_1(\tilde{t}, y(\tilde{t}), c) - s_1 \neq 0$, then due to the continuity

of α_1 and the continuity of DDE-IVP solutions there can be at most one consistent choice of the discontinuity interval indicator in the neighborhood of $(\tilde{t}, y(\tilde{t}))$.

Care must be taken, however, when the deviating argument is located exactly at the discontinuity point s_1. Here, in general both "assumed values" for the discontinuity interval indicator are possible for the continuation of the IVP solution, namely 1 (i.e. the deviating argument is to the left of s_1) and 2 (i.e. the deviating argument is to the right of s_1). Both assumed values lead to different solutions of the ODE-IVP (4.23), which may or may not be such that the actual discontinuity interval indicator equals the assumed value. If it does, it is called a consistent choice, otherwise it is called an inconsistent choice.

It is clear that for existence of a DDE-IVP solution it is necessary that there exists at least one consistent choice of the discontinuity interval indicators for all t, and conversely, that existence of a solution implies that at least one such consistent choice exists for all t. In order to formulate a theorem on the uniqueness of DDE-IVP solution, it is natural to request uniqueness of the discontinuity interval indicators in all time points t where any of the deviating arguments is located at one of the discontinuity points of order 0 in the past. This is the case in the following theorem.

Theorem 4.12 (Global Uniqueness of DDE-IVP Solutions)

Consider a DDE-IVP as in Definition 1.12 with a vector c of fixed parameter values and let $\mathcal{D}^y \subset \mathbb{R}^{n_y}$ be some open domain. Let $y \in \mathscr{PD}(\mathcal{T}^f(c), \mathcal{D}^y)$ be a solution of the problem, where $s_1 < s_2 < \cdots < s_{n_s} \leqslant t^{ini}(c)$ are the time points of discontinuity of order 0 in y. Further, define the set

$$\mathcal{D}^t_{\alpha,0}(\mathcal{T}(c)) := \{t \in \mathcal{T}(c) \backslash \{t^{ini}(c)\} \mid \alpha_i(t, y^-(t), c) = s_j \text{ for at least one } (i,j) \in \{1,\ldots,n_\tau\} \times \{1,\ldots,n_s\}\}, \tag{4.26}$$

i.e. $\mathcal{D}^t_{\alpha,0}(\mathcal{T}(c))$ represents all times for which at least one deviating argument is located at one of the time points of discontinuity of order 0 in y.

Assume that the following conditions are fulfilled

(C) Continuity*: The right-hand-side function $f(t, y, c, \{v_i\}_{i=1}^{n_\tau})$ is continuous with respect to t, y, and $\{v_i\}_{i=1}^{n_\tau}$ for $(t, y, \{v_i\}_{i=1}^{n_\tau}) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau}$. The delay functions $\tau_i(t, y, c)$, $1 \leqslant i \leqslant n_\tau$, are continuous with respect to t and y for $(t, y) \in \mathcal{T}(c) \times \mathcal{D}^y$.*

(L) Lipschitz-Continuity*: The right-hand-side function $f(t, y, c, \{v_i\}_{i=1}^{n_\tau})$ is uniformly Lipschitz continuous with respect to y and $\{v_i\}_{i=1}^{n_\tau}$ for $(t, y, \{v_i\}_{i=1}^{n_\tau}) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau}$. The delay functions $\tau_i(t, y, c)$ are uniformly Lipschitz continuous with respect to y for $(t, y) \in \mathcal{T}(c) \times \mathcal{D}^y$. The initial function ϕ is piecewise continuously differentiable, i.e. $\phi(\cdot, c) \in \mathscr{PD}(\mathcal{T}^f(c), \mathcal{D}^y)$, and between two successive discontinuity points it is uniformly Lipschitz continuous with respect to t.*

(B) Boundedness*: The right-hand-side function is bounded by*

$$\|f(t, y, c, \{v_i\}_{i=1}^{n_\tau})\|_\infty < M_f < \infty \tag{4.27}$$

for $(t, y, \{v_i\}_{i=1}^{n_\tau}) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau}$.

(NVD) Non-Vanishing Delays*: It holds for the delay functions τ_i, $1 \leqslant i \leqslant n_\tau$ that $\tau_i(t, y(t), c) \geqslant \underline{\tau} > 0$ for $t \in \mathcal{T}(c)$ and the considered solution $y(t)$.*

(CI) Consistent Indicators*: For $(t, y) = (t^{ini}(c), y^{ini}(c))$ and each $(t, y(t))$ with $t \in \mathcal{D}^t_{\alpha,0}(\mathcal{T}(c))$ there is exactly one consistent choice of the discontinuity interval indicators.*

Then the solution $y(t)$ of the DDE-IVP is unique on the interval $\mathcal{T}(c)$.

Proof

The proof works by contradiction. Assume that there exists a solution $\bar{y}(t)$ of the DDE-IVP on $\mathcal{T}^f(c)$ different from $y(t)$. However, $\bar{y}$ and y are identical for $t \leqslant t^{ini}(c)$, and there exists a unique consistent choice $\xi^{\alpha'}$ of the discontinuity interval indicators at the initial time. Consequently, for some interval $(t^{ini}(c), t^{ini}(c) + \Delta t)$ the past state arguments can be replaced by evaluations of the same deduced functions of y. Due to assumption (NVD), the DDE-IVP is locally equivalent to an ODE-IVP with right-hand-side function

$$f_{ODE}(t, \mathfrak{y}(t), c) = f(t, \mathfrak{y}(t), c, z_{\xi_i^{\alpha'}}(t - \tau_i(t, \mathfrak{y}(t), c))) \tag{4.28}$$

Due to Lipschitz continuity of f, ϕ, and τ_i, the function $f_{ODE}(t, y, c)$ is Lipschitz continuous with respect to y, and because of assumption (B) it is also bounded. Hence in some sufficiently small interval $[t^{ini}(c), t^{ini}(c) + \delta t]$, Theorem 4.2 can be applied, which gives $\bar{y}(t) = y(t)$.

Since it holds for the given solution that $y(t) \in \mathcal{D}^y$, with $\mathcal{D}^y$ being an open set, the same argument can be applied for a sequence of intervals and closed neighborhoods on the interval $[t^{ini}(c), s_{n_s+1})$, where s_{n_s+1} denotes the earliest time point where any of the deviating arguments becomes equal to one of the time points of discontinuity of order 0 in the past. At s_{n_s+1}, the state is continuous, because DDE-IVP solutions have to be continuous for $t \in \mathcal{D}_1^t(\mathcal{T}(c))$. Hence, $\bar{y}(t) = y(t)$ for $t \leqslant s_{n_s+1}$.

For times to the right of s_{n_s+1} the assumption (CI) guarantees that there is a unique choice of the discontinuity interval indicators. Hence, an ODE-IVP can be formulated, where the past state arguments of the DDE right-hand-side function are evaluated at the current (possibly updated) set of deduced functions. The previous arguments can now be repeated, which results in the conclusion $\bar{y}(t) = y(t)$ on the full time interval $\mathcal{T}^f(c)$, so the solution is unique. ■

Note that, similar to Theorem 4.8, the assumptions (C), (L) and (B) are formulated for a domain $\mathcal{D}^y \subset \mathbb{R}^{n_y}$, but that it is in fact sufficient if the assumptions hold in a tubular neighborhood of the considered solution $y(t)$.

The conditions (NVD) and (CI), which are the main differences in the assumptions of Theorem 4.12 as compared to the related ODE-Theorem 4.2, can easily be checked in practice (at least in principle). The check is obvious for condition (NVD). In order to check condition (CI) it is required to go, at a time point where a deviating argument is located at a point of discontinuity of order 0 in the past, through all possible values of the discontinuity interval indicators and check whether equation (4.24) holds.

4.6. The General Case: IHDDEs

IHDDE-IVPs combine the difficulties of implicitly defined root discontinuities as they occur in IHODE-IVPs with difficulties related to the dependency of the right-hand-side function on past states as in DDE-IVPs. Accordingly, a theorem that guarantees uniqueness of an existing IHDDE-IVP solution has to combine the assumptions on the model functions that were formulated in the Theorems 4.8 and 4.12:

Theorem 4.13 (Global Uniqueness of IHDDE-IVP Solutions)

Consider an IHDDE-IVP as in Definition 1.2 with a vector c of fixed parameter values and let $\mathcal{D}^y \subset \mathbb{R}^{n_y}$ be some open domain. Let $y \in \mathscr{PD}(\mathcal{T}^f(c), \mathcal{D}^y)$ be a solution of the problem, where $s_1 < s_2 < \cdots < s_{n_s}$ are the time points of discontinuity of order 0 in y. Further, define the set

$$\mathcal{D}_{\alpha,0}^t(\mathcal{T}(c)) := \{t \in \mathcal{T}(c) \backslash \{t^{ini}(c)\} \mid \alpha_i(t, y^-(t), c) = s_j \text{ for at least one } (i,j) \in \{1, \ldots, n_\tau\} \times \{1, \ldots, n_s\}\}, \tag{4.29}$$

i.e. $\mathcal{D}_{\alpha,0}^t(\mathcal{T}(c))$ represents all times in which at least one deviating argument is located at one of the time points of discontinuity of order 0 in y.

Assume that the following conditions are fulfilled

- *(C)* Continuity*: The right-hand-side function $f(t, y, c, \{v_i\}_{i=1}^{n_\tau}, \zeta)$ is continuous with respect to t, y, and $\{v_i\}_{i=1}^{n_\tau}$ for $(t, y, \{v_i\}_{i=1}^{n_\tau}, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_1^\zeta$. The switching functions $\sigma_i(t, y, c, \{v_j\}_{j=1}^{n_\tau})$ are continuous with respect to t, y, and $\{v_j\}_{j=1}^{n_\tau}$ for $(t, y, \{v_j\}_{j=1}^{n_\tau}) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau}$. The delay functions $\tau_i(t, y, c)$, $1 \leqslant i \leqslant n_\tau$, are continuous with respect to t and y for $(t, y) \in \mathcal{T}(c) \times \mathcal{D}^y$.*

- *(L)* Lipschitz-Continuity*: The right-hand-side function $f(t, y, c, \{v_i\}_{i=1}^{n_\tau}, \zeta)$ is uniformly Lipschitz continuous with respect to y and $\{v_i\}_{i=1}^{n_\tau}$ for $(t, y, \{v_i\}_{i=1}^{n_\tau}, \zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_1^\zeta$. The delay functions $\tau_i(t, y, c)$ are uniformly Lipschitz continuous with respect to y for $(t, y) \in \mathcal{T}(c) \times \mathcal{D}^y$. The initial function ϕ is piecewise continuously differentiable, i.e. $\phi(\cdot, c) \in \mathscr{PD}(\mathcal{T}^f(c), \mathcal{D}^y)$, and between two successive discontinuity points it is Lipschitz continuous with respect to t.*

(B) Boundedness*: The right-hand-side function is bounded by*

$$\|f(t,y,c,\{v_i\}_{i=1}^{n_\tau},\zeta)\|_\infty < M_f < \infty \tag{4.30}$$

for $(t,y,\{v_i\}_{i=1}^{n_\tau},\zeta) \in \mathcal{T}(c) \times \mathcal{D}^y \times (\mathcal{D}^y)^{n_\tau} \times \mathcal{I}_1^\zeta$.

(NVD) Non-Vanishing Delays*: It holds for the delay functions* τ_i, $1 \leqslant i \leqslant n_\tau$ *that* $\tau_i(t,y(t),c) \geqslant \underline{\tau} > 0$ *for* $t \in \mathcal{T}(c)$ *and the considered solution* $y(t)$.

(CIS) Consistent Indicators and Signs*: For* $(t,y) = (t^{ini}(c), y^{ini}(c))$ *and for each* $(t,y(t))$ *with* $t \in \mathcal{D}_0^t(\mathcal{T}(c)) \cup \mathcal{D}_{\alpha,0}^t(\mathcal{T}(c))$ *there is exactly one consistent choice of the switching function signs and exactly one consistent choice of the discontinuity interval indicators.*

Then the solution $y(t)$ *is the unique solution of the IHDDE-IVP on the interval* $\mathcal{T}(c)$.

Proof
The proof is obtained by the obvious generalization of the proofs of the Theorems 4.8 and 4.12.■

5. Numerical Solution

> *The discontinuity of the dense output at ζ* (i.e. at a past time point of discontinuity) *may prevent fast convergence or even lead to divergence. [...] To avoid this difficulty, we propose to use the extrapolated dense output also for the arguments beyond ζ.*
>
> Guglielmi and Hairer, in their paper "Computing breaking points in implicit delay differential equations" [124], suggesting the use of extrapolations beyond past discontinuities.

In the previous chapter, existence and uniqueness results for solutions of initial value problems in impulsive hybrid discrete-continuous delay differential equations (IHDDE-IVPs) were presented. In practice, however, it is of course insufficient to know whether a solution $y(t)$ of a given IHDDE-IVP (or an IVP in a simpler subclass of differential equations) exists, but it is necessary to determine the solution in some way. For many real-world applications, the IVP solutions cannot be obtained analytically, but it is instead necessary to rely on numerical methods for the computation of an approximate solution.

The output of the numerical method may either be a sequence of numerical approximations y_l of $y(t_l)$ at a finite set of mesh points t_l (in the case of a so-called *discrete integration method*), or, preferably, a continuous approximation $\eta(t)$ of $y(t)$ for $t^{ini}(c) \leqslant t \leqslant t^{fin}(c)$ (in the case of a so-called *continuous integration method*). The presentation of discrete and continuous integration methods for solving IVPs in differential equations as well as the analysis of these methods – in particular, the convergence of the approximations y_l and $\eta(t)$ to the exact values $y(t_l)$ and $y(t)$ – is the topic of this chapter.

Literature Survey

For initial value problems in ordinary differential equations (ODE-IVPs), the literature on numerical methods is very rich. The reader is therefore referred, e.g., to the standard textbooks by Hairer, Nørsett, and Wanner [126], Hairer and Wanner [127], Butcher [55], Stoer and Bulirsch [241], Deuflhard and Bornemann [76], Petzold and Ascher [208] and the references therein.

With respect to ordinary differential equations with discontinuous right-hand-side functions, some papers have discussed the general case that the numerical method has no information on the location of discontinuities, see Gear and Østerby [112], Enright et al. [98], and Calvo, Montijano, and Rández [57]. The greater fraction of works, which is referenced further below, assumes that the locations of discontinuities are characterized as zeros of switching functions and that the numerical method has access to the switching functions. Away from the zero sets of the switching functions, the right-hand-side function is assumed to be "sufficiently" smooth, where the meaning of "sufficiently" is related to the smoothness requirements of the numerical method. These assumptions are also made in this thesis, and it is recalled that the use of switching functions as indicators for the discontinuity points corresponds to what is in this thesis called a hybrid discrete-continuous ordinary differential equation (HODE).

The main part of Chapter 2 was concerned with the issue of finding a definition of a solution. As discussed therein, this thesis is concerned with solutions for which the switching functions have finitely many roots, which excludes, in particular, so-called Filippov solutions. Accordingly, this chapter is restricted to the topic of numerically computing solutions that obey this restriction, but it is appropriate to at least mention some of the works that have been concerned with the numerical computation of Filippov solutions: Piiroinen and Kuznetzov [210] and Dieci and Lopez [78].

For problems with a finite number of root discontinuities, it is clear that the time evolution of an HODE-IVP solution is that of an ODE-IVP solution between two successive root discontinuities. Hence, the numerical solution of an HODE-IVP has to be based on methods for the solution of ODE-IVPs. While taking an integration step with an ODE-method, one approach is to evaluate only smooth "branches" of the right-hand-side function f. This means that the numerical method

calls the right-hand-side function f, for all necessary evaluations in one integration step, with a fixed vector of switching function signs even if the switching function changes its sign during the integration step. It should be noted that the use of this approach relies on the mild assumption that the smooth "branch" is evaluable beyond the point of the root discontinuity (see Dieci and Lopez [79] for a numerical method that abstains from this mild assumption).

Evaluating smooth branches of f has been used since the 1970's and 1980's, see e.g. Hay, Crosbie, and Chaplin [143], Ellison [87], and Bock [37]. Later it has been called the *discontinuity locking* mechanism, see e.g. Park and Barton [200], Bahl and Linninger [8], and Compere [67]. This approach can be regarded as the standard procedure for the evaluation of f in the numerical solution of HODE-IVPs, see e.g. Bock, Schlöder, and Schulz [45] and the textbooks by Stoer and Bulirsch [241], page 184, and by Hairer, Nørsett, and Wanner [126], page 198.

It remains to discuss approaches for the treatment of the root discontinuities, which are, in the literature, occasionally called *events*. Numerical methods proposed in the literature typically distinguish between two phases: First, recognizing that a switch has occured since the last mesh point (*event detection*), and second, locating the switch (*event location*). The methods that can be used for these tasks are intimitely related to the nature of integration method in use.

- Discrete integration methods: It is possible to construct HODE-IVP solvers on entirely discrete integration methods, i.e. methods that approximate the solution $y(t)$ of the HODE-IVP only at the mesh point.

  For example, one may check after, each integration step, whether the signs of the switching functions have changed. If not, the integration is continued. Otherwise, the location of the root discontinuity can be approximated by linearly interpolating the values of the switching function at the mesh points, and determining the zero of this linear function. This approach is taken in the early paper by Hay, Crosbie, and Chapin [143]. By using, in addition, also the time derivatives of the switching function at the mesh points, a higher order interpolation polynomial can be constructed and used, see Cellier [61]. Optionally, the zero of a higher order interpolation polynomial can only be used as an initial guess for a subsequent iterative procedure that employs additional trial steps with the discrete integration method, see Bulirsch [51].

- Continuous integration methods: A more straightforward approach to the numerical solution of HODE-IVPs is the use of continuous integration methods that provide, besides the approximation of the solution at the mesh points, also a continuous approximation of the solution between the mesh points. Such a continuous approximation is usually called *continuous representation* or *dense output*.

  Using continuous integration methods implies the opportunity to evaluate the switching functions for time points that are not part of mesh. This allows to apply any zero finding procedure for the determination of the discontinuity point and has become the standard technique for event location, see Park and Barton [200], Meijaard [188], Kirches [160], Wunderlich [261], and also the textbooks by Eich-Soellner and Führer [86], page 202, and Stoer and Bulirsch [241], page 184. In addition, continuously evaluable switching functions can also be used as a basis for the development of sophisticated event detection strategies that aim at detecting multiple zero crossings of a single switching function in a single integration step. For works in this direction, it is referred to Shampine, Gladwell, and Brankin [231] and Park and Barton [200].

The historical developments of numerical methods for solving initial value problems in delay differential equations (DDE-IVPs) and HODE-IVPs exhibit some similarities with regard to the fact that several early approaches are based on discrete integration methods.

For example, Bellman [27] and Bellman, Buell, and Kalaba [29] present the technique that is today referred to as the *method of steps*, see also Remark 4.5. The method of steps can be realized in such a way that a sequence of ODE-IVPs of increasing dimension is solved. The benefit of this realization is that all past states that are needed for the evaluation of f are available as components of the augmented state vector at the current time. Hence, by using the method of steps, storing past values can be avoided and any standard ODE-IVP solver can be used.

Another approach that allows to apply discrete integration methods is the use of so-called *constrained meshes*, which is suggested, in El'sgol'ts [90], page 284, El'sgol'ts [91], page 165, and elaborated in Cryer [72]. In this approach, the approximate discrete solution at a finite number of

time points is stored, and the stepsizes of the integration method are chosen in such a way that the past states that are needed in the current step are among the stored values.

Both the method of steps and the use of constrained meshes have severe disadvantages. First, they are not suitable for all classes of DDE-IVPs, e.g. the application to DDE-IVPs with state-dependent delays or vanishing delays is not possible. Second, they are inherently inefficient. The method of steps, on the one hand, requires a redundant computation of the DDE-IVP solution on the same interval over and over again. On the other hand, the use of constrained meshes does not allow the use of so-called variable-stepsize strategies, which are crucial for the efficiency of an integration method.

An alternative approach for the numerical solution of DDE-IVPs is to use continuous integration methods. Early works that rely on this approach are Neves [193], Oppelstrup [198], Bock and Schlöder [43], and Oberle and Pesch [196]. Later, this idea has been adopted by Paul [201], Enright and Hayashi [96], Guglielmi and Hairer [122] Shampine and Thompson [233], and many others. These use of continuous integration methods has the big advantage that they can be combined with variable-stepsize strategies in a straightforward way. The vast majority of modern algorithms and computer codes make use of continuous integration methods, therefore it has become custom to call it the *standard approach* for solving DDE-IVPs, see Bellen and Zennaro [26].

As discussed in Chapter 2, DDE-IVP solutions typically exhibit propagated discontinuities. It is well-known that in order to obtain accurate numerical approximations of DDE-IVP solutions, the numerical method needs to include the time points of the propagated discontinuities into the mesh. However, there are different approaches on how to handle this issue in practice.

One approach is to rely on the assumption that the lack of smoothness will lead to repeated stepsize rejections in the vicinity of the time point of a propagated discontinuity. Based on this assumption, Oppelstrup [198], Enright and Hayashi [96], and Shampine [230] have developed algorithms that analyze the sequence of accepted and rejected stepsizes. If the analysis raises suspicion that a discontinuity may be present, it is attempted to include the discontinuity point approximately into the mesh.

In other papers, a rigorous *tracking of discontinuities* has been favoured, i.e. including all points of discontinuity up to the order of the method into the mesh by finding the zeros of the propagation switching functions. This approach has been taken, e.g., in Bock and Schlöder [43], Feldstein and Neves [103], Willé and Baker [256], and Paul [204]. Furthermore, Guglielmi and Hairer [123] suggest a "relaxed" variant of discontinuity tracking, i.e. an algorithm that calls the root finding strategy only if certain conditions are met, e.g. if a stepsize has been rejected.

It is remarkable that even those research works that advocate the use of discontinuity tracking do typically not describe or mention the DDE-analogon of the discontinuity locking mechanism. Algorithms that track discontinuities but do not use discontinuity locking ensure that the time points of the propagated discontinuities are included in the mesh, but the computation of past states for the trial stepsizes is done in such a way that the smoothness assumptions of the employed numerical method are violated. To the best of the authors knowledge, only the recent works by Guglielmi and Hairer [124], ZivariPiran [271], ZivariPiran and Enright [272], and Ernst [101] have proposed to use the DDE-analogon of discontinuity locking, which means to use extrapolations beyond past discontinuity points if a deviating argument crosses such a past discontinuity point in the present integration step.

Novel Results Presented in This Chapter

The use of extrapolations beyond past discontinuity points (as suggested in Guglielmi and Hairer [124], ZivariPiran [271], ZivariPiran and Enright [272], and Ernst [101]) constitutes a solution approach for DDE-IVPs that is not any longer a realization of the standard approach. In this chapter, the use of extrapolations is therefore formally introduced as an integral part of a new solution approach called the *modified standard approach*. This modified standard approach is formulated in two versions. First, as an "idealized" variant that relies on the assumption that the discontinuity points of the exact solution are known and included into the mesh. This assumption can typically only be fulfilled for constant and time-dependent delays. Therefore, as a second variant, a practically realizable version of the modified standard approach is presented that employs numerically determined approximations of the discontinuity points.

For the idealized variant of the modified standard approach novel theoretical results are presented that ensure, for the case of continuous Runge-Kutta (CRK) methods, existence and uniqueness of the numerical solution in each integration step and convergence of the numerical solution to the

exact solution. The presented theorems are also applicable to problems where the initial function has discontinuities of order 0. Further, a subtle pitfall in the convergence proof for the standard approach in Bellen and Zennaro [26] is described, which can effectively be circumvented by the use of extrapolations.

The practical variant of the modified standard approach is defined, and its convergence properties and aspects for the implementation are discussed, also in view of the more general case of IHDDE-IVPs.

Organization of This Chapter

Section 5.1 presents basic definitions and results for continuous one-step methods for ODEs, with a specific focus on the popular subclass of CRK methods. In Section 5.2 the application of continuous one-step methods to DDE-IVPs is discussed. The standard approach for solving DDE-IVPs is recalled, and the idealized variant of the modified standard approach is introduced. Well-posedness of the numerical method and convergence to the exact solution is shown for CRK methods realized in the framework of the modified standard approach.

Section 5.3 discusses the extension of the modified standard approach for the solution of IHDDE-IVPs. The practical variant of the modified standard approach is introduced in Section 5.4. Eventually, Section 5.5 discusses several practically important aspects like error estimation, error control, and efficient selection of stepsizes.

Notation

The solution of DDE-IVPs with numerical methods often requires a higher degree of smoothness of the model functions than theorems regarding the existence and uniqueness of solutions. Therefore, in order to express the necessary assumptions as compact as possible, the following notation is used in this chapter. For any function $g : (x_1, x_2) \to \mathbb{R}^{n_g}$, $x_1 \in \mathbb{R}^{n_{x_1}}$, $x_2 \in \mathbb{R}^{n_{x_2}}$, the notation $g(\cdot,\cdot) \in \mathscr{C}^p(A_{x_1} \times A_{x_2}, \mathbb{R}^{n_g})$ means that the function g is p-times continuously differentiable with respect to both arguments x_1 and x_2 on the sets A_{x_1}, A_{x_2}. Further, the notation $g(\cdot, x_2) \in \mathscr{C}^p(A_{x_1}, \mathbb{R}^{n_g})$ means that for a given fixed x_2, the function g is p-times continuously differentiable with respect to x_1 on the set A_{x_1}.

Further, it is remarked that within this chapter, the symbol $\|\cdot\|$ represents any of the norms $\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$ on a finite-dimensional space.

The Landau symbol $\mathscr{O}(g(x))$ is used throughout the chapter, where $f(x) = \mathscr{O}(g(x))$ for $x \to a$ means that $\limsup_{x\to a} |f(x)/g(x)| < \infty$. Equivalently, $f = \mathscr{O}(g(x))$ can also be understood as the existence of $C > 0$ and $\epsilon > 0$ such that for all x with $|x - a| < \epsilon$ it holds that $|f(x)| \leqslant C|g(x)|$.

5.1. Continuous One-Step Methods for ODE-IVPs

Consider an ODE-IVP as in Definition 1.4, i.e.

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c) \tag{5.1a}$$

$$\mathfrak{y}(t^{ini}(c)) = y^{ini}(c). \tag{5.1b}$$

Throughout the chapter the parameters c shall thereby be arbitrary but fixed. Further, it is assumed that there exists a unique solution $y(t)$. The task is to find a numerical method for computing an approximation a) of the solution at some given final time, $y(t^{fin}(c))$, and b) of the solution $y(t)$ on the whole time interval $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$. For this purpose, a mesh $t_0 < t_1 < \cdots < t_{n_m}$ is considered with $t_0 = t^{ini}(c)$, $t_{n_m} = t^{fin}(c)$, where the total number of mesh points is $n_m + 1$.

Definition 5.1 (Discrete One-Step Method, Stepsize, Discrete Increment Function)

A numerical method for computing an approximation y_{l+1} of the solution $y(t_{l+1})$ at the mesh point t_{l+1} that is of the form

$$y_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; f), \tag{5.2}$$

is called a discrete one-step method. *The difference between two successive mesh points,* $h_{l+1} = t_{l+1} - t_l$*, for* $l \in \{0, \dots, n_m - 1\}$*, is called the* stepsize (in step $l+1$)*, and the function* Φ *is called the* discrete increment function.

The method is called *discrete*, because it provides approximations of $y(t)$ only at the mesh points t_l, and it is called a *one-step method*, because the right hand side of equation (5.2) depends only on the mesh point t_l and on the state y_l at this mesh point, but not on states $y_{l'}$ or mesh points $t_{l'}$ for $l' < l$. The latter would be the case for so-called multi-step methods, in particular for linear multi-step methods and backward differentiation formulae, see e.g. Hairer and Wanner [127].

With regard to the stepsizes h_l, it is distinguished between *fixed stepsizes* and *variable stepsizes*.

Definition 5.2 (Fixed Stepsize, Variable Stepsize)

If the stepsize is identical for all steps, $h_l \equiv h$ *for* $1 \leqslant l \leqslant n_m$*, then the method is called a* fixed-stepsize method. *Otherwise it is called a* variable-stepsize method.

For computing an approximation of $y(t^{fin}(c))$ it is sufficient to use a discrete method, whereas an approximation of the solution $y(t)$ for all $t \in \mathcal{T}(c)$, requires to endow the discrete one-step method with a *continuous extension*.

Definition 5.3 (Continuous Extension, Continuous Increment Function)

A continuous extension *of the one-step method is a function* $\eta : \mathcal{T}(c) \to \mathbb{R}^{n_y}$ *that is defined piecewise by* $\eta(t) = \eta_{l+1}(t)$ *on the interval* $[t_l, t_{l+1}]$ *for* $0 \leqslant l \leqslant n_m - 1$*, and* $\eta_{l+1}(t)$ *is of the form*

$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; f), \tag{5.3}$$

where $\theta \in [0, 1]$*. The function* $\eta_{l+1}(t)$ *satisfies the continuity conditions*

$$\eta_{l+1}(t_l) = y_l \quad \textit{and} \quad \eta_{l+1}(t_{l+1}) = y_{l+1}. \tag{5.4}$$

The function Ψ *is called the* continuous increment function.

Definition 5.4 (Continuous One-Step Method)

The discrete one-step method of Definition 5.1, together with a continuous extension of Definition 5.3, is called a continuous one-step method.

In practical methods, the function Ψ is typically a polynomial function of the variable θ, and for the remainder of this thesis only polynomial functions Ψ are considered. Accordingly, the continuous extension $\eta(t)$ is a piecewise polynomial function, and the polynomials are continuously linked at the mesh points.

Due to the continuity conditions (5.4) on the continuous extension, it is clear that the increment functions fulfill the relations

$$\Psi(t_l, y_l, h_{l+1}, 0; f) = 0 \tag{5.5a}$$
$$\Psi(t_l, y_l, h_{l+1}, 1; f) = \Phi(t_l, y_l, h_{l+1}; f). \tag{5.5b}$$

In order for a continuous one-step method to be useful, it is clearly necessary that, for $h_l \to 0$, and $n_m \to \infty$, and $\sum_{l=1}^{n_m} h_l = t^{fin}(c) - t^{ini}(c) = const.$, it should hold that $y_l \to y(t_l)$ for all $l = 0, \dots, n_m$, and that $\eta(t) \to y(t)$ for $t \in \mathcal{T}(c)$, i.e. the discrete and continuous approximations should converge to the unique solution (subsequently also called "exact solution") $y(t)$ of the ODE-IVP. Consistency of one-step methods as defined below is a crucial property in this context.

Definition 5.5 (Consistency, Discrete Local Order, Uniform Local Order)

Let $p \geqslant 1$ *be the largest integer number such that for all ODE-IVPs with* $f(\cdot, \cdot, c) \in \mathscr{C}^p(\mathcal{T}(c) \times \mathbb{R}^{n_y}, \mathbb{R}^{n_y})$ *and for all* $l = 0, \dots, n_m - 1$ *it holds that*

$$\|u_{l+1}(t_{l+1}) - y_{l+1}\| = \mathscr{O}(h_{l+1}^{p+1}), \tag{5.6}$$

where $u_{l+1}(t)$ *is the exact solution of the local ODE-IVP*

$$\dot{\mathbf{u}}_{l+1}(t) = f(t, \mathbf{u}_{l+1}(t), c) \tag{5.7a}$$
$$\mathbf{u}_{l+1}(t_l) = y_l, \tag{5.7b}$$

and y_{l+1} is the numerical approximation of $u_{l+1}(t_{l+1})$ computed with a discrete one-step method, i.e.

$$y_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; f). \tag{5.8}$$

Then p is called the discrete order of consistency *or the* discrete local order *of the one-step method, and the one-step method is called* consistent (of discrete order p).

Similarly, let $q \geqslant 1$ be the largest integer number such that for all ODE-IVPs with $f(\cdot,\cdot,c) \in \mathscr{C}^q(\mathcal{T}(c) \times \mathbb{R}^{n_y}, \mathbb{R}^{n_y})$ and for all $l = 0, \ldots, n_m - 1$ it holds that

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|u_{l+1}(t) - \eta_{l+1}(t)\| = \mathscr{O}(h_{l+1}^{q+1}), \tag{5.9}$$

where $\eta_{l+1}(t)$ is the continuous numerical approximation of $u_{l+1}(t)$ on $[t_l, t_{l+1}]$ given by the continuous extension of the continuous one-step method, i.e.

$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; f). \tag{5.10}$$

Then q is called the uniform order of consistency *or the* uniform local order *of the one-step method, and the one-step method is called* consistent (of uniform order q).

Due to the continuity conditions (5.4), the maximum error in equation (5.9) on the interval $[t_l, t_{l+1}]$ can, at best, go to zero with the same order of h_{l+1} as the error at the end of the interval, which is given by equation (5.6). Hence, it is clear that $q \leqslant p$.

It is noted that the formulation $f(\cdot,\cdot,c) \in \mathscr{C}^p(\mathcal{T}(c) \times \mathbb{R}^{n_y}, \mathbb{R}^{n_y})$ means that the partial derivatives of f with respect to both t and y exist and are continuous up to derivative order p. However, the p-th order total derivative of $f(t, y(t), c)$ with respect to t requires, by application of the chain rule, the existence of $d^p y(t)/dt^p$, which is given by the $p-1$-th order total derivative of f. Hence, by recursion, the existence of the p-th order partial derivatives also implies the existence of the p-th order total derivative of f (which is the $p+1$-th order total derivative of y) with respect to t.

The concept of consistency allows to formulate a convergence theorem for continuous one-step methods.

Theorem 5.6 (Convergence of Continuous One-Step Methods for ODE-IVPs, Convergence Order)

Consider the ODE-IVP (5.1) for some arbitrary but fixed parameter values c, and a continuous one-step method defined by equations (5.2), (5.3). Let the following assumptions be fulfilled:

- *(S)* Smoothness (of the right-hand-side function)*: The right-hand-side function f of the ODE-IVP is such that $f(\cdot,\cdot,c) \in \mathscr{C}^p(\mathcal{T}(c) \times \mathbb{R}^{n_y}, \mathbb{R}^{n_y})$.*
- *(L)* Lipschitz continuity (of the increment functions)*: The increment function $\Phi(t, y, h; f)$ and $\Psi(t, y, h, \theta; f)$ are Lipschitz continuous with respect to y.*
- *(E)* Existence of a unique solution*: There exists a unique solution $y(t)$ of the ODE-IVP.*
- *(C)* Consistency*: The continuous one-step method is consistent of discrete local order p and consistent of uniform local order q.*

Then it holds that

$$\max_{1 \leqslant l \leqslant n_m} \|y(t_l) - y_l\| = \mathscr{O}(h^p), \tag{5.11}$$

and the one-step method is called convergent with discrete order p. *Alternatively, it is said that the one-step method has* discrete global order p. *Further, it holds that*

$$\max_{t \in \mathcal{T}(c)} \|y(t) - \eta(t)\| = \mathscr{O}(h^r) \tag{5.12}$$

with $r = \min(p, q+1)$, and the one-step method is called convergent with uniform order r. *Alternatively, it is said that the one-step method has* uniform global order r.

Proof

See Theorem 3.2.8. in Bellen and Zennaro [26], page 44ff. ■

5.1.1. Continuous Runge-Kutta Methods

As an important subclass of continuous one-step methods, continuous Runge-Kutta methods (CRK) are introduced.

Definition 5.7 (Discrete Runge-Kutta Method, Continuous Runge-Kutta Method)

A continuous one-step method in which the increment functions take the special form

$$\Phi(t_l, y_l, h_{l+1}; f) = \sum_{i=1}^{\nu} \beta_i f(t_{l+1}^i, y_{l+1}^i, c) \tag{5.13a}$$

$$\Psi(t_l, y_l, h_{l+1}, \theta; f) = \sum_{i=1}^{\nu} b_i(\theta) f(t_{l+1}^i, y_{l+1}^i, c), \tag{5.13b}$$

with

$$t_{l+1}^i = t_l + \gamma_i h_{l+1} \tag{5.14a}$$

$$y_{l+1}^i = y_l + h_{l+1} \sum_{j=1}^{\nu} a_{i,j} f(t_{l+1}^j, y_{l+1}^j, c) \tag{5.14b}$$

for $1 \leqslant i \leqslant \nu$ is called a ν-stage continuous Runge-Kutta method *(CRK method). Further, y_{l+1}^i are called the* stage values*, and the procedure $y_{l+1} = y_l + h_{l+1} \sum_{i=1}^{\nu} \beta_i f(t_{l+1}^i, y_{l+1}^i, c)$ is called a* discrete Runge-Kutta method*. The numbers $\gamma_i \in [0, 1]$ are called the* abscissae*, β_i are the* weights*, $b_i(\theta)$ are the* continuous weight functions *and $a_{i,j}$ are the* coefficients *of the CRK method.*

Definition 5.8 (Butcher Tableau)

The coefficients, weights, and abscissae of a discrete Runge-Kutta method can be expressed in a Butcher Tableau *as follows:*

$$\begin{array}{c|cccc} \gamma_1 & a_{11} & a_{12} & \dots & a_{1\nu} \\ \gamma_2 & a_{21} & a_{22} & \dots & a_{2\nu} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_\nu & a_{\nu 1} & a_{\nu 2} & \dots & a_{\nu\nu} \\ \hline & \beta_1 & \beta_2 & \dots & \beta_\nu \end{array}, \tag{5.15}$$

or, in short, as

$$\begin{array}{c|c} \gamma & A \\ \hline & \beta^T \end{array}. \tag{5.16}$$

In order to fulfill the continuity conditions for the increment functions Φ and Ψ (equations (5.5)) for arbitrary right-hand-side functions f, it is obvious that $b_i(0) = 0$ for $1 \leqslant i \leqslant \nu$ and that $b_i(1) = \beta_i$ for $1 \leqslant i \leqslant \nu$.

In general, e.g. if all $a_{i,j}$ are non-zero, the equations (5.14b) are implicit in the variables y_{l+1}^i. In this case, a fix-point method is needed to solve the equations. However, in the special case that $a_{i,j} = 0$ for $j \geqslant i$, then the equations are explicit and the stage values y_{l+1}^i can be computed by simple recursion. This special property leads to the following definition.

Definition 5.9 (Explicit CRK Method, Implicit CRK Method)

If $a_{i,j} = 0$ for $j \geqslant i$, then the CRK method is called explicit*, otherwise it is called* implicit*.*

For explicit CRK methods the existence of a numerical solution for y_{l+1}^i is trivial. For implicit CRK methods, it is stated by the following theorem.

Theorem 5.10 (Existence and Uniqueness of the Numerical Solution for Implicit CRK Methods)

Consider an implicit CRK method applied to the ODE-IVP (5.1) for some arbitrary but fixed parameter values c. Let the right-hand-side function $f(t, y, c)$ be continuous in t and y, and Lipschitz continuous with respect to y with Lipschitz constant L_f. If the stepsize h_{l+1} is chosen as

$h_{l+1} < 1/L_f a_{max}$ with $a_{max} := \max_{1\leqslant i\leqslant\nu} \sum_{j=1}^{\nu} |a_{i,j}|$, then there exists a unique solution of the equations (5.14b).

Proof
See Hairer, Nørsett, and Wanner [126], p. 206. ■

As an alternative to the computation of y_{l+1}^i for $1 \leqslant i \leqslant \nu$ as solution of equation (5.14b), it is also possible to define $g_{l+1}^i := f(t_{l+1}^i, y_{l+1}^i, c)$ and to determine g_{l+1}^i as solution of the following equation system:

$$g_{l+1}^i = f\left(t,\ y_l + h_{l+1} \sum_{j=1}^{\nu} a_{i,j} g_{l+1}^j,\ c\right), \quad 1 \leqslant i \leqslant \nu. \tag{5.17}$$

The quantities g_{l+1}^i are also referred to as *stage values*.

For implicit methods, the system (5.17) can be uniquely solved under the same conditions as those for the system (5.14b), i.e. under the conditions given in Theorem 5.10. For the practical solution of ODE-IVPs, the two formulations are equivalent, but for solving DDE-IVPs it is implementationally convenient to use g_{l+1}^i as variables of the CRK method, as is shown later in Section 6.4.

It is clear that a CRK method is practically useful only if the discrete and continuous approximations y_l and $\eta(t)$ converge to the exact ODE-IVP solution. Hence, the convergence theorem (Theorem 5.6) should apply. From the assumptions of this theorem, (L) and (C) are checked in the following for the special case of CRK methods.

With regard to the assumption (L), i.e. Lipschitz continuity of the increment functions Φ and Ψ with respect to y, the following lemma is considered.

Lemma 5.11 (Lipschitz Continuity of CRK Increment Functions)

Under the conditions of Theorem 5.10, the CRK increment functions Φ and Ψ are Lipschitz continuous with respect to their y argument, i.e. they fulfill the assumption (L) of Theorem 5.6.

Proof
Follows by considering the difference between two evaluations of the increment functions at different y arguments, i.e.

$$\begin{aligned} &\|\Psi(t_l, y_l^1, h_{l+1}, \theta; f) - \Psi(t_l, y_l^2, h_{l+1}, \theta; f)\| \\ &= \left\| \sum_{i=1}^{\nu} b_i(\theta) f(t_{l+1}^i, \{y^1\}_{l+1}^i, c) - \sum_{i=1}^{\nu} b_i(\theta) f(t_{l+1}^i, \{y^2\}_{l+1}^i, c) \right\| \end{aligned} \tag{5.18}$$

where

$$\{y^1\}_{l+1}^i = y_l^1 + h_{l+1} \sum_{j=1}^{\nu} a_{i,j} f(t_{l+1}^j, \{y^1\}_{l+1}^j, c) \tag{5.19a}$$

$$\{y^2\}_{l+1}^i = y_l^2 + h_{l+1} \sum_{j=1}^{\nu} a_{i,j} f(t_{l+1}^j, \{y^2\}_{l+1}^j, c) \tag{5.19b}$$

are the stage values for two different approximations y_l^1 and y_l^2 at the mesh point t_l. Please note that the curly braces, e.g. in $\{y^1\}_{l+1}^i$, are used here in order to visually distinguish the inner index (which gives the index of the argument of the increment function) from the outer indices (which correspond, as usual, to the integration step and to the stage).

With the Lipschitz continuity of f it follows that

$$\|\Psi(t_l, y_l^1, h_{l+1}, \theta; f) - \Psi(t_l, y_l^2, h_{l+1}, \theta; f)\| \leqslant b_{max} L_f \max_{1\leqslant i\leqslant\nu} \|\{y^1\}_{l+1}^i - \{y^2\}_{l+1}^i\|, \tag{5.20}$$

where $b_{max} = \max_{1\leqslant i\leqslant\nu, 0\leqslant\theta\leqslant 1} |b_i(\theta)|$.

Further, from the definitions of $\{y^1\}_{l+1}^i$ and $\{y^2\}_{l+1}^i$ and with the Lipschitz condition on f, the following relation follows:

$$\max_{1\leqslant i\leqslant\nu} \|\{y^1\}_{l+1}^i - \{y^2\}_{l+1}^i\| \leqslant \|y_l^1 - y_l^2\| + h_{l+1} L_f a_{max} \max_{1\leqslant i\leqslant\nu} \|\{y^1\}_{l+1}^i - \{y^2\}_{l+1}^i\|. \tag{5.21}$$

Herein, $a_{\max} = \max_{1 \leqslant i \leqslant \nu} \sum_{j=1}^{\nu} |a_{i,j}|$. From $h_{l+1} < 1/(L_f \cdot a_{max})$ it follows that $1 - h_{l+1} L_f a_{\max} \neq 0$, and hence

$$\max_{1 \leqslant i \leqslant \nu} \|\{y^1\}_{l+1}^i - \{y^2\}_{l+1}^i\| \leqslant \frac{1}{1 - h_{l+1} L_f a_{\max}} \|y_l^1 - y_l^2\| \tag{5.22}$$

with a prefactor less than infinity. Insertion of this expression into equation (5.20) yields, as desired,

$$\|\Psi(t_l, y_l^1, h_{l+1}, \theta; f) - \Psi(t_l, y_l^2, h_{l+1}, \theta; f)\| \leqslant b_{max} L_f \frac{1}{1 - h_{l+1} L_f a_{\max}} \|y_l^1 - y_l^2\|, \tag{5.23}$$

i.e. Lipschitz continuity of Ψ with respect to y_l with Lipschitz constant $L_\Psi = b_{max} L_f / (1 - h_{l+1} L_f a_{\max}) < \infty$. With an analogous derivation, the Lipschitz continuity of Φ can be shown as well. ■

In order to obtain a convergence result for CRK methods, it remains to discuss the consistency condition (C) of Theorem 5.6.

Lemma 5.12 (Order Conditions for CRK Methods)

Consider the following equations for the abscissae, coefficients, and continuous weight functions of the CRK method:

$$\sum_{i=1}^{\nu} b_i(\theta) = \theta \tag{5.24a}$$

$$\sum_{i=1}^{\nu} b_i(\theta) \gamma_i = \frac{1}{2} \theta^2 \tag{5.24b}$$

$$\sum_{i=1}^{\nu} b_i(\theta) \gamma_i^2 = \frac{1}{3} \theta^3 \tag{5.24c}$$

$$\sum_{i=1}^{\nu} b_i(\theta) a_{i,j} \gamma_i = \frac{1}{6} \theta^3. \tag{5.24d}$$

If $(a_{i,j}, b_i(\theta), \gamma_{i,j})$ of a CRK method satisfy

- *the first of these equations, then it has at least uniform local order $q = 1$,*
- *the first and the second of these equations, then it has at least uniform local order $q = 2$,*
- *all four equations, then it has at least uniform local order $q = 3$.*

Analogous conditions for the discrete local order p of a method are obtained by setting $\theta = 1$ and by recalling that $\beta_i = b_i(1)$ for $1 \leqslant i \leqslant \nu$.

Proof
See Hairer, Nørsett, and Wanner [126], and Hairer, Wanner, and Lubich [128] for the conditions for discrete Runge-Kutta methods, and Bellen and Zennaro [26] for the extension to continuous Runge-Kutta methods. ■

The given references also contain conditions for CRK methods of order 4. In addition, Hairer Nørsett, Wanner [126] and Hairer, Wanner, and Lubich [128] also present a general technique to derive the conditions for orders $\geqslant 5$.

In the literature, many discrete Runge-Kutta methods can be found that are not endowed with a continuous extension. Then an apparent question is whether, for a given discrete Runge-Kutta method of order p, there exists a continuous extension which has at least uniform local order $q = 1$ or possibly even $q = p$. Two general results in this context are as follows:

- Without additional stages, it is always possible to construct a continuous extension of order $q = \lfloor \frac{p+1}{2} \rfloor$, see Bellen and Zennaro [26], page 118.

- Further, by allowing additional stages, it is always possible to construct, successively, continuous extensions of higher order until a uniform order of consistency $q = p$ is reached. However, additional stages formally lead to a CRK method with $\nu' > \nu$ stages, whose Butcher tableau is

$$\begin{array}{c|cc} \gamma & A & 0 \\ \tilde{\gamma} & \tilde{A}_1 & \tilde{A}_2 \\ \hline & \beta^T & 0 \end{array}, \tag{5.25}$$

  where $\tilde{\gamma}$ contains the additional abscissae. Since the weights β_j are zero for $\nu+1 \leqslant j \leqslant \nu'$, the discrete Runge-Kutta method is unaffected by the additional stages, whose result therefore does not change.

A particular method for constructing a continuous extension of order $q = p$ is the so-called uniform correction procedure developed by Zennaro [269] and Bellen and Zennaro [25, 26]. Since this method is used for the design of the numerical method implemented in Colsol-DDE, it is presented in detail in Section 6.2.

For the polynomial continuous extensions, the order q must, in general, not be identical to the polynomial degree δ. Some facts on the relation between q and δ are as follows.

- The order conditions (5.24) imply that $\delta \geqslant q$.
- If a CRK method with continuous extension of degree $\delta > q$ is given, then it is possible to construct a different continuous extension of degree q without a need for computing additional stage values (see Bellen and Zennaro [26], page 116).
- If, in addition to the assumptions of Theorem 5.6, the right-hand-side function $f(\cdot,\cdot,c)$ is $\mathscr{C}^{\max(\delta,p)}(\mathcal{T}(c)\times\mathbb{R}^{n_y},\mathbb{R}^{n_y})$, and it holds for the degree of the continuous extension that $\delta \geqslant q$, then

$$\max_{t\in\mathcal{T}(c)} \left\| \frac{d^j}{dt^j} y^j(t) - \frac{d^j}{dt^j}\eta(t) \right\| = \mathscr{O}(h^{q+1-j}) \quad \text{for} \quad 1 \leqslant j \leqslant \delta. \tag{5.26}$$

  At the mesh points, where $\eta(t)$ is only continuous, this relation holds for both the left-sided and the right-sided time derivatives. See Bellen and Zennaro [26], page 114 for the theorem and its proof.

These facts make clear that, on the one hand, $\delta \geqslant q$ is needed in order to obtain the uniform local order q, but that, on the other hand, $\delta \geqslant q+1$ is unnecessary and, for $\delta \geqslant q+2$, leads to divergent approximations of higher order time derivatives. Hence, it is generally preferable to use polynomial continuous extensions with degree $\delta = q$, and it is therefore assumed in the following that this is the case.

The section is concluded by a preparatory step toward the treatment of DDE-IVPs: the investigation of the error propagation in CRK methods applied to ODE-IVPs with an additional argument. It is noted that the following lemma is a variation of Bellen and Zennaro [26], page 84f. The difference is that the lemma presented here is not for general one-step methods but specialized to CRK methods. At the same time, it yields a stronger assertion for the error propagation, i.e. a more restrictive bound, and this more restrictive bound becomes important in the theory of CRK methods applied to DDE-IVPs (Section 5.2).

Lemma 5.13 (Limited Error Propagation)

Consider, at the mesh point t_l, two local IVPs for some arbitrary but fixed parameter values c, with different initial values y_l^1, y_l^2, and with different additional input functions $v_1(t)$, $v_2(t)$:

$$\dot{\mathbf{u}}_{l+1}^1(t) = f(t, \mathbf{u}_{l+1}^1(t), c, v^1(\alpha(t, \mathbf{u}_{l+1}^1(t)))) \tag{5.27a}$$
$$\mathbf{u}_{l+1}^1(t_l) = y_l^1 \tag{5.27b}$$

and

$$\dot{\mathbf{u}}_{l+1}^2(t) = f(t, \mathbf{u}_{l+1}^2(t), c, v^2(\alpha(t, \mathbf{u}_{l+1}^2(t)))) \tag{5.28a}$$
$$\mathbf{u}_{l+1}^2(t_l) = y_l^2, \tag{5.28b}$$

with $\alpha(t,y) \leqslant t$ for $\mathcal{T}(c) \times \mathbb{R}^{n_y}$, and observe that despite the dependence on an additional input function these IVPs are still ODE-IVPs. Further, consider a CRK method, applied to both these problems, for the step $t_l \to t_{l+1}$.

Let the following assumptions be fulfilled:

(S) Smoothness (of the model functions)*: It holds that $f(\cdot,\cdot,c,\cdot) \in \mathscr{C}^p(\mathcal{T}(c)\times\mathbb{R}^{n_y}\times\mathbb{R}^{n_y},\mathbb{R}^{n_y})$, and the right-hand-side function f is globally Lipschitz continuous with respect to the arguments y and v:*

$$\|f(t,y_1,c,v_1) - f(t,y_2,c,v_2)\| \leqslant L_f(\|y_1 - y_2\| + \|v_1 - v_2\|). \tag{5.29}$$

Further, v_1 and v_2 are globally Lipschitz continuous with Lipschitz constants L_{v_1} and L_{v_2}, and $\alpha(t,y)$ is globally Lipschitz continuous with respect to y with Lipschitz constant L_α.

(B) Boundedness (of the stepsize)*: Assume that the stepsize h_{l+1} is bounded by*

$$h_{l+1} \leqslant \frac{1}{(2L_f\alpha_{max}(1+\max(L_{v_1},L_{v_2})L_\alpha))}. \tag{5.30}$$

Let $\eta^1_{l+1}(t)$ and $\eta^2_{l+1}(t)$ denote the continuous numerical solutions and let y^1_{l+1} and y^2_{l+1} be the discrete numerical solutions of the ODE-IVPs obtained with the CRK method applied to the local IVPs (5.27) and (5.28), respectively:

$$\eta^i_{l+1}(t_l + \theta h_{l+1}) = y^i_l + h_{l+1}\sum_{j=1}^{\nu} b_j(\theta)\bar{f}^i(t^j_{l+1},\{y^i\}^j_{l+1},c), \quad i = 1,2 \tag{5.31a}$$

$$y^i_{l+1} = y^i_l + h_{l+1}\sum_{j=1}^{\nu} \beta_j\bar{f}^i(t^j_{l+1},\{y^i\}^j_{l+1},c), \qquad i = 1,2 \tag{5.31b}$$

$$\{y^i\}^j_{l+1} = y^i_l + h_{l+1}\sum_{k=1}^{\nu} \alpha_{j,k}\bar{f}^i(t^k_{l+1},\{y^i\}^k_{l+1},c), \qquad i = 1,2. \tag{5.31c}$$

Herein, the standard-form ODE right-hand-side functions $\bar{f}^1$ and $\bar{f}^2$ are defined by

$$\bar{f}^i(t,y,c) := f(t,y,c,v^i(\alpha(t,y))), \quad i = 1,2, \tag{5.32}$$

and curly braces have been used to visually separate the inner index i (of the considered local IVP) from the outer indices $l+1$ and j (for the integration step and CRK stage, respectively).

Then there exist constants A_i, $1 \leqslant i \leqslant 6$, only dependent on the various Lipschitz constants and $b_{max} = \max_{1\leqslant i\leqslant\nu, 0\leqslant\theta\leqslant 1}|b_i(\theta)|$, such that the following holds:

$$\|y^1_{l+1} - y^2_{l+1}\| \leqslant (1 + h_{l+1}A_1)\|y^1_l - y^2_l\| + h_{l+1}A_2 v_{diff} \tag{5.33a}$$

$$\max_{t_l\leqslant t\leqslant t_{l+1}} \|\eta^1_{l+1}(t) - \eta^2_{l+1}(t)\| \leqslant (1 + h_{l+1}A_3)\|y^1_l - y^2_l\| + h_{l+1}A_4 v_{diff} \tag{5.33b}$$

$$\max_{t_l\leqslant t\leqslant t_{l+1}} \left\| \frac{d^j}{dt^j}\left(\eta^1_{l+1}(t) - \eta^2_{l+1}(t)\right)\right\| \leqslant h^{1-j}_{l+1}A_5\|y^1_l - y^2_l\| + h^{1-j}_{l+1}A_6 v_{diff} \quad \textit{for} \quad 1 \leqslant j \leqslant q \tag{5.33c}$$

with

$$v_{diff} := \max_{1\leqslant i\leqslant\nu} \|v^1(\alpha(t^i_{l+1},\{y^2\}^i_{l+1})) - v^2(\alpha(t^i_{l+1},\{y^2\}^i_{l+1})\|. \tag{5.34}$$

Proof

By subtracting $\eta^1_{l+1}(t)$ and $\eta^2_{l+1}(t)$, the expression

$$\begin{aligned}
&\max_{t_l\leqslant t\leqslant t_{l+1}} \|\eta^1_{l+1}(t) - \eta^2_{l+1}(t)\| \\
&\leqslant \|y^1_l - y^2_l\| + h_{l+1}\max_{t_l\leqslant t\leqslant t_{l+1}} \|\Psi(t_l,y^1_l,h_{l+1},\theta;\bar{f}^1) - \Psi(t_l,y^2_l,h_{l+1},\theta;\bar{f}^2)\| \\
&\leqslant \|y^1_l - y^2_l\| + h_{l+1}b_{max}\max_{1\leqslant i\leqslant\nu} \|f(t^i_{l+1},\{y^1\}^i_{l+1},c,v^1(\alpha(t^i_{l+1},\{y^1\}^i_{l+1}))) \\
&\qquad\qquad - f(t^i_{l+1},\{y^2\}^i_{l+1},c,v^2(\alpha(t^i_{l+1},\{y^2\}^i_{l+1})))\|
\end{aligned} \tag{5.35}$$

is obtained. With the Lipschitz continuity of f it follows that

$$\begin{aligned}
&\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta_{l+1}^1(t) - \eta_{l+1}^2(t)\| \\
\leqslant &\|y_l^1 - y_l^2\| \\
&+ h_{l+1} b_{max} L_f \max_{1 \leqslant i \leqslant \nu} \big(\|\{y^1\}_{l+1}^i - \{y^2\}_{l+1}^i\| + \|v^1(\alpha(t_{l+1}^i, \{y^1\}_{l+1}^i)) - v^1(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i))\| \\
&\qquad + \|v^1(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i)) - v^2(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i))\| \big)
\end{aligned} \tag{5.36}$$

and with the Lipschitz continuity of the functions v_i and α,

$$\begin{aligned}
&\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta_{l+1}^1(t) - \eta_{l+1}^2(t)\| \\
\leqslant &\|y_l^1 - y_l^2\| + h_{l+1} b_{max} L_f (1 + L_{v_1} L_\alpha) \max_{1 \leqslant i \leqslant \nu} \big(\|\{y^1\}_{l+1}^i - \{y^2\}_{l+1}^i\| \big) \\
&\qquad + h_{l+1} b_{max} L_f \max_{1 \leqslant i \leqslant \nu} \|v^1(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i)) - v^2(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i))\|
\end{aligned} \tag{5.37}$$

By using the definitions of $\{y^1\}_{l+1}^i$ and $\{y^2\}_{l+1}^i$ as the stage values of the CRK method for the two problems, it can further be shown that

$$\begin{aligned}
&\|\{y^1\}_{l+1}^i - \{y^2\}_{l+1}^i\| \\
\leqslant &\frac{1}{1 - h_{l+1} L_f \alpha_{max} (1 + L_{v_1} L_\alpha)} \\
&\cdot \left(\|y_l^1 - y_l^2\| + h_{l+1} \alpha_{max} L_f \max_{1 \leqslant i \leqslant \nu} \|v^1(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i)) - v^2(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i))\| \right),
\end{aligned} \tag{5.38}$$

where the prefactor is, according to assumption (B), bounded by 2. Hence, insertion into equation (5.37) and using again assumption (B) gives

$$\begin{aligned}
&\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta_{l+1}^1(t) - \eta_{l+1}^2(t)\| \\
\leqslant &(1 + 2 h_{l+1} b_{max} L_f (1 + L_{v_1} L_\alpha)) \|y_l^1 - y_l^2\| \\
&+ 2 h_{l+1} b_{max} L_f \max_{1 \leqslant i \leqslant \nu} \|v^1(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i)) - v^2(\alpha(t_{l+1}^i, \{y^2\}_{l+1}^i))\|.
\end{aligned} \tag{5.39}$$

By defining $A_3 := 2 b_{max} L_f (1 + L_{v_1} L_\alpha)$ and $A_4 := 2 b_{max} L_f$, the proof for equation (5.33b) is completed.

The proof for relation (5.33a) for the discrete method is analogous. Moreover, the proof for the time derivatives of the continuous representation, equation (5.33c), follows from the fact that

$$\frac{d^j}{dt^j} \eta_{l+1}^k(t) = \sum_{i=1}^{\nu} h_{l+1} \frac{d^j}{d\theta^j} b_i(\theta) h_{l+1}^{-j} f(t_{l+1}^i, \{y^k\}_{l+1}^i, c, v^k(\alpha(t_{l+1}^i, \{y^k\}_{l+1}^i))), \tag{5.40}$$

for $k = 1, 2$ and $1 \leqslant j \leqslant q$, because every differentiation of $b_i(\theta)$ gives, due to the inner derivative of θ with respect of t, a factor of h_{l+1}^{-1}. ■

In short, Lemma 5.13 gives a bound on the difference of the two discrete numerical solutions (equation (5.33a)), a bound on the difference of the two continuous representations on the entire interval (equation (5.33b)), and a bound on the difference of the time derivatives of the two continuous representations (equation (5.33c)). All bounds are thereby given in terms of the different initial values y_l^1, y_l^2, and in terms of the different input functions v^1, v^2.

It is remarked that, due to the interchangeable roles of v_1 and v_2, it is also possible to express v_{diff} by

$$v_{diff} := \max_{1 \leqslant i \leqslant \nu} \|v^1(\alpha(t_{l+1}^i, \{y^1\}_{l+1}^i)) - v^2(\alpha(t_{l+1}^i, \{y^1\}_{l+1}^i)\|, \tag{5.41}$$

where the evaluations take place at the stage values $\{y^1\}_{l+1}^i$ of the CRK method applied to the first IVP (5.27a) rather than at the stage values $\{y^2\}_{l+1}^i$ of the CRK method applied to the second IVP (5.28a).

Lemma 5.13 still holds if several evaluations of the functions v^1 and v^2 enter the right-hand-side function f, i.e. if the right hand side in the equations (5.27) and (5.28) is replaced by $f(t, \mathfrak{u}^k_{l+1}(t), c, \{v^k(\alpha_i(t, \mathfrak{u}^k_{l+1}(t)))\}_{i=1}^{n_\alpha})$ for $k = 1, 2$ respectively. If the right-hand-side function f is Lipschitz continuous with respect to each of the n_α additional arguments, then it is only necessary to add, in the step from equation (5.35) to equation (5.36), a suitable zero $v^1(\alpha_j(t^i_{l+1}, \{y^2\}^i_{l+1})) - v^1(\alpha_j(t^i_{l+1}, \{y^2\}^i_{l+1}))$ for all $j = 1, \dots, n_\alpha$. The rest of the proof then works as before, with the sole modification that a factor of n_τ enters the constants A_j, $1 \leqslant j \leqslant 6$.

In the next section, continuous one-step methods are applied to DDE-IVPs. The theorems therein are stated for the general case of multiple delays, but the proofs are given for the notationally simpler case of a single delay. The extension to the general case of multiple delays can always be done in a way similar to the generalization of Lemma 5.13 for multiple evaluations of the additional input functions v_1 and v_2.

5.2. Continuous One-Step Methods for DDE-IVPs

5.2.1. The Standard Approach for Solving DDE-IVPs

Consider the task of numerically solving DDE-IVPs as in Definition 1.12, i.e.

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c, \{\mathfrak{y}(t - \tau_i(t, \mathfrak{y}(t), c))\}_{i=1}^{n_\tau}) \tag{5.42a}$$

$$\mathfrak{y}(t^{ini})(c) = y^{ini}(c) \tag{5.42b}$$

$$\mathfrak{y}(t) = \phi(t, c) \quad \text{for} \quad t < t^{ini}(c), \tag{5.42c}$$

for arbitrary but fixed parameter values c. Consider a mesh $t^{ini}(c) = t_0 < t_1 < \dots < t_{n_m} = t^{fin}(c)$, and assume that the problem has been solved by a continuous one-step method until the mesh point t_l, which implies that a continuous extension $\eta(t)$ is available for $t \in [t_0, t_l]$ and that discrete values $y_{l'}$ are available at $t_{l'}$ for $l' \leqslant l$. In such a setting, it would be ideal to find the exact solution $u_{l+1}(t)$ of the following local DDE-IVP:

$$\dot{\mathfrak{u}}_{l+1}(t) = f(t, \mathfrak{u}_{l+1}(t), c, \{w_{\eta, u_{l+1}}(t - \tau_i(t, \mathfrak{u}_{l+1}(t), c))\}_{i=1}^{n_\tau}) \tag{5.43a}$$

$$\mathfrak{u}_{l+1}(t_l) = y_l, \tag{5.43b}$$

where the computation of past states is done by evaluating the function

$$w_{\eta, u_{l+1}}(t) = \begin{cases} \phi(t) & \text{for} \quad t < t^{ini}(c) \\ \eta(t) & \text{for} \quad t^{ini}(c) \leqslant t \leqslant t_l \\ u_{l+1}(t) & \text{for} \quad t_l < t \leqslant t_{l+1} \end{cases}. \tag{5.44}$$

The subscripts of the function $w_{\eta, u_{l+1}}$ indicate the use of the function η in the time interval $[t^{ini}(c), t_l]$, and the use of the exact solution u_{l+1} of the local problem (5.43) in the time interval $(t_l, t_{l+1}]$.

This notation is generalized as follows: For any functions $\nu_1(t)$ and $\nu_2(t)$, the function $w_{\nu_1, \nu_2}(t)$ is defined by

$$w_{\nu_1, \nu_2}(t) = \begin{cases} \phi(t) & \text{for} \quad t < t^{ini}(c) \\ \nu_1(t) & \text{for} \quad t^{ini}(c) \leqslant t \leqslant t_l \\ \nu_2(t) & \text{for} \quad t_l < t \leqslant t_{l+1} \end{cases}, \tag{5.45}$$

and formally a right-hand-side function in the standard form of ODEs is obtained by the definition

$$\bar{f}_{\nu_1, \nu_2}(t, y, c) := f(t, y, c, \{w_{\nu_1, \nu_2}(t - \tau_i(t, y, c))\}_{i=1}^{n_\tau}). \tag{5.46}$$

If both $\nu_1(t)$ and $\nu_2(t)$ are known functions – or, for theoretical analyses, assumed to be available – then the right-hand-side function $\bar{f}_{\nu_1, \nu_2}(t, y, c)$ can indeed be regarded as the right-hand-side function of an ODE, and the application of a numerical method can be done in a straightforward way.

For the IVP (5.43), rewritten as

$$\dot{\mathbf{u}}_{l+1}(t) = \bar{f}_{\eta,u_{l+1}}(t, \mathbf{u}_{l+1}(t), c) \tag{5.47a}$$
$$\mathbf{u}_{l+1}(t_l) = y_l, \tag{5.47b}$$

this is not the case because the exact solution $u_{l+1}(t)$ is typically unknown (or otherwise it would not be necessary to use numerical methods for its approximation).

However, if all deviating arguments assume values to the left of t_l for all $t \in [t_l, t_{l+1}]$ then it is not necessary to evaluate $u_{l+1}(t)$ for the computation of the past states. Instead, the past states are obtained by evaluations either of the initial function, or of the already-computed continuous extension $\eta(t)$ in some interval $[t_{l'}, t_{l'+1}]$, $l' < l$, which is locally given by $\eta_{l'+1}(t)$. Since both the initial function and the continuous extension in past integrations steps are, in practice, available, the application of a continuous one-step method for the solution of the ODE-IVP is straightforward. Accordingly, a discrete approximation y_{l+1} and a continuous approximation $\eta_{l+1}(t)$ of $u_{l+1}(t)$ can be computed by

$$y_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; \bar{f}_{\eta,\cdot}) \tag{5.48a}$$
$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}_{\eta,\cdot}). \tag{5.48b}$$

The dot in the second subscript argument of $\bar{f}$ thereby indicates that the function used for computing past states at times $t > t_l$ does not need to be specified, because it is not evaluated. Note that in this case it is possible to compute the discrete step y_{l+1} and the continuous extension $\eta_{l+1}(t)$ independently of each other, if desired.

In the general case that one or several deviating arguments assume values to the right of t_l, an interpretation of the IVP (5.47) as an ODE is not possible. However, the application of the continuous one-step method itself defines a continuous representation $\eta_{l+1}(t)$ on the current interval $[t_l, t_{l+1}]$, which can in practice be used to numerically approximate $u_{l+1}(t)$ for $t > t_l$. Hence, instead of $w_{\eta,u_{l+1}}(t)$, the function $w_{\eta,\eta}(t)$ defined by

$$w_{\eta,\eta}(t) = \begin{cases} \phi(t) & \text{for} \quad t < t^{ini}(c) \\ \eta(t) & \text{for} \quad t^{ini}(c) \leqslant t \leqslant t_{l+1} \end{cases} \tag{5.49}$$

is used for the evaluation of past states. Practically, the equations

$$y_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; \bar{f}_{\eta,\eta}) \tag{5.50a}$$
$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}_{\eta,\eta}). \tag{5.50b}$$

are solved by the continuous one-step method, where the right-hand-side function is defined by

$$\bar{f}_{\eta,\eta}(t, y, c) = f(t, y, c, \{w_{\eta,\eta}(t - \tau_i(t, y, c))\}_{i=1}^{n_\tau}). \tag{5.51}$$

Note that the defining expression (5.50b) for η_{l+1} is *implicit*, because the function η_{l+1} is also needed in the right hand side of the equation. For the same reason, it is not longer possible to apply the discrete one-step method independently from the computation of the continuous representation.

Since the issue whether or not the values assumed by the deviating arguments lie in the current interval has consequences for the construction of numerical methods, the following is defined:

Definition 5.14 (Overlapping)

If it holds for $t \in [t_l, t_{l+1}]$ and for some delay τ_i that $t - \tau_i(t, \eta_{l+1}(t), c) \in [t_l, t_{l+1}]$, then this phenomenon is called overlapping (in the numerical solution)*; likewise, if it holds for $t \in [t_l, t_{l+1}]$ and for some delay τ_i that $t - \tau_i(t, u_{l+1}(t), c) \in [t_l, t_{l+1}]$, then this phenomenon is called* overlapping (in the exact solution of the local problem)*.*

Further, in agreement with Bellen and Zennaro [26], the previously-described method for solving DDE-IVPs, for both the overlapping and the non-overlapping case, is called the *standard approach for solving DDE-IVPs.*

Definition 5.15 (Standard Approach for Solving DDE-IVPs)

Solving DDE-IVPs by computing, in each step $t_l \to t_{l+1}$, the solution of the equations (5.50) as an approximation of the solution of the local IVP (5.47) is called the standard approach for solving DDE-IVPs. *Thereby, $\bar{f}_{\eta,\eta}$ is defined by equation (5.51) and $w_{\eta,\eta}$ is defined by equation (5.49).*

Bellen and Zennaro [26], page 78ff, formulate a theorem on the convergence of the so-defined standard approach to the unique solution $y(t)$ of the DDE-IVP. Unfortunately, the proof of this theorem contains a (very subtle) error, see Remark 5.19 below.

In the following, it is shown that the pitfall in the proof of Bellen and Zennaro [26] can be bypassed by using extrapolations beyond past discontinuities if the corresponding deviating arguments cross such discontinuity points during the integration step $t_l \to t_{l+1}$. However, the use of extrapolations constitutes a new solution approach for DDEs that differs from the standard approach. In the following, the use of extrapolations is therefore first formally defined as *modified standard approach*.

The subsequently presented convergence result is also more general then the one formulated in Bellen and Zennaro [26] in the sense that discontinuous initial functions are allowed. Furthermore, the convergence theorem for the modified standard approach provides the rigorous mathematical basis for the use of extrapolations, which has previously been used in practical DDE codes, e.g. REBUS by Bock and Schlöder [43, 44], RADAR5 by Guglielmi and Hairer [122, 123], and DDEM by ZivariPiran [271].

5.2.2. The Modified Standard Approach for Solving DDE-IVPs

A basic assumption for the new approach is that the exact solution $y(t)$, $t \in \mathcal{T}^f(c)$, has a finite number of discontinuities up to order p, where p is the discrete local order of the employed numerical method. The time points of these discontinuities are denoted by $s_{-n_s^\phi} < s_{-n_s^\phi+1} < \cdots < s_0 < s_1 < \cdots < s_{n_s}$. Thereby, s_i, $-n_s^\phi \leqslant i \leqslant -1$, are the discontinuity points of the initial function for $t < t^{ini}(c)$, and s_i, $1 \leqslant i \leqslant n_s$ represent the discontinuity points at $t > t^{ini}(c)$. Assume further, without loss of generality, that $s_0 = t^{ini}(c)$, i.e. the initial time should be included in the set $\{s_{-n_s^\phi}, \ldots, s_{n_s}\}$ regardless of whether or not the initial function is smoothly linked (up to derivative order p) to $y(t)$ for $t \geqslant t^{ini}(c)$.

The intervals $(-\infty, s_{-n_s^\phi})$, $[s_{-n_s^\phi}, s_{-n_s^\phi+1})$, $\ldots$, $[s_{n_s}, t^{fin}(c)]$ are called *discontinuity intervals* (cf. Definition 4.9 in Chapter 4).

In this setting, for any discontinuity point s_j, $-n_s^\phi \leqslant j \leqslant n_s$, and for all deviating arguments α_i, $1 \leqslant i \leqslant n_\tau$, recall the definition of *propagation switching functions* (Definition 2.10)

$$\sigma_{i,s_j}^{\alpha}(t, y(t), c) = \alpha_i(t, y(t), c) - s_j \tag{5.52}$$

and their signs

$$\zeta_{i,s_j}^{\alpha,+}(t) = \operatorname{sign}^+(\alpha_i(t, y^-(t), c) - s_j). \tag{5.53}$$

Thereby, sign^+ is the simplified version of the sign function as defined by equation (2.14), i.e.

$$\operatorname{sign}^+(x) := \begin{cases} +1 & \text{for} \quad x \geqslant 0 \\ -1 & \text{for} \quad x < 0, \end{cases} \tag{5.54}$$

which attributes the value 1 to the argument 0.

Further, define the corresponding *discontinuity interval indicators* as

$$\xi_i^{\alpha}(t) = (n_s + 1) + \frac{1}{2} \sum_{j=-n_s^\phi}^{n_s} (\zeta_{i,s_j}^{\alpha,+}(t) - 1). \tag{5.55}$$

Note that the terms "discontinuity interval" and "discontinuity interval indicator" are used with a slightly different meaning compared to Chapter 4. The difference is that not only the time points of discontinuity of order 0 in y are taken into account, but instead all time points of discontinuity up to the discrete local order p of the numerical method. The discontinuity interval indicator $\xi_i^{\alpha}(t)$

assumes the following values depending on the value of the deviating argument:

$$\xi_i^\alpha(t) = \begin{cases} -n_s^\phi & \text{for} \quad \alpha_i(t, y^-(t), c) < s_{-n_s^\phi} \\ -n_s^\phi + 1 & \text{for} \quad s_{-n_s^\phi} \leqslant \alpha_i(t, y^-(t), c) < s_{-n_s^\phi+1} \\ \vdots & \\ 0 & \text{for} \quad s_{-1} \leqslant \alpha_i(t, y^-(t), c) < s_0 \\ 1 & \text{for} \quad s_0 \leqslant \alpha_i(t, y^-(t), c) < s_1 \\ \vdots & \\ n_s & \text{for} \quad s_{n_s-1} \leqslant \alpha_i(t, y^-(t), c) < s_{n_s} \\ n_s + 1 & \text{for} \quad s_{n_s} \leqslant \alpha_i(t, y^-(t), c) \end{cases} . \tag{5.56}$$

This means that the value $\xi_i^\alpha(t)$ is given by the index of the discontinuity point that lies to the right of the past time point.

Consider now a continuous one-step method of discrete local order p and uniform local order q, which is applied to solve the DDE-IVP (5.42) by using a mesh $t_0 < t_1 < \cdots < t_{n_m}$. Assume that there is a finite number of discontinuities in the functions $\xi_i^\alpha(t)$ for $t \in \mathcal{T}(c)$, which implies, in particular, that there is a finite number of discontinuities of order $p+1$ in the exact solution $y(t)$. Assume further that the numerical method has access to the discontinuity interval indicators $\xi_i^\alpha(t)$ and that the mesh contains all (finitely many) time points where $\xi_i^\alpha(t)$ is discontinuous. Hence, the indicators $\xi_i^\alpha(t)$, $1 \leqslant i \leqslant n_\tau$, are constant between two mesh points.

Remark that, despite these fairly restrictive assumptions, the right-hand-side function $\bar{f}_{\eta,u_{l+1}}$ in the local IVP (5.47) is not necessarily p-times continuously differentiable, because the mesh is constrained by the requirement to include the discontinuity points of the exact DDE-IVP solution $y(t)$, but not the discontinuity points of the exact solution of the local problem $u_{l+1}(t)$. Therefore it may happen, for example, that at some $t \in [t_l, t_{l+1}]$ the expression $t - \tau_i(t, u_{l+1}(t), c)$ crosses a time point where the initial function ϕ has a discontinuity of order 0. As a consequence, the ODE right-hand-side function $\bar{f}_{\eta,u_{l+1}}$ is discontinuous as well, and the exact solution of the local problem $u_{l+1}(t)$, if it exists, may be just continuous but not differentiable. The same conclusion generally also holds for the practically used right-hand-side function in the standard approach, i.e. for $\bar{f}_{\eta,\eta}$.

For the construction of a local IVP with continuous right-hand-side function, functions different from $w_{\eta,u_{l+1}}$ and $w_{\eta,\eta}$ have to be used for the computation of past states. One approach to achieve this is to use the analogon of the discontinuity locking mechanism known from the numerical solution of HODE-IVPs. Discontinuity locking means that smooth branches of the right-hand-side function f are evaluated in each integration step, see e.g. Hay, Crosbie, and Chaplin [143], Ellison [87], Park and Barton [200], and also the textbooks by Eich-Soellner and Führer [86], page 198, Stoer and Bulirsch [241], page 184, or Hairer, Nørsett, and Wanner [126], page 198. In case one steps over a discontinuity point, a smooth continuation of the solution beyond the discontinuity point is computed first. Subsequently, the discontinuity point can be localized and the step can be repeated in order to include the discontinuity point in the mesh.

In order to transfer this approach to the numerical solution of DDE-IVPs, it is assumed that there exists a representation of the initial function $\phi(t, c)$

$$\phi(t, c) = \begin{cases} \phi_{-n_s^\phi}(t, c) & \text{for} \quad t < s_{-n_s^\phi} \\ \phi_{-n_s^\phi+1}(t, c) & \text{for} \quad s_{-n_s^\phi} \leqslant t < s_{-n_s^\phi+1} \\ \vdots & \\ \phi_0(t, c) & \text{for} \quad s_{-1} \leqslant t < s_0 \end{cases} , \tag{5.57}$$

with functions $\phi_i(t, c)$, $-n_s^\phi \leqslant i \leqslant 0$, that are Lipschitz continuous and p-times continuously differentiable with respect to t on the intervals $(-\infty, s_{-n_s^\phi}]$, $[s_{-n_s^\phi}, s_{-n_s^\phi+1}]$, $\ldots$, $[s_{-1}, t^{fin}(c)]$. These function ϕ_i are called the *smooth branches (of the initial function ϕ)*.

Further, let $J : \{0, \ldots, n_s\} \to \{0, \ldots, n_m\}$ be a function that maps each index k of a discontinuity point $s_k \geqslant t^{ini}(c)$ up to order p to the index l of the mesh point t_l at which it occurs.

With these preparations, reconsider the task of taking the step from t_l to t_{l+1}, under the usual assumption that discrete approximations $y_{l'}$ and continuous extensions $\eta_{l'}(t)$ are available for $l' \leqslant l$.

Let $j'+1$ denote the index of the discontinuity interval in which the current integration step is located, i.e. $[t_l, t_{l+1}] \subset [s_{j'}, s_{j'+1}]$. Let further $\xi_i^\alpha[l+1]$ denote the value of the discontinuity interval indicator for the exact solution in the considered time interval, i.e. $\xi_i^\alpha[l+1]$ represents $\xi_i^\alpha(t')$ for $t' \in (t_l, t_{l+1})$ arbitrary. Furthermore, let $\xi^\alpha[l+1] = (\xi_1^\alpha[l+1], \ldots, \xi_{n_\tau}^\alpha[l+1])^T$. Then, instead of the local IVP (5.47), consider the following alternative:

$$\dot{\mathbf{u}}_{l+1}^1(t) = \bar{f}_{\eta,u_{l+1}^1}^d(t, \mathbf{u}_{l+1}^1(t), c, \xi^\alpha[l+1]) \tag{5.58a}$$

$$\mathbf{u}_{l+1}^1(t_l) = y_l, \tag{5.58b}$$

where

$$\bar{f}_{\eta,u_{l+1}^1}^d(t, y, c, \xi^\alpha) := f(t, y, c, \{z_{\eta,u_{l+1}^1}^{\xi_i^\alpha}(t - \tau_i(t, y, c))\}_{i=1}^{n_\tau}) \tag{5.59}$$

Thereby, $z_{\eta,u_{l+1}^1}^j$, with $-n_s^\phi \leqslant j \leqslant j'+1$, denotes one of the following *deduced functions*:

$$z_{\eta,u_{l+1}^1}^{-n_s^\phi}(t) = \begin{cases} \phi_{-n_s^\phi}(t) & \text{for} \quad t < s_{-n_s^\phi} \\ \phi_{-n_s^\phi}(s_{-n_s^\phi}) + \sum\limits_{i=1}^{p} \frac{1}{i!} \left.\frac{d^i \phi_{-n_s^\phi}(t,c)}{dt^i}\right|_{t=s_{-n_s^\phi}} (t - s_{-n_s^\phi})^i & \text{for} \quad t \geqslant s_{-n_s^\phi} \end{cases} \tag{5.60a}$$

$$z_{\eta,u_{l+1}^1}^{j}(t) = \begin{cases} \phi_j(s_{j-1}) + \sum\limits_{i=1}^{p} \frac{1}{i!} \left.\frac{d^i \phi_j(t,c)}{dt^i}\right|_{t=s_{j-1}} (t - s_{j-1})^i & \text{for} \quad t < s_{j-1} \\ \phi_j(t) & \text{for} \quad s_{j-1} \leqslant t \leqslant s_j \\ \phi_j(s_j) + \sum\limits_{i=1}^{p} \frac{1}{i!} \left.\frac{d^i \phi_j(t,c)}{dt^i}\right|_{t=s_j} (t - s_j)^i & \text{for} \quad t > s_j \end{cases}$$

$$\text{for} \quad -n_s^\phi + 1 \leqslant j \leqslant 0 \tag{5.60b}$$

$$z_{\eta,u_{l+1}^1}^{j}(t) = \begin{cases} \eta_{J(j-1)+1}(t) & \text{for} \quad t < s_{j-1} \\ \eta(t) & \text{for} \quad s_{j-1} \leqslant t \leqslant s_j \\ \eta_{J(j)}(t) & \text{for} \quad t > s_j \end{cases} \quad \text{for} \quad 1 \leqslant j \leqslant j' \tag{5.60c}$$

$$z_{\eta,u_{l+1}^1}^{j'+1}(t) = \begin{cases} \eta_{J(j')+1}(t) & \text{for} \quad t < s_{j'} \\ \eta(t) & \text{for} \quad s_{j'} \leqslant t \leqslant t_l \\ u_{l+1}^1(t) & \text{for} \quad t > t_l \end{cases} \tag{5.60d}$$

The meaning of the term "deduced functions" is, compared to Chapter 4, altered in such a way that the deduced functions now exhibit a higher degree of smoothness when extrapolation beyond the discontinuity points s_i is used. For example, the functions $z_{\eta,u_{l+1}^1}^j$, $-n_s^\phi \leqslant j \leqslant 0$, are p-times continuously differentiable extensions of the functions ϕ_j, $-n_s^\phi \leqslant j \leqslant 0$, which themselves are p-times continuously differentiable on their time domains of definition.

Similarly, the functions $z_{\eta,u_{l+1}^1}^j$, $1 \leqslant j \leqslant j'$, are defined in such a way that they use, for $t < s_{j-1}$ and $t > s_j$, smooth extrapolations of the polynomial continuous extensions in the first mesh interval $[t_{J(j-1)}, t_{J(j-1)+1}]$ and in the last mesh interval $[t_{J(j)-1}, t_{J(j)}]$ that are comprised within the discontinuity interval $[s_{j-1}, s_j]$. Within the discontinuity interval $[s_{j-1}, s_j]$, the continuous extension $\eta(t)$ is used. It is recalled at this point that the continuous extension $\eta(t)$ is a piecewise polynomial function that is only continuous at the mesh points. This issue is uncritical for the convergence proof (proof of Theorem 5.18), and also for the orders of the local errors of practical integration methods, see Theorem 5.21.

Finally, for $z_{\eta,u_{l+1}^1}^{j'+1}$, the deduced function again uses a smooth extrapolation to the left of $s_{j'}$ and the continuous extension $\eta(t)$ for all times in $[s_{j'}, t_l]$. For $t > t_l$, the exact solution $u_{l+1}^1(t)$ of the local problem (5.58) is used for the representation of past states.

The fact that deduced functions are used is notationally reflected by the superscript d in $\bar{f}_{\eta,u_{l+1}^1}^d$.

The deduced functions $z_{\eta,u_{l+1}^1}^j(t)$ in equation (5.60) are piecewise identical to the functions $w_{\eta,u_{l+1}}(t)$ defined by equation (5.44). For example, for $j = -n_s^\phi$ and $t < s_{-n_s^\phi}$, the evaluations of both functions are given by evaluations of the initial function.

In practice, the exact solution $u_{l+1}^1(t)$ of the local problem (5.58) is typically unknown. Therefore,

in a practical numerical method the past states are, for $t > t_l$, computed from the continuous extension $\eta(t)$ that is induced by the numerical method itself. This leads to discrete and continuous numerical approximations that are determined by

$$y_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; \bar{f}^d_{\eta,\eta}) \tag{5.61a}$$

$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}^d_{\eta,\eta}), \tag{5.61b}$$

where

$$\bar{f}^d_{\eta,\eta}(t, y, c, \xi^\alpha) = f(t, y, c, \{z^{\xi^\alpha_i}_{\eta,\eta}(t - \tau_i(t, y, c))\}^{n_\tau}_{i=1}). \tag{5.62}$$

Thereby, the deduced functions $z^j_{\eta,\eta}$, with $-n^\phi_s \leqslant j \leqslant j' + 1$, are given by

$$z^j_{\eta,\eta}(t) = z^j_{\eta,u^1_{l+1}}(t) \quad \text{for} \quad -n^\phi_s \leqslant j \leqslant j' \tag{5.63a}$$

$$z^{j'+1}_{\eta,\eta}(t) = \begin{cases} \eta_{J(j')+1}(t) & \text{for} \quad t < s_{j'} \\ \eta(t) & \text{for} \quad s_{j'} \leqslant t \leqslant t_l \\ \eta_{l+1}(t) & \text{for} \quad t_l < t. \end{cases} \tag{5.63b}$$

Hence, $z^j_{\eta,\eta}(t) \neq z^j_{\eta,u^1_{l+1}}(t)$ only if $j = j' + 1$ and $t > t_l$, i.e. only if the corresponding past state should be obtained from the current discontinuity interval and if the past time point given by the deviating argument is located within the current integration interval $[t_l, t_{l+1}]$. As indicated by the subscripts, the function $z^j_{\eta,u^1_{l+1}}$ then uses the exact solution u^1_{l+1} of the local problem (5.47), whereas $z^j_{\eta,\eta}$ uses the continuous extension η of the numerical method.

Formally, the approach of using deduced functions for the computation of past states in the aforementioned way is part of the formal definition of the *idealized variant of the modified standard approach for solving DDE-IVPs.*

Definition 5.16 (Idealized Variant of the Modified Standard Approach for Solving DDE-IVPs)

Assume that there is a unique "exact" solution $y(t)$ of a given DDE-IVP on the interval $\mathcal{T}^f(c) = (-\infty, t^{fin}(c)]$ and assume that this solution has finitely many discontinuity points s_i, $-n^\phi_s \leqslant i \leqslant n_s$, of order up to the discrete local order p of a given numerical method. Assume further that the discontinuity interval indicators of the exact solution, denoted by $\xi^\alpha(t)$, are piecewise constant with only finitely many discontinuities (in each of its n_τ components), and that both s_i and $\xi^\alpha(t)$ are known to the numerical method.

Then, the idealized variant of the modified standard approach is defined as follows: Select a mesh $t^{ini}(c) = t_0 < t_1 < \cdots < t_{n_m} = t^{fin}(c)$ in such a way that it contains all finitely many discontinuity points of $\xi^\alpha(t)$, and compute, in each step $t_l \to t_{l+1}$, the solution of the equations (5.61) as an approximation of the solution of the local IVP (5.58). Thereby, $\bar{f}^d_{\eta,\eta}$ is defined by equation (5.62) (with $\xi^\alpha = \xi^\alpha[l+1] = \xi^\alpha(t')$ for $t' \in (t_l, t_{l+1})$) and the deduced functions $z^j_{\eta,\eta}$ are defined by equation (5.63).

In the modified standard approach, the computation of past states depends primarily on the discontinuity interval indicators $\xi^\alpha[l+1]$ of the exact solution in the current integration interval $[t_l, t_{l+1}]$, and only secondarily on the past time points given by the deviating arguments. In contrast, in the original standard approach, the computation of past states is solely determined by the past time points given by the deviating arguments, whereas the indicators $\xi^\alpha[l+1]$ play no role in the construction of the method.

Defintion 5.16 is called the "idealized variant" of the modified standard approach, because it relies on the assumption that the numerical method has access to the discontinuity points s_i of the exact solution $y(t)$, and further that it has access to the corresponding discontinuity interval indicators $\xi^\alpha(t)$ of the exact solution. Both these assumptions are, for the case of state-dependent delays, unrealistic. However, with some modifications, the idea of using extrapolations can also be implemented in a practical code. This is the case for the *practical variant of the modified standard approach*, which is defined later in this chapter, see Section 5.4.

5.2.3. Continuous Runge-Kutta Methods for DDE-IVPs

Irrespective of the fact that the idealized variant of the modified standard approach is, for state-dependent delays, only a theoretical construct, two main results are established for the special case of CRK methods. These two results are, on the one hand, the existence of a unique numerical solution, and on the other hand the convergence of the modified standard approach to the exact solution $y(t)$ of the DDE-IVP. The restriction to CRK methods is thereby necessary because at some point in the proof of the convergence theorem Lemma 5.13 on the limited error propagation is exploited.

Theorem 5.17 (Existence and Uniqueness of the Numerical Solution in the Modified Standard Approach)

Consider a DDE-IVP (5.42) for some arbitrary but fixed parameter values c, and assume that the following condition holds:

(L) Lipschitz Continuity (of the model functions)*: The right-hand-side function $f(t, y, c, \{v_i\}_{i=1}^{n_\tau})$ is Lipschitz continuous with respect to y and $\{v_i\}_{i=1}^{n_\tau}$ with Lipschitz constant L_f, i.e.*

$$\|f(t, y_1, c, \{(v_1)_i\}_{i=1}^{n_\tau}) - f(t, y_2, c, \{(v_2)_i\}_{i=1}^{n_\tau})\| = L_f \left(\|y_1 - y_2\| + \sum_{i=1}^{n_\tau} \|(v_1)_i - (v_2)_i\| \right). \quad (5.64)$$

The delay functions $\tau_i(t, y, c)$, $1 \leqslant i \leqslant n_\tau$, are Lipschitz continuous with respect to y with Lipschitz constant L_τ. Further, there exists a representation (5.57) of the initial function ϕ such that the functions ϕ_i are Lipschitz continuous with respect to t on the intervals $(-\infty, s_1]$, $[s_i, s_{i+1}]$, for $1 \leqslant i \leqslant n_s^\phi - 1$, and $[s_{n_s^\phi}, t^{fin}(c)]$, respectively.

Then, for sufficiently small stepsize h_{l+1}, there exists a polynomial solution η_{l+1} of equation (5.61b), and the solution is unique in the space of all polynomials of the same degree.

The theorem, and its proof below, are based on results of Bellen and Zennaro [26], pages 79f, but transferred to the special case of CRK methods, which allows some simplifications.

Proof

According to the discussion of the proof of Lemma 5.13, it is sufficient to give the proof for the case of a single delay τ_1. Accordingly, there is only one discontinuity interval indicator, and its value $\xi_1^\alpha[l+1]$ in the step $t_l \to t_{l+1}$ is, for simplicity of notation, in this proof denoted by ξ.

Consider the continuous operator F that is defined by the right hand side of the equation (5.61b) such that it transforms a polynomial function $\mu(t)$ to $F(\mu)(t)$, $t \in [t_l, t_{l+1}]$:

$$F(\mu)(t) = y_l + h_{l+1} \Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}_{\eta,\mu}^d). \quad (5.65)$$

Further, let θ be defined as

$$\theta = \frac{t - t_l}{h_{l+1}}. \quad (5.66)$$

The goal is to find a fixed point $\mu^\star$ of the continuous operator F, because if $F(\mu^\star)(t) = \mu^\star(t)$, then $\mu^\star(t)$ is apparently a numerical solution of equation (5.61b).

Consider, for this purpose, two arbitrary polynomial functions $\mu_1(t)$, $\mu_2(t)$, $t \in [t_l, t_{l+1}]$, with the sole restriction that $\mu_1(t_l) = \mu_2(t_l) = y_l$ such that they obey the continuity condition at the mesh point t_l. By definition, the application of the operator F yields

$$F(\mu_1)(t) = y_l + h_{l+1} \sum_{i=1}^{\nu} b_i(\theta) f(t_{l+1}^i, \{y^1\}_{l+1}^i, c, z_{\eta,\mu_1}^\xi(t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^1\}_{l+1}^i, c))) \quad (5.67a)$$

$$F(\mu_2)(t) = y_l + h_{l+1} \sum_{i=1}^{\nu} b_i(\theta) f(t_{l+1}^i, \{y^2\}_{l+1}^i, c, z_{\eta,\mu_2}^\xi(t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c))) \quad (5.67b)$$

with

$$\{y^1\}^i_{l+1} = y_l + h_{l+1} \sum_{j=1}^{\nu} a_{i,j} f(t^j_{l+1}, \{y^1\}^j_{l+1}, c, z^{\xi}_{\eta,\mu_1}(t^j_{l+1} - \tau_1(t^j_{l+1}, \{y^1\}^j_{l+1}, c))) \tag{5.68a}$$

$$\{y^2\}^i_{l+1} = y_l + h_{l+1} \sum_{j=1}^{\nu} a_{i,j} f(t^j_{l+1}, \{y^2\}^j_{l+1}, c, z^{\xi}_{\eta,\mu_2}(t^j_{l+1} - \tau_1(t^j_{l+1}, \{y^2\}^j_{l+1}, c))). \tag{5.68b}$$

Please note that curly braces have been used in order to visually distinguish the inner index (which corresponds to the index of the employed polyomial function) from the outer indices (which correspond to the indices of the integration step and to the index of the stage of the CRK method).

Investigate the behavior of $\|F(\mu_1)(t) - F(\mu_2)(t)\|$:

$$\begin{aligned}
\|F(\mu_1)(t) - F(\mu_2)(t)\| \leqslant & h_{l+1} b_{max} \max_{1\leqslant i\leqslant \nu} \|f(t^i_{l+1}, \{y^1\}^i_{l+1}, c, z^{\xi}_{\eta,\mu_1}(t^i_{l+1} - \tau_1(t^i_{l+1}, \{y^1\}^i_{l+1}, c))) \\
& - f(t^i_{l+1}, \{y^1\}^i_{l+1}, c, z^{\xi}_{\eta,\mu_1}(t^i_{l+1} - \tau_1(t^i_{l+1}, \{y^2\}^i_{l+1}, c))) \\
& + f(t^i_{l+1}, \{y^1\}^i_{l+1}, c, z^{\xi}_{\eta,\mu_1}(t^i_{l+1} - \tau_1(t^i_{l+1}, \{y^2\}^i_{l+1}, c))) \\
& - f(t^i_{l+1}, \{y^2\}^i_{l+1}, c, z^{\xi}_{\eta,\mu_2}(t^i_{l+1} - \tau_1(t^i_{l+1}, \{y^2\}^i_{l+1}, c)))\| \\
\leqslant & h_{l+1} b_{max} L_f L_z L_{\tau_1} \max_{1\leqslant i\leqslant \nu} \|\{y^1\}^i_{l+1} - \{y^2\}^i_{l+1}\| \\
& + h_{l+1} b_{max} L_f \left(\max_{1\leqslant i\leqslant \nu} \|\{y^1\}^i_{l+1} - \{y^2\}^i_{l+1}\| + \right. \\
& \left. \max_{t_l\leqslant t\leqslant t_{l+1}} \|\mu_1(t) - \mu_2(t)\| \right) \text{ for all } t \in [t_l, t_{l+1}]
\end{aligned} \tag{5.69}$$

where, in the second step, L_f, L_z, and L_{τ_1} denote the Lipschitz constants of the functions f, z^{ξ}_{η,μ_1}, and τ_1, respectively. Further, it was exploited in the last term that $z^{\xi}_{\eta,\mu_1}(t)$ and $z^{\xi}_{\eta,\mu_2}(t)$ differ only if $t^i_{l+1} - \tau_1(t^i_{l+1}, \{y^2\}^i_{l+1}, c) > t_l$ and if $\xi = j' + 1$, where j' is such that $[t_l, t_{l+1}] \subset [s_{j'}, s_{j'+1}]$.

In a similar manner, it can be shown that

$$\begin{aligned}
\max_{1\leqslant i\leqslant \nu} \|\{y^1\}^i_{l+1} - \{y^2\}^i_{l+1}\| \leqslant & h_{l+1} a_{max} \left(L_f (1 + L_z L_{\tau_1}) \max_{1\leqslant i\leqslant \nu} \|\{y^1\}^i_{l+1} - \{y^2\}^i_{l+1}\| \right. \\
& \left. + L_f \max_{t_l\leqslant t\leqslant t_{l+1}} \|\mu_1(t) - \mu_2(t)\| \right).
\end{aligned} \tag{5.70}$$

with $a_{max} := \max_{1\leqslant i\leqslant \nu} \sum_{j=1}^{\nu} |a_{i,j}|$. Hence, for h_{l+1} sufficiently small, e.g. $h_{l+1} < (2 a_{max} L_f (1 + L_z L_{\tau_1}))^{-1}$, it follows that $h_{l+1} a_{max} L_f (1 + L_z L_{\tau_1}) < 1/2$ and thus

$$\max_{1\leqslant i\leqslant \nu} \|\{y^1\}^i_{l+1} - \{y^2\}^i_{l+1}\| \leqslant 2 h_{l+1} a_{max} L_f \max_{t_l\leqslant t\leqslant t_{l+1}} \|\mu_1(t) - \mu_2(t)\|. \tag{5.71}$$

Insertion into equation (5.69) gives

$$\|F(\mu_1)(t) - F(\mu_2)(t)\| \leqslant 2 h_{l+1} b_{max} L_f \max_{t_l\leqslant t\leqslant t_{l+1}} \|\mu_1(t) - \mu_2(t)\| \text{ for all } t \in [t_l, t_{l+1}]. \tag{5.72}$$

Apparently, for sufficiently small h_{l+1}, the mapping F is a contraction, and therefore a unique fixed point $\mu^\star(t)$ exists in the space of all polynomials of the same degree. The continuous representation of the CRK method in $[t_l, t_{l+1}]$ is thus uniquely determined by $\eta_{l+1}(t) = \mu^\star(t)$.

For the sake of completeness it is mentioned that the deduced functions z^{ξ}_{η,μ_1}, due to the use extrapolations in the modified standard approach, is not globally but only locally Lipschitz continuous in the second step of equation (5.69). However, the function z^{ξ}_{η,μ_1} is evaluated, for $h_{l+1} \to 0$, only in a neighborhood of $t_l - \tau_1(t_l, y_l, c)$ and hence the local Lipschitz continuity of z^{ξ}_{η,μ_1} is sufficient.■

The next theorem established convergence of CRK methods in the framework of the modified standard approach.

Theorem 5.18 (Convergence of CRK Methods for DDE-IVPs)

Consider a DDE-IVP (5.42) for some arbitrary but fixed parameter values c, and a CRK method as in Definition 5.7. Let the following assumptions be fulfilled for $p \geqslant 1$:

(S) Smoothness (of the model functions)*: The right-hand-side function* $f(t, y, c, \{v_i\}_{i=1}^{n_\tau})$ *is* $f(\cdot,\cdot,c,\cdot) \in \mathscr{C}^p(\mathcal{T}(c) \times \mathbb{R}^{n_y} \times \{\mathbb{R}^{n_y}\}^{n_\tau}, \mathbb{R}^{n_y})$ *and Lipschitz continuous with respect to* y *and* $\{v_i\}_{i=1}^{n_\tau}$ *with Lipschitz constant* L_f*, i.e.*

$$\|f(t, y_1, c, \{(v_1)_i\}_{i=1}^{n_\tau}) - f(t, y_2, c, \{(v_2)_i\}_{i=1}^{n_\tau})\| = L_f \left(\|y_1 - y_2\| + \sum_{i=1}^{n_\tau} \|(v_1)_i - (v_2)_i\| \right). \quad (5.73)$$

The delay functions $\tau_i(t, y, c)$, $1 \leqslant i \leqslant n_\tau$*, are such that* $\tau_i(\cdot,\cdot,c) \in \mathscr{C}^p(\mathcal{T}(c) \times \mathbb{R}^{n_y}, \mathbb{R}^{n_y})$*, and they are Lipschitz continuous with respect to* y *with Lipschitz constant* L_τ*. Further, there exists a representation (5.57) of the initial function* ϕ *such that the functions* ϕ_i *are* p*-times continuously differentiable and Lipschitz continuous with respect to* t *on the corresponding intervals.*

(B) Boundedness (of the right-hand-side function)*: The right-hand-side function is bounded by*

$$\|f(t, y, c, \{v_i\}_{i=1}^{n_\tau})\| \leqslant M_f < \infty \quad (5.74)$$

(E) Existence of a unique solution*: There exists a unique solution* $y : \mathcal{T}^f(c) \to \mathbb{R}^{n_y}$ *of the DDE-IVP, which has a finite number of discontinuities up to order* p*. The time points of these discontinuities are denoted by* $s_{-n_s^\phi} < s_{-n_s^\phi+1} < \cdots < s_{n_s}$*. Of these discontinuity points, the time points* $s_{-n_s^\phi}, \ldots, s_{-1}$ *are to the left of* $t^{ini}(c)$ *(i.e. they indicate discontinuities in the initial function), and* $s_0, \ldots, s_{n_s}$ *are equal to or to the right of* $t^{ini}(c)$*.*

(M) Mesh condition*: The set of mesh points* $\{t_0, t_1, \ldots, t_{n_m}\}$ *is such that between two mesh points the discontinuity interval indicators* $\xi_i^\alpha(t)$, $1 \leqslant i \leqslant n_\tau$*, defined by equation (5.55), are all constant. In particular, this implies that the mesh contains all time points of discontinuity up to order* $p + 1$*.*

(C) Consistency*: The CRK method is consistent of discrete order* p *and consistent of uniform order* q*.*

Then the CRK method, realized in the framework of the idealized variant of the modified standard approach for solving DDE-IVPs, converges with discrete global order and uniform global order $r = \min(p, q + 1)$*, and the time derivatives of the continuous representation,* $d^j\eta(t)/dt^j$*, converge with uniform global order* $q + 1 - j$*, i.e.*

$$\max_{1 \leqslant l \leqslant n_m} \|y(t_l) - y_l\| = \mathscr{O}(h_{max}^r) \quad (5.75a)$$

$$\max_{t^{ini}(c) \leqslant t \leqslant t^{fin}(c)} \|y(t) - \eta(t)\| = \mathscr{O}(h_{max}^r) \quad (5.75b)$$

$$\max_{t^{ini}(c) \leqslant t \leqslant t^{fin}(c)} \left\| \frac{d^j}{dt^j} y(t) - \frac{d^j}{dt^j} \eta(t) \right\| = \mathscr{O}(h_{max}^{q+1-j}), \quad \text{for} \quad 1 \leqslant j \leqslant q, \quad (5.75c)$$

where $h_{max} = \max_{1 \leqslant l \leqslant n_m} h_l$*.*

Proof

The proof is given for the notationally simpler case of a single delay τ_1. The associated discontinuity interval indicator is $\xi_1^\alpha(t)$, which is constant between two succesive mesh points t_l and t_{l+1}. Therefore, for simplicity of notation, ξ is used throughout the proof as a short notation for $\xi_1^\alpha(t')$ for $t' \in (t_l, t_{l+1})$.

The basic idea is to consider, on the one hand, the local IVP in equation (5.58), i.e.

$$\dot{\mathbf{u}}_{l+1}^1(t) = \bar{f}_{\eta, u_{l+1}^1}^d(t, \mathbf{u}_{l+1}^1(t), c, \xi) \quad (5.76a)$$

$$\mathbf{u}_{l+1}^1(t_l) = y_l, \quad (5.76b)$$

in which the initial value is given by the numerical approximation y_l of $y(t_l)$, and the past states in the DDE are computed by evaluating the deduced function $z_{\eta, u_{l+1}^1}^\xi$ defined in equation (5.60).

On the other hand, the local problem

$$\dot{\mathbf{u}}^2_{l+1}(t) = \bar{f}^d_{y,u^2_{l+1}}(t, \mathbf{u}^2_{l+1}(t), c, \xi) \tag{5.77a}$$

$$\mathbf{u}^2_{l+1}(t_l) = y(t_l) \tag{5.77b}$$

is considered, in which the initial value is given by the value $y(t_l)$ of the exact DDE-IVP solution at the time point t_l, and where

$$\bar{f}^d_{y,u^2_{l+1}}(t, y, c) = f(t, y, c, z^{\xi}_{y,u^2_{l+1}}(t - \tau_1(t, y, c))). \tag{5.78}$$

Therein, $z^j_{y,u^2_{l+1}}$, $-n^{\phi}_s \leqslant j \leqslant j'+1$, is one of the following deduced functions ($j'+1$ being the index of the current discontinuity interval, i.e. $[t_l, t_{l+1}] \in [s_{j'}, s_{j'+1}]$):

$$z^j_{y,u^2_{l+1}}(t) = z^j_{\eta,u^1_{l+1}}(t) \qquad \text{for} \quad -n^{\phi}_s \leqslant j \leqslant 0 \tag{5.79a}$$

$$z^j_{y,u^2_{l+1}}(t) = \begin{cases} y(s_{j-1}) + \sum\limits_{i=1}^{p} \frac{1}{i!} \left.\frac{d^i y^+(t,c)}{dt^i}\right|_{t=s_{j-1}} (t - s_{j-1})^i & \text{for} \quad t < s_{j-1} \\ y(t) & \text{for} \quad s_{j-1} \leqslant t \leqslant s_j \\ y(s_j) + \sum\limits_{i=1}^{p} \frac{1}{i!} \left.\frac{d^i y^-(t,c)}{dt^i}\right|_{t=s_j} (t - s_j)^i & \text{for} \quad t > s_j \end{cases}$$

$$\text{for} \quad 1 \leqslant j \leqslant j' \tag{5.79b}$$

$$z^{j'+1}_{y,u^2_{l+1}}(t) = \begin{cases} y(s_{j'}) + \sum\limits_{i=1}^{p} \frac{1}{i!} \left.\frac{d^i y^+(t,c)}{dt^i}\right|_{t=s_{j'}} (t - s_{j'})^i & \text{for} \quad t < s_{j'} \\ y(t) & \text{for} \quad s_{j'} \leqslant t \leqslant t_l \\ u^2_{l+1}(t) & \text{for} \quad t > t_l \end{cases} \tag{5.79c}$$

For $-n^{\phi}_s \leqslant j \leqslant 0$, the deduced functions $z^j_{y,u^2_{l+1}}$ are identical to the deduced functions $z^j_{\eta,u^1_{l+1}}$, i.e. they are smooth extensions of the functions ϕ_j. For $1 \leqslant j \leqslant j'$, the deduced functions are equal to the exact solution $y(t)$ between two successive discontinuity points s_{j-1}, s_j, and outside of the discontinuity interval a p-th order Taylor expansion is used. The deduced function $z^{j'+1}_{y,u^2_{l+1}}$ makes use of the exact solution $y(t)$ between $s_{j'}$ and t_l, a p-th order Taylor expansion is used to the left of $s_{j'}$, and the exact solution $u^2_{l+1}(t)$ of the local problem (5.77) is used for $t_l < t \leqslant t_{l+1}$.

The exact solution of the second local problem, equation (5.77), is evidently $u^2_{l+1}(t) = y(t)$ for $t \in [t_l, t_{l+1}]$, because its initial value is $y(t_l)$ and the past states are obtained by evaluations of $\phi(t)$ for $t < t^{ini}(c)$ and by evaluations of $y(t)$ for $t \geqslant t^{ini}(c)$.

Consider now, on the one hand, an application of the CRK method, realized in the framework of the idealized variant of the modified standard approach, to problem (5.76), and on the other hand the application of the same CRK method to the IVP (5.77). Assume, for the second local problem, that the exact solution $u^2_{l+1}(t) = y(t)$ is available for the definition of the numerical method, so that the local problem can be considered as an ODE-IVP with additional input argument.

In this setting, the discrete and continuous numerical approximations y^1_{l+1}, y^2_{l+1}, η^1_{l+1} and η^2_{l+1} are given by

$$y^1_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; \bar{f}^d_{\eta,\eta^1_{l+1}}) \tag{5.80a}$$

$$y^2_{l+1} = y(t_l) + h_{l+1}\Phi(t_l, y(t_l), h_{l+1}; \bar{f}^d_{y,u^2_{l+1}}) \tag{5.80b}$$

$$\eta^1_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}^d_{\eta,\eta^1_{l+1}}) \tag{5.80c}$$

$$\eta^2_{l+1}(t_l + \theta h_{l+1}) = y(t_l) + h_{l+1}\Psi(t_l, y(t_l), h_{l+1}, \theta; \bar{f}^d_{y,u^2_{l+1}}). \tag{5.80d}$$

Herein, the evaluation of the function $\bar{f}^d_{\eta,\eta^1_{l+1}}$ is identical to an evaluation of $\bar{f}^d_{\eta,u^1_{l+1}}$, except if $\xi = j'+1$ and if the deviating argument assumes a value such that $t - \tau_1(t, y, c) > t_l$. In this special case, the continuous representation $\eta^1_{l+1}(t)$ implied by the numerical method is used instead of the exact local solution $u^1_{l+1}(t)$.

The discrete and continuous numerical solution y^1_{l+1} and $\eta^1_{l+1}(t)$ of the IVP (5.76) exist and are unique according to Theorem 5.17 for sufficiently small h_{l+1}. For the discrete and continuous

numerical solution y_{l+1}^2 and $\eta_{l+1}^2(t)$, this follows trivially for explicit CRK methods and from Theorem 5.10 for implicit CRK methods, because the problem is an ODE-IVP.

In order to prove the theorem, consider

$$\|y(t_{l+1}) - y_{l+1}^1\| \leqslant \|y(t_{l+1}) - y_{l+1}^2\| + \|y_{l+1}^2 - y_{l+1}^1\| \tag{5.81a}$$

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|y(t) - \eta_{l+1}^1(t)\| \leqslant \max_{t_l \leqslant t \leqslant t_{l+1}} \|y(t) - \eta_{l+1}^2(t)\| + \max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta_{l+1}^2(t) - \eta_{l+1}^1(t)\| \tag{5.81b}$$

$$\begin{aligned}\max_{t_l \leqslant t \leqslant t_{l+1}} \left\| \frac{d^j}{dt^j} \left(y(t) - \eta_{l+1}^1(t)\right)\right\| \leqslant & \max_{t_l \leqslant t \leqslant t_{l+1}} \left\| \frac{d^j}{dt^j} \left(y(t) - \eta_{l+1}^2(t)\right)\right\| \\ & + \max_{t_l \leqslant t \leqslant t_{l+1}} \left\| \frac{d^j}{dt^j} \left(\eta_{l+1}^2(t) - \eta_{l+1}^1(t)\right)\right\|.\end{aligned} \tag{5.81c}$$

In the first term on the right hand side of the equations, a difference between the exact and the numerical solutions of the ODE-IVP (5.77) occurs. Due to condition (M), $t - \tau_1(t, y(t), c)$ remains in a domain where $y(t)$ is smooth. However, y_{l+1}^2 and $\eta_{l+1}^2(t)$ are defined by the equations (5.80b) and (5.80d), which imply that the past states are obtained by evaluations of a deduced function at $t_{l+1}^i - \tau_1(t_{l+1}^i, \eta_{l+1}^2(t_{l+1}^i), c)$, $1 \leqslant i \leqslant \nu$, where t_{l+1}^i are the abscissae of the CRK method. These past time points may lie outside of the discontinuity interval indicated by ξ. The use of deduced functions in the modified standard approach ensures, at this point, that the smoothness assumptions of the numerical method are nevertheless fulfilled. It is thus possible to exploit the property that the CRK method is consistent of discrete local order p and uniform local order q, which allows to conclude that the first terms in the three equations (5.81) are $\mathscr{O}(h_{l+1}^{p+1})$, $\mathscr{O}(h_{l+1}^{q+1})$, and $\mathscr{O}(h_{l+1}^{q+1-j})$, respectively.

Remark 5.19

The proof for the convergence of numerical methods realized in the framework of the standard approach given in Bellen and Zennaro [26] fails at this point. Only the use of extrapolations allows to construct a sufficiently smooth local IVP.

For the second term in the right hand sides of equations (5.81), Lemma 5.13 is used in the setting $y_l^1 = y_l$, $y_l^2 = y(t_l)$, $v_1 = z^{\xi}_{\eta,\eta_{l+1}^1}$ and $v_2 = z^{\xi}_{y,y}$. This gives

$$\|y_{l+1}^2 - y_{l+1}^1(t)\| \leqslant (1 + A_0 h_{l+1})\|y(t_l) - y_l\| + h_{l+1} A_1 z_{diff} \tag{5.82a}$$

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta_{l+1}^2(t) - \eta_{l+1}^1(t)\| \leqslant (1 + A_0^0 h_{l+1})\|y(t_l) - y_l\| + h_{l+1} A_1^0 z_{diff} \tag{5.82b}$$

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \left\| \frac{d^j}{dt^j} \left(\eta_{l+1}^2(t) - \eta_{l+1}^1(t)\right)\right\| \leqslant A_0^j h_{l+1}^{1-j}\|y(t_l) - y_l\| + h_{l+1}^{1-j} A_1^j z_{diff}, \quad \text{for} \quad 1 \leqslant j \leqslant q \tag{5.82c}$$

where

$$z_{diff} := \max_{1 \leqslant i \leqslant \nu} \|z^{\xi}_{\eta,\eta_{l+1}^1}(t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c)) - z^{\xi}_{y,y}(t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c))\|. \tag{5.83}$$

Therein, $\{y^2\}_{l+1}^i$ are the stage values of the CRK method applied to problem (5.77), for which it holds, by the boundedness condition (B) on f, that $\|\{y^2\}_{l+1}^i - y(t_{l+1}^i)\| = \mathscr{O}(h_{l+1})$.

Obviously, for $-n_s^{\phi} \leqslant \xi \leqslant 0$, the expression z_{diff} is zero because the deduced functions $z^{\xi}_{\eta,\eta_{l+1}^1}$ and $z^{\xi}_{y,y}$ are identical for this case.

Remark further that, for h_{l+1} sufficiently small, the past time points $t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c)$ that occur in the equation for z_{diff} are, due to $\|\{y^2\}_{l+1}^i - y(t_{l+1}^i)\| = \mathscr{O}(h_{l+1})$, within the interval $[s_{\xi-1}, s_{\xi}]$ indicated by the discontinuity interval indicator whenever neither t_l nor t_{l+1} is a propagation of either $s_{\xi-1}$ or s_{ξ}. In this situation the use of deduced functions in equation (5.83) is redundant and it can be concluded that

$$z_{diff} \leqslant \max_{t^{ini}(c) \leqslant t \leqslant t_{l+1}} \|y(t) - \eta(t)\|. \tag{5.84}$$

It remains to consider the general case that the past time points given by the deviating arguments are located outside of the discontinuity interval indicated by ξ. According to the arguments given above, this can happen only in one of the following four cases: $t_l - \tau_1(t_l, y(t_l), c) = s_{\xi-1}$,

$t_l - \tau_1(t_l, y(t_l), c) = s_\xi$, $t_{l+1} - \tau_1(t_{l+1}, y(t_{l+1}), c) = s_{\xi-1}$, and $t_{l+1} - \tau_1(t_{l+1}, y(t_{l+1}), c) = s_\xi$. As an example, the second case is considered, but the other cases can be treated analogously.

As an abbreviation, let $t_{past} = t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c)$ for any $1 \leqslant i \leqslant \nu$, and observe that according to the definition of the deduced functions, it follows that

$$\begin{aligned}
&\|z^\xi_{\eta,\eta^1_{l+1}}(t_{past}) - z^\xi_{y,y}(t_{past})\| \\
&= \left\| \eta(s_\xi) + \sum_{i=1}^{q} \frac{1}{i!} \frac{d^i}{dt^i}\eta(t)\Big|_{t=s_\xi} (t_{past} - s_\xi)^i - y(s_\xi) - \sum_{i=1}^{p} \frac{1}{i!} \frac{d^i}{dt^i} y(t)\Big|_{t=s_\xi} (t_{past} - s_\xi)^i \right\| \\
&\leqslant \max_{t^{ini}(c)\leqslant t\leqslant t_l} \|\eta(t) - y(t)\| + \sum_{i=1}^{q} \frac{1}{i!} \max_{t^{ini}(c)\leqslant t\leqslant t_l} \left\| \frac{d^i}{dt^i}\left(\eta(t) - y(t)\right)\right\| |t_{past} - s_\xi|^i \\
&\quad + \sum_{i=q+1}^{p} \frac{1}{i!} \max_{t^{ini}(c)\leqslant t\leqslant t_l} \left\| \frac{d^i}{dt^i} y(t)\right\| |t_{past} - s_\xi|^i. \qquad (5.85)
\end{aligned}$$

Herein, the maximum of the differences of the time derivatives in the second term and the maximum of the time derivatives in the third term has to be interpreted in such a way that it is applied to both the left-sided and right-sided limit at the mesh points.

Since y is a piecewise p-times continuously differentiable function on the bounded interval $[t^{ini}(c), t_l]$, its time derivatives up to order p are also bounded. Therefore, it follows with a suitable choice of constants N_i, $1 \leqslant i \leqslant p$,

$$\begin{aligned}
&\|z^\xi_{\eta,\eta^1_{l+1}}(t_{past}) - z^\xi_{y,y}(t_{past})\| \\
&\leqslant \max_{t^{ini}(c)\leqslant t\leqslant t_l} \|\eta(t) - y(t)\| + \sum_{i=1}^{q} N_i \max_{t^{ini}(c)\leqslant t\leqslant t_l} \left\| \frac{d^i}{dt^i}\left(\eta(t) - y(t)\right)\right\| |t_{past} - s_\xi|^i \\
&\quad + \sum_{i=q+1}^{p} N_i |t_{past} - s_\xi|^i. \qquad (5.86)
\end{aligned}$$

By considering

$$\begin{aligned}
|t_{past} - s_\xi| &= |t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c) - t_l - \tau_1(t_l, y(t_l), c)| \\
&\leqslant (1 + L_{\tau_1})|t_{l+1}^i - t_l| + L_{\tau_1}\|\{y^2\}_{l+1}^i - y(t_l)\| \qquad (5.87)
\end{aligned}$$

and

$$\|\{y^2\}_{l+1}^i - y(t_l)\| = \left\| h_{l+1} \sum_{j=1}^{\nu} a_{i,j} f(t_{l+1}^i, \{y^2\}_{l+1}^i, c, z^\xi_{y,y}(t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c)))\right\| \qquad (5.88)$$

it follows, by the boundedness of f, that $|t_{past} - s_\xi| \leqslant K h_{l+1}$ for some $K < \infty$, and thus

$$z_{diff} \leqslant \max_{t^{ini}(c)\leqslant t\leqslant t_l} \|\eta(t) - y(t)\| + \sum_{i=1}^{q} N_i \max_{t^{ini}(c)\leqslant t\leqslant t_l} \left\| \frac{d^i}{dt^i}\left(\eta(t) - y(t)\right)\right\| h_{l+1}^i + \sum_{i=q+1}^{p} N_i h_{l+1}^i. \qquad (5.89)$$

Define the following quantities:

$$d_l^0 := \max_{0\leqslant l'\leqslant l} \|y(t_{l'}) - y_{l'}\| \qquad (5.90a)$$

$$\bar{d}_l^0 := \max_{t^{ini}(c)\leqslant t\leqslant t_l} \|y(t) - \eta(t)\| \qquad (5.90b)$$

$$\bar{d}_l^i := \max_{t^{ini}(c)\leqslant t\leqslant t_l} \left\| \frac{d^i}{dt^i}\left(\eta(t) - y(t)\right)\right\| \quad \text{for} \quad 1 \leqslant i \leqslant q. \qquad (5.90c)$$

This allows to bound z_{diff} shortly by

$$z_{diff} \leqslant \bar{d}_{l+1}^0 + \sum_{i=1}^{q} N_i \bar{d}_{l+1}^i h_{l+1}^i + \sum_{i=q+1}^{p} N_i h_{l+1}^i. \qquad (5.91)$$

It has thereby been exploited that $\bar{d}_l^i \leqslant \bar{d}_{l+1}^i$ for $0 \leqslant i \leqslant q$.

Further, by taking the maximum for $l \leqslant \bar{l}$ in equations (5.81), and using the relations (5.82) and (5.91), the expressions

$$d_{\bar{l}+1}^0 \leqslant B_0 h^{p+1} + (1 + A_0 h) d_{\bar{l}}^0 + h A_1 \left(\bar{d}_{\bar{l}+1}^0 + \sum_{i=1}^{q} N_i \bar{d}_{\bar{l}+1}^i h^i + \sum_{i=q+1}^{p} N_i h^i \right) \tag{5.92a}$$

$$\bar{d}_{\bar{l}+1}^0 \leqslant B_0^0 h^{q+1} + (1 + A_0^0 h) d_{\bar{l}}^0 + h A_1^0 \left(\bar{d}_{\bar{l}+1}^0 + \sum_{i=1}^{q} N_i \bar{d}_{\bar{l}+1}^i h^i + \sum_{i=q+1}^{p} N_i h^i \right) \tag{5.92b}$$

$$\bar{d}_{\bar{l}+1}^j \leqslant B_0^j h^{q+1-j} + A_0^j h^{1-j} d_{\bar{l}}^0 + h^{1-j} A_1^j \left(\bar{d}_{\bar{l}+1}^0 + \sum_{i=1}^{q} N_i \bar{d}_{\bar{l}+1}^i h^i + \sum_{i=q+1}^{p} N_i h^i \right), \quad \text{for} \quad 1 \leqslant j \leqslant q \tag{5.92c}$$

are obtained, with $h := \max_{1 \leqslant l \leqslant \bar{l}+1} h_l$, and suitable constants B_0, A_0, A_1 and B_0^j, A_0^j, A_1^j, for $0 \leqslant j \leqslant q$. For simplicity of notation, the symbol $\bar{l}$ is in the following replaced by l.

For the remainder of the proof terms of higher order in h are consequently neglected. For example, in equation (5.92b), it holds that for sufficiently small h the last sum, whose terms are proportional to h^i, $i \geqslant q+2$, is smaller than the first term, which is proportional to h^{q+1}. Hence, for suitably chosen constants C_0, C_0^j, $0 \leqslant j \leqslant q$, it can be concluded that

$$d_{l+1}^0 \leqslant C_0 h^{r+1} + (1 + C_0 h) d_l^0 + h C_0 \bar{d}_{l+1}^0 + C_0 \sum_{i=1}^{q} \bar{d}_{l+1}^i h^{i+1} \tag{5.93a}$$

$$\bar{d}_{l+1}^0 \leqslant C_0^0 h^{q+1} + (1 + C_0^0 h) d_l^0 + h C_0^0 \bar{d}_{l+1}^0 + C_0^0 \sum_{i=1}^{q} \bar{d}_{l+1}^i h^{i+1} \tag{5.93b}$$

$$\bar{d}_{l+1}^j \leqslant C_0^j h^{q+1-j} + C_0^j h^{1-j} d_l^0 + h^{1-j} C_0^j \bar{d}_{l+1}^0 + C_0^j \sum_{i=1}^{q} \bar{d}_{l+1}^i h^{i+1-j} \quad \text{for} \quad 1 \leqslant j \leqslant q, \tag{5.93c}$$

with $r = \min(p, q+1)$.

Observe that the last term in all three equations vanishes if it holds that $t_{l+1}^i - \tau_1(t_{l+1}^i, \{y^2\}_{l+1}^i, c) \in [s_\xi, s_{\xi+1}]$, i.e. if the past time points are in the "correct" discontinuity interval in all integration steps. In this simpler case, it follows that

$$\bar{d}_{l+1}^0 \leqslant C_0^0 h^{q+1} + (1 + C_0^0 h) d_l^0 + h C_0^0 \bar{d}_{l+1}^0. \tag{5.94}$$

For sufficiently small h, $1 - hC_0^0 > 0$ and therefore

$$\bar{d}_{l+1}^0 \leqslant \frac{1}{1 - hC_0^0} \left(C_0^0 h^{q+1} + (1 + C_0^0 h) d_l^0 \right). \tag{5.95}$$

Further, there exists C_1^0 and $h^\star$, depending only on C_0^0, such that for all $h \leqslant h^\star$ it holds that $1/(1 - hC_0^0) \leqslant 1 + hC_1^0$. Hence,

$$\bar{d}_{l+1}^0 \leqslant C_0^0 h^{q+1} + C_1^0 C_0^0 h^{q+2} + (1 + C_0^0 h + C_1^0 h + C_0^0 C_1^0 h^2) d_l^0, \tag{5.96}$$

and by neglecting higher order terms,

$$\bar{d}_{l+1}^0 \leqslant C_2^0 h^{q+1} + (1 + C_2^0 h) d_l^0. \tag{5.97}$$

Insertion into equation (5.93a) – with the last term neglected – gives

$$\begin{aligned} d_{l+1}^0 &\leqslant C_0 h^{r+1} + (1 + 2C_0 h + C_0 C_2^0 h^2) d_l^0 + h C_0 C_2^0 h^{q+2} \\ &\leqslant C_1 h^{r+1} + (1 + C_1 h) d_l^0 \end{aligned} \tag{5.98}$$

for sufficiently small h and some constant C_1. By recursion, this gives

$$d_{l+1}^0 \leqslant (1+C_1h)^{l+1} \underbrace{d_0^0}_{=0} + \sum_{i=0}^{l} (1+C_1h)^i (hC_1h^r). \tag{5.99}$$

Since the second term is a finite geometric series, the relation

$$\begin{aligned} d_{l+1}^0 \leqslant & \frac{(1+C_1h)^{l+1}-1}{1+C_1h-1}(hC_1h^r) \\ \leqslant & (1+C_1h)^{l+1}h^r \end{aligned} \tag{5.100}$$

follows. Since $1+C_1h \leqslant \exp(C_1h)$ and $l+1 \leqslant n_m$, i.e. the number of mesh points, the relation

$$d_{l+1}^0 \leqslant \exp(C_1h \cdot n_m)h^r \tag{5.101}$$

follows. In addition, $h = h_{max}$ for $l+1 = n_m$. Then, if the limit of the stepsizes is taken in the setting that the ratio of the maximum stepsize h_{max} to the minimum stepsize $h_{min} = \min_{1 \leqslant l \leqslant n_m} h_l$ is bounded by $M < \infty$, then also $h_{max}n_m$ is bounded by $M \cdot (t^{fin}(c) - t^{ini}(c))$. Hence, the desired result $d_{l+1}^0 = \mathscr{O}(h_{max}^r)$ is obtained. Further, insertion of this result into equation (5.97) yields the same asymptotic behavior for $\bar{d}_{l+1}^0$.

For the case that the last two terms in equations (5.93) do not vanish because the deviating argument assumes values that are outside the discontinuity interval $[s_{\xi-1}, s_\xi]$, consider equation (5.93c) for the case $j = q$:

$$\bar{d}_{l+1}^q \leqslant C_0^q h + C_0^q h^{1-q} d_l^0 + h^{1-q} C_0^q \bar{d}_{l+1}^0 + C_0^q \sum_{i=1}^{q-1} \bar{d}_{l+1}^i h^{i+1-q} + C_0^q \bar{d}_{l+1}^q h \tag{5.102}$$

At this point it is again possible to argue that for sufficiently small h there exists a constant $\tilde{C}_1^q$ such that $1 - C_0^q h > 0$ and that $1/(1-C_0^q h) < 1 + \tilde{C}_1^q h$. However, here it is sufficient to use the even more general approximation that $1/(1-C_0^q h)$ can be bounded by a constant C_1^q such that it follows

$$\bar{d}_{l+1}^q \leqslant C_2^q h + C_2^q h^{1-q} d_l^0 + h^{1-q} C_2^q \bar{d}_{l+1}^0 + C_2^q \sum_{i=1}^{q-1} \bar{d}_{l+1}^i h^{i+1-q} \tag{5.103}$$

for a suitable constant C_2^q. Insertion of this relation in equation (5.93c) for $1 \leqslant j \leqslant q-1$ yields, for a suitable constant C_1^j, the expression

$$\begin{aligned} \bar{d}_{l+1}^j \leqslant & C_1^j h^{q+1-j} + C_1^j h^{1-j} d_l^0 + h^{1-j} C_1^j \bar{d}_{l+1}^0 + C_1^j \sum_{i=1}^{q-1} \bar{d}_{l+1}^i h^{i+1-j} \\ & + C_1^j h^{q+1-j} \left(h + h^{1-q} d_l^0 + h^{1-q} \bar{d}_{l+1}^0 + \sum_{i=1}^{q-1} \bar{d}_{l+1}^i h^{i+1-q} \right). \end{aligned} \tag{5.104}$$

Observe that the four terms in the second row, originating from $\bar{d}_{l+1}^q$, are proportional to h^{q+2-j}, h^{2-j}, h^{2-j}, and h^{i+2-j}, and thus in each case of one order higher than the corresponding terms in the first row.

It is thus sufficient to consider the effect of the first order term $\bar{d}_{l+1}^1$ when it is inserted into the equations (5.93a) and (5.93b). At first, it follows from equation (5.93c) for $\bar{d}_{l+1}^1$, by neglecting the contributions from $\bar{d}_{l+1}^j$ for $j \geqslant 2$, that

$$\bar{d}_{l+1}^1 \leqslant C_0^1 h^q + C_0^1 d_l^0 + C_0^1 \bar{d}_{l+1}^0 + C_0^1 \bar{d}_{l+1}^1 h. \tag{5.105}$$

With the usual arguments, it follows

$$\bar{d}_{l+1}^1 \leqslant C_2^1 h^q + C_2^1 d_l^0 + C_2^1 \bar{d}_{l+1}^0 \tag{5.106}$$

and insertion into equation (5.93b) gives

$$\begin{aligned} \bar{d}_{l+1}^0 &\leqslant C_0^0 h^{q+1} + (1 + C_0^0 h) d_l^0 + h C_0^0 \bar{d}_{l+1}^0 + C_0^0 h^2 (C_2^1 h^q + C_2^1 d_l^0 + C_2^1 \bar{d}_{l+1}^0) \\ &\leqslant C_1^0 h^{q+1} + (1 + C_1^0 h) d_l^0 + h C_1^0 \bar{d}_{l+1}^0 \end{aligned} \tag{5.107}$$

for a suitable constant C_1^0. This equation is structurally equivalent to equation (5.94), from which on the proof can then be continued as before. Hence, $d_{l+1}^0 = \mathscr{O}(h^r)$ and $\bar{d}_{l+1}^0 = \mathscr{O}(h^r)$.

It is now also easy to see that insertion of $d_{l+1}^0 = \mathscr{O}(h^r)$ and $\bar{d}_{l+1}^0 = \mathscr{O}(h^r)$ into equation (5.106) gives $\bar{d}_{l+1}^1 = \mathscr{O}(h^q)$. Further, by recursion, $\bar{d}_{l+1}^j = \mathscr{O}(h^{q+1-j})$. ■

Theorem 5.18 guarantees that the "global error" $\max_{t_0 \leqslant t \leqslant t_{nm}} \|y(t) - \eta(t)\|$ approaches zero proportional to h_{max}^r as the maximum stepsize goes to zero. However, equally important for practical purposes is the behavior of the *discrete local error* and the *uniform local error*, defined as follows:

Definition 5.20 (Discrete Local Error, Uniform Local Error)

Let u_{l+1}^1 be the exact solution of the local problem (5.58), and let, as usual, the discrete and continuous numerical approximations of this solution be defined by

$$y_{l+1}^1 = y_l + h_{l+1} \Phi(t_l, y_l, h_{l+1}; \bar{f}_{\eta, \eta_{l+1}^1}^d) \tag{5.108a}$$

$$\eta_{l+1}^1(t_l + \theta h_{l+1}) = y_l + h_{l+1} \Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}_{\eta, \eta_{l+1}^1}^d) \tag{5.108b}$$

Then the two quantities

$$\delta_{l+1} := \|u_{l+1}^1(t_{l+1}) - y_{l+1}^1\| \tag{5.109a}$$

$$\bar{\delta}_{l+1} := \max_{t_l \leqslant t \leqslant t_{l+1}} \|u_{l+1}^1(t) - \eta_{l+1}^1(t)\| \tag{5.109b}$$

are called the discrete local error *and the* uniform local error*.*

The discrete and uniform local error represent the newly introduced error in the step from t_l to t_{l+1}. Investigation of the behavior of these errors is particularly relevant with regard to the construction of variable-stepsize methods in Section 5.5.

For CRK methods applied to ODEs with smooth right-hand-side functions, the behavior of these local errors is determined by the discrete local order p and by the uniform local order q of the CRK method itself; according to the definition of consistency, it holds that $\delta_{l+1} = \mathscr{O}(h_{l+1}^{p+1})$ and that $\bar{\delta}_{l+1} = \mathscr{O}(h_{l+1}^{q+1})$. However, despite the use of deduced functions, the right-hand-side function $\bar{f}_{\eta, u_{l+1}^1}^d$ in equation (5.58) (and, similarly, $\bar{f}_{\eta, \eta_{l+1}^1}^d$) does not exhibit the necessary smoothness because the continuous extension is only continuous in the mesh points.

At this point it is instructive to recall the proof of Theorem 5.18, which was carried out despite the lacking smoothness of $\bar{f}_{\eta, u_{l+1}^1}^d$. The crucial point was a suitable use of the triangular inequality and the construction of an additional smooth local problem. The same idea can also be used in order to investigate the behavior of the discrete and uniform local error of CRK methods applied to DDE-IVPs by using the modified standard approach.

Theorem 5.21 (Local Errors of CRK Methods for Solving DDE-IVPs with the Modified Standard Approach)

Consider the local problem (5.58) and denote its exact solution as $u_{l+1}^1(t)$. Consider further a CRK method with discrete local order p and uniform local order q, which is applied to numerically solve the local problem with the modified standard approach, such that y_{l+1}^1 and $\eta_{l+1}^1(t_l + \theta h_{l+1})$ are given by equations (5.108).

Assume that the conditions of Theorem 5.18 hold and further that the continuous extension in the steps $t \leqslant t_l$ are computed with uniform global order r:

$$\max_{t \leqslant t_l} \|\eta(t) - y(t)\| = \mathscr{O}(h^r), \tag{5.110}$$

with $h = \max_{1 \leqslant i \leqslant l} h_i$.

Then it holds that

$$\|y^1_{l+1} - u^1_{l+1}(t_{l+1})\| = \mathscr{O}(h_{l+1}^{p'+1}) \tag{5.111a}$$

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta^1_{l+1}(t) - u^1_{l+1}(t)\| = \mathscr{O}(h_{l+1}^{q'+1}) \tag{5.111b}$$

where $p' = \min(p, r)$, $q' = \min(q, r)$. *Specifically, in the light of Theorem 5.18, if the uniform global order is given by* $r = \min(p, q+1)$, *then it follows that* $p' = \min(p, q+1)$ *and* $q' = q$.

Proof
As in the previous proofs of this section, the case of a single delay τ_1 is considered, and ξ is used a short notation for the discontinuity interval indicator in the current step $t_l \to t_{l+1}$, i.e. for $\xi_1^\alpha(t')$ with $t' \in (t_l, t_{l+1})$.

The sufficiently smooth problem that is employed for the proof is as follows:

$$\dot{\mathbf{u}}^2_{l+1}(t) = \bar{f}^d_{y,y}(t, \mathbf{u}^2_{l+1}(t), c, \xi) \tag{5.112a}$$

$$\mathbf{u}^2_{l+1}(t_l) = y_l. \tag{5.112b}$$

with

$$\bar{f}^d_{y,y}(t, y, c, \xi) = f(t, y, c, z^\xi_{y,y}(t - \tau_1(t, y, c))). \tag{5.113}$$

As indicated by the notation $z^\xi_{y,y}$, the exact solution $y(t)$ of the DDE-IVP is used for the computation of past states if the deviating argument assumes a value between the two discontinuity points indicated by the discontinuity interval indicator; otherwise, smooth extensions of $y(t)$ are used, see the proof of Theorem 5.18 (and in particular equation (5.79)) for details.

Denote the exact solution of the local problem (5.112) by $u^2_{l+1}(t)$ and define the discrete and continuous numerical approximations y^2_{l+1} and $\eta^2_{l+1}(t)$ of $u^2_{l+1}(t)$ as follows:

$$y^2_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; \bar{f}^d_{y,y}) \tag{5.114a}$$

$$\eta^2_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}^d_{y,y}). \tag{5.114b}$$

Note that y^2_{l+1} and $\eta^2_{l+1}(t)$ (which are not actually computed in practice) are defined in such a way that they make use of the exact solution $y(t)$ of the DDE-IVP (or a smooth extension thereof) for the evaluation of past states. They are thus considered as solutions of an ODE-IVP.

Consider, for the continuous extension, the following relation:

$$\begin{aligned}\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta^1_{l+1}(t) - u^1_{l+1}(t_{l+1})\| \leqslant & \underbrace{\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta^1_{l+1}(t) - \eta^2_{l+1}(t)\|}_{A_1} + \underbrace{\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta^2_{l+1}(t) - u^2_{l+1}(t_{l+1})\|}_{A_2} \\ & + \underbrace{\max_{t_l \leqslant t \leqslant t_{l+1}} \|u^2_{l+1}(t_{l+1}) - u^1_{l+1}(t_{l+1})\|}_{A_3}.\end{aligned} \tag{5.115}$$

The second term, A_2, is clearly $\mathscr{O}(h_{l+1}^{q+1})$, because the problem (5.112) is sufficiently smooth. For the first term, which represents the difference between two continuous extensions, Lemma 5.13 is used in the setting $y^1_l = y_l$, $y^2_l = y_l$, $v^1 = z^\xi_{\eta,\eta^1_{l+1}}$, $v^2 = z^\xi_{y,y}$, which gives

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta^1_{l+1}(t) - \eta^2_{l+1}(t)\| \leqslant h_{l+1} B z_{diff} \tag{5.116}$$

for some constant $B < \infty$ and with

$$z_{diff} = \max_{1 \leqslant i \leqslant \nu} \|z^\xi_{\eta,\eta^1_{l+1}}(t^i_{l+1} - \tau_1(t^i_{l+1}, \{y^2\}^i_{l+1}, c)) - z^\xi_{y,y}(t^i_{l+1} - \tau_1(t^i_{l+1}, \{y^2\}^i_{l+1}, c))\|. \tag{5.117}$$

With an analogous argumentation as in the proof of Theorem 5.18, by using equation (5.110) and by considering the limit $h_i \to 0$ in a way such that

$$\frac{h_{max}}{h_{min}} = K < \infty, \tag{5.118}$$

it follows from equation (5.89) that $z_{diff} = \mathscr{O}(h_{l+1}^r)$, and hence $A_1 = \mathscr{O}(h_{l+1}^{r+1})$.

For the term A_3, consider the integral representation of the exact solutions:

$$A_3 \leqslant \int_{t_l}^{t_{l+1}} \Big\| f(t, u_{l+1}^2(t), c, z_{y,y}^{\xi}(t - \tau_1(t, u_{l+1}^2(t), c))) - f(t, u_{l+1}^1(t), c, z_{\eta,\eta_{l+1}^1}^{\xi}(t - \tau_1(t, u_{l+1}^1(t), c))) \Big\| \, dt' \tag{5.119}$$

With the usual arguments on the Lipschitz continuity of f, $z_{\eta,\eta_{l+1}^1}^{\xi}$, and τ_1, and the analysis of z_{diff}, it follows that also $A_3 = \mathscr{O}(h_{l+1}^{r+1})$

The proof for the discrete method can be carried out in the same way, except that the second term, i.e. the difference between the exact and the approximate solution of the smooth local problem (5.112), is $\mathscr{O}(h_{l+1}^{p+1})$. ■

For the application of a CRK method to ODE-IVPs, the discrete local error is $\mathscr{O}(h_{l+1}^{p+1})$. In order to obtain the same result for DDE-IVPs, it follows Theorem 5.21 that the discrete and uniform local orders p and q of the method should be such that $q = p - 1$ or $q = p$.

Nevertheless, it should be noted that the discrete local error can be $\mathscr{O}(h_{l+1}^{p+1})$ in some integration steps even if $q \leqslant p - 2$. To see this, follow again the arguments of the proof and observe that $A_1 = A_3 = 0$ if the discontinuity interval indicator ξ is less than or equal to 0. In this case, the evaluations of the functions $z_{\eta,\eta_{l+1}^1}^{\xi}$ and $z_{y,y}^{\xi}$ yield the same result because the same smooth function $\phi_i(t, c)$ (or extension thereof) is used. Hence, also the two local problems (5.76) and (5.112) are identical and the discrete local error is $\mathscr{O}(h_{l+1}^{p+1})$.

In general, this leads to the important observation that the discrete local error may, for a given CRK method and a given DDE-IVP, vary from one step to another if $q \leqslant p - 2$. Therefore, it is highly desirable for the practical construction of methods to use continuous extensions with uniform local order $q = p - 1$ or $q = p$.

5.3. Continuous One-Step Methods for IHDDE-IVPs

The results for the modified standard approach for solving DDE-IVPs presented in the previous section, i.e. Theorems 5.18 and 5.21, can be generalized to the more general case of IHDDE-IVPs by the following modifications. The presence of switching functions makes it necessary to request that the numerical method has access to both the discontinuity interval indicators $\xi^{\alpha}(t)$ as well as to the signs $\zeta(t)$ of the switching functions of the exact solution. The mesh condition then has to be formulated in such a way that between two mesh points both $\xi^{\alpha}(t)$ and $\zeta(t)$ are constant. This ensures that the time points of all propagated discontinuities up to order p are included in the mesh, and further that the time points of all root discontinuities are included in the mesh.

For non-zero impulse functions, it is further necessary to apply the impulses

$$\eta^+(s) = \eta^-(s) + \omega(s, \eta^-(s), c, \{z_{\eta,\eta}^{\xi_i^{\alpha}}(s - \tau_i(s, \eta^-(s), c))\}_{i=1}^{n_\tau}, \zeta(s)) \tag{5.120}$$

in the time point s of a root discontinuity. If Lipschitz continuity of the impulse functions with respect to the current and past state arguments is assumed, then the error $\|\eta^+(s) - y^+(s)\|$ is of the same order (of the maximum stepsize h) as the error $\|\eta^-(s) - y^-(s)\|$.

5.4. Numerical Computation of Discontinuity Points

The idealized variant of the modified standard approach relies on the assumption that the exact solution $y(t)$ has finitely many discontinuities up to order p, and that the time points of these discontinuities and the corresponding discontinuity interval indicators are known to the numerical method. Further, the mesh is expected to be chosen in such a way that it comprises all time points of discontinuity of the discontinuity interval indicator of the exact solution, i.e. all discontinuity points of $\xi^{\alpha}(t)$.

These conditions can easily be fulfilled by a practical method as long as there are no state-dependencies in the delay and switching functions, because then the locations of both root discontinuities and propagated discontinuities can be computed a priori. However, for the case of

state-dependent switching or delay functions, the location of the discontinuities in the exact DDE-IVP solution $y(t)$ and the associated discontinuity interval indicators are usually unknown.

Hence, for state-dependencies in the delay or switching functions, a *practical variant of the modified standard approach* is needed. Similar to Definition 5.16, this practical variant is formulated for general continuous one-step methods in the case of DDE-IVPs, i.e. without switching and impulse functions:

Definition 5.22 (Practical Variant of the Modified Standard Approach for Solving DDE-IVPs)

Consider a continuous one-step method of discrete local order p and uniform local order q, with a discrete increment function Φ and a continuous increment function Ψ.

For the start of the practical variant, assume that the time points of the discontinuities up to order p in the initial function are known, and denote them by $\hat{s}_i$, $-n_s^\phi \leqslant i \leqslant -1$. Further, let $\hat{s}_0 = t^{ini}(c)$ and apply the following algorithm:

1. *Set $l = 0$ and $n_s = 0$.*
2. *Solve the equations*

$$y_{l+1} = y_l + h_{l+1}\Phi(t_l, y_l, h_{l+1}; \bar{f}^d_{\eta,\eta}) \tag{5.121a}$$

$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\Psi(t_l, y_l, h_{l+1}, \theta; \bar{f}^d_{\eta,\eta}), \tag{5.121b}$$

with

$$\bar{f}^d_{\eta,\eta}(t, y, c, \xi^\alpha[l+1]) = f(t, y, c, \{z_{\eta,\eta}^{\xi_i^\alpha[l+1]}(t - \tau_i(t, y, c))\}_{i=1}^{n_\tau}). \tag{5.122}$$

Therein, $\xi^\alpha[l+1] = (\xi_1^\alpha[l+1], \ldots, \xi_{n_\tau}^\alpha[l+1])^T$ is the unique consistent choice of the discontinuity interval indicator in the interval (t_l, t_{l+1}) for the numerical solution, i.e. it holds

$$\xi_i^\alpha[l+1] \equiv \hat{\xi}_i^\alpha(t) := n_s + 1 + \sum_{j=-n_s^\phi}^{n_s} (\hat{\zeta}_{i,\hat{s}_j}^{\alpha,+}(t) - 1) \tag{5.123a}$$

$$\hat{\zeta}_{i,\hat{s}_j}^{\alpha,+}(t) := sign^+(\alpha_i(t, \eta_{l+1}(t), c) - \hat{s}_j), \tag{5.123b}$$

for $t \in (t_l, t_{l+1})$, $1 \leqslant i \leqslant n_\tau$. If no $h_{l+1} > 0$ can be found such that there is a unique consistent choice for the discontinuity interval indicator, terminate the integration with an error message.

3. *If t_{l+1} is no point of discontinuity in $\hat{\xi}_i^\alpha(t)$, then set $l = l + 1$ and continue with step 2. Otherwise determine the order of the discontinuity of the solution at t_{l+1}. If the order is less than or equal to p, continue with step 4, otherwise set $l = l + 1$ and continue with step 2.*
4. *Set $\hat{s}_{n_s+1} = t_{l+1}$, $n_s = n_s + 1$, $l = l + 1$, and continue with step 2.*

The practical variant of the modified standard approach makes, in contrast to the idealized variant, no assumptions on the exact solution $y(t)$. Instead, it is only requested that the discontinuity interval indicator for the numerical solution, $\hat{\xi}_i^\alpha(t)$, is constant between two mesh points. As a consequence, the application of the practical variant yields a sequence of time points $\hat{s}_k$, $k \geqslant 0$, such that

$$\sigma_{i,\hat{s}_j}^\alpha(\hat{s}_k, \eta(\hat{s}_k), c) = \alpha_i(\hat{s}_k, \eta(\hat{s}_k), c) - \hat{s}_j = 0 \tag{5.124}$$

for at least one combination $(i, j) \in \{1, \ldots, n_\tau\} \times \{-n_s^\phi, \ldots, k-1\}$. The set of so-determined points $\hat{s}_k$ is suspected to include approximations of all time points of discontinuity in the exact solution $y(t)$ up to order $p + 1$.

In practice, it is of course not possible to select the sequence of stepsizes $h_{l+1} = t_{l+1} - t_l$ a priori in such a way that the time points $\hat{s}_k$ of propagated discontinuities are included as mesh points. Instead, it is typical to continue the integration from t_l by taking, at first, a trial step with stepsize h'. For this trial step, all evaluations of f are done with some given value of the

discontinuity interval indicator. For the practical realization of the trial step in the special case of CRK methods, see Section 6.4.

After having computed the trial step, it has to be checked whether the discontinuity interval indicator matches the assumed (fixed) value over the whole integration step, and whether this value is the unique consistent choice. Demanding this property ensures that the practical variant of the modified standard approach yields, for sufficiently small stepsizes, in each integration step the unique numerical solution. However, ensuring existence and uniqueness of the discontinuity interval indicator is, in general, non-trivial, see question Q4 below and the corresponding answer further below.

In order to set the context, it is appropriate to mention that the practical variant of the modified standard approach is inherently coupled to the use of so-called *discontinuity tracking*, see Willé and Baker [256]. Tracking of discontinuities means explicitly locating the numerical discontinuities by root finding techniques and including them into the mesh.

The practical variant of the modified standard approach differs from the idealized variant (Definition 5.16) insofar as it makes use of the numerically determined discontinuity points $\hat{s}_i$ instead of the exact discontinuity points s_i, and, accordingly of the discontinuity interval indicator $\hat{\xi}^\alpha(t)$ for the numerical solution instead of the discontinuity interval indicator $\xi^\alpha(t)$ for the exact solution. This gives rise to the following questions:

- Q1 How can the order of a discontinuity be determined in practice in order to decide whether or not a newly found discontinuity needs to be propagated further?
- Q2 Is the numerically determined discontinuity point $\hat{s}_k$ an approximation of the discontinuity point s_k in the exact solution such that, for $h_{max} \to 0$, $\hat{s}_k \to s_k$?
- Q3 Is it possible, in the special case of CRK methods, to transfer the convergence result of Theorem 5.18 to the practical variant of the modified standard approach?
- Q4 How can existence and uniqueness of a consistent choice for the numerically determined discontinuity interval indicators be guaranteed?

As a starting point for discussing these issues, assume that there is a unique solution $y(t)$ of the DDE-IVP (called "exact solution") with the property that the zeros of the propagation switching functions

$$\sigma^\alpha_{i,s_j}(t, y(t), c) = \alpha_i(t, y(t), c) - s_j \tag{5.125}$$

are distinct. This means that $t \in \mathcal{T}(c)$ is the zero of at most one function σ^α_{i,s_j}. Further, if s_k is a zero of the function σ^α_{i,s_j} for some specific indices i, j, then it is assumed that

$$\frac{d}{dt}\sigma^\alpha_{i,s_j}(s_k, y(s_k), c) \neq 0. \tag{5.126}$$

In the case that s_j is a time point of discontinuity of order 0 in the initial function, the total time derivative of the solution y at s_k may not exist, and, as a consequence, also $d\sigma^\alpha_{i,s_j}(t, y(t), c)/dt$ may not exist at s_k. In this case, equation (5.126) is assumed independently for both the left-sided and for the right-sided time derivative; more specifically, consistency of the discontinuity interval indicators for the exact solution $y(t)$ then also implies that the signs of the left-sided and right-sided time derivative of the switching function are identical.

In the following, answers to the questions Q1–Q4 are given for both the above-established setting with distinct zeros of propagation switching functions with non-zero time derivatives (equation (5.126)) and for the general situation that these assumptions are not fulfilled. For simplicity of the discussion, the two cases are subsequently referred to as "the special case" and as "the general case".

Determination of the Order of Discontinuities

With regard to the question Q1, it is first mentioned that the orders of the initial discontinuities should in general be clear from the context, as they are part of the problem formulation. They can therefore be assumed to be provided as input from the user to the numerical method.

For the orders of the propagated discontinuities, it is ensured in the special case, by condition (5.126), that the roots of the propagation switching functions have multiplicity one. Then, in the

context of scalar DDEs, it is well-known that the order of the child discontinuity is one higher than the order of the parent discontinuity. In systems of DDEs, higher order smoothing may occur depending on the structure of the system. Willé and Baker [256] describe this issue in detail and develop a general approach to compute the orders of propagated discontinuities in DDE systems accurately. A numerical code based on this approach requires a significant amount of user input in order to exploit the higher order smoothing.

It is, however, easily possible to give a lower bound for the orders of the discontinuity in the following way without the need for additional user input. The numerical method may assume that the orders of all initial discontinuities are 0, and that each propagation increases the order only by 1. By using the so-obtained lower bound, it can be guaranteed that all those discontinuities are correctly propagated whose order is less or equal to the discrete local order of the numerical method

In the general case, zeros of higher multiplicity are allowed for the exact solution $y(t)$. In the numerical solution, such zeros will typically split up or vanish due to the presence of numerical integration errors, i.e. it rarely occurs in practice that both $\sigma_{i,j}^{\alpha}$ and its time derivative are exactly zero. Similarly, zeros of two propagation switching functions – which are allowed to coincide in the general case for the exact solution – will typically not exactly coincide for the numerical solution. In other words, zeros of higher multiplicity and coinciding zeros are numerically ill-defined.

However, in the numerical practice threshold values are used such that all values below the threshold are considered as "practically zero". In this context, coinciding propagated discontinuities or zeros of higher multiplicity may certainly occur.

It is clear that the lower bound, obtained as described above, still holds for zeros of higher multiplicity. Further, for coinciding zeros of several propagation switching functions at a time point s_k, a lower bound for the order of the discontinuity at s_k is obtained by increasing the lowest order of all parent discontinuities by 1.

Convergence of Numerically Determined Discontinuity Points to Exact Discontinuity Points

In the special case, the answer to question Q2 is yes, and it is even possible to determine the order of convergence.

In order to discuss the issue in more detail, let the discontinuity at s_k in the exact solution be the child of the discontinuity at s_j. Due to the fact that the discontinuity points are distinct and because the propagation switching function crosses the time point s_k with non-zero time derivative, it follows that the propagation switching function changes its sign also along the numerical solution, if sufficiently small stepsizes are taken. Accordingly, when the practical variant of the modified standard approach is used, then there exists a mesh point $t_l = \hat{s}_k$, where $\hat{s}_k$ denotes the time point of the child discontinuity of the discontinuity at $\hat{s}_j$.

For this situation, Guglielmi and Hairer [124] show by means of the implicit function theorem that

$$|\hat{s}_k - s_k| \leqslant C \left(\|y_l - y(t_l)\| + |\hat{s}_j - s_j| \right), \tag{5.127}$$

where C is a constant. For initial discontinuities, it holds that $\hat{s}_j = s_j$, and hence the numerically determined point $\hat{s}_k$ of the child discontinuity converges to the exact point s_k of the child discontinuity with an approximation order that is equal to the uniform global order of the numerical method. Further, by recursion over all k, $1 \leqslant k \leqslant n_s$, this property follows for all numerically determined discontinuity points $\hat{s}_k$.

This conclusion cannot be transferred to the general case. The reason is that both coinciding propagated discontinuities as well as zeros of higher multiplicity are numerically ill-defined, as discussed in the answer to question Q1. For illustration, consider the case that the exact solution $y(t)$ exhibits coinciding discontinuities that are children of a discontinuity of order 0. Then, in the numerical integration, any arbitrary small numerical error may lead to a situation such that only one of the propagation switching function becomes zero, which leads to a child discontinuity of order 1. After this discontinuity, the time evolution of the other propagation switching function may be such that it never becomes zero.

Therefore, even for asymptotically small stepsizes, it cannot be guaranteed that each discontinuity point s_j in the exact solution is represented in the numerical solution, and, as a result, the numerical solution may end up far off the exact solution.

Convergence of the Numerical Solution

In the special case, the answer to question Q3 is yes, as is shown in Guglielmi and Hairer [124]. This means that the numerical solution obtained with the practical variant of the modified standard approach converges to the exact solution with uniform global order $r = \min(p, q + 1)$. The proof of this result crucially depends on the fact that, in the special case, the numerically determined discontinuity points converge to the exact discontinuity points with the uniform global order r of the numerical method (recall the answer to question Q2 above).

For this reason, no similar convergence result for the general case is known, and the development of a theory for this case is beyond the scope of this thesis.

Consistency of the Discontinuity Interval Indicators

In the special case, it is clear that, for sufficiently small stepsizes, there is at most one sign change per integration step in any of the propagation switching functions. Since the numerical solution converges to the exact solution, it is sufficient to check after a trial step whether the sign of any of the propagation switching functions has changed.

If a propagation switching function has changed its sign, an iterative procedure can be executed that adapts the stepsizes in such a way that equation (5.124) is eventually fulfilled for $t_{l+1} = \hat{s}_k$ within the range of a small numerical tolerance. After that, the integration is continued with the sign of the corresponding propagation switching function negated. If the parent discontinuity has order 0, then it is additionally necessary to verify that the right-sided limit of the total time derivative of the propagation switching function is such that the new choice of the indicator is consistent.

Section 6.9 contains a practical algorithm (namely Algorithm 6.21) that includes discontinuity points into the mesh. The algorithm is realized in the new software package Colsol-DDE. The software checks, however, not rigorously that the assumptions of the special case are fulfilled, because coinciding discontinuities are quite frequent in practice and not generally harmful. Therefore, only coinciding discontinuities of order 1 (or 0, in the case of IVPs with impulses) are avoided in Colsol-DDE, and a set of numerical checks is employed to test whether the propagation switching functions have a "regular" behavior ("sufficiently" non-zero time derivative, leaving the zero set of the switching function into the "correct" direction) in the vicinity of the discontinuity points. These numerical checks, which are also motivated by the differentiability theory developed in Chapter 7, are described in more detail in Section 6.9 and in Subsection 9.1.11. The checks were found to be suitable to detect most irregularities in practical situations.

However, the development of an algorithm that rigorously ensures existence and uniqueness of the discontinuity interval indicator in each integration step also in the general case is an interesting topic for future research.

5.5. Error Control and Adaptive Stepsizes

5.5.1. Global Error Control vs. Local Error Control

In Subsection 5.2.3, and in particular in Theorem 5.18, the convergence of the modified standard approach was investigated in the limit that the stepsizes go to zero. In that context, the stepsizes were allowed to be different in every integration step, but they were treated as if they were some a priori sequence of numbers. The only conditions were that the exact discontinuity points should be included in the mesh, and that the ratio of the maximum stepsize to the minimum stepsize is bounded.

This section is concerned with practical methods for the selection of stepsizes. Knowledge of the asymptotic behavior of the global error $\max_{t_0 \leqslant t \leqslant t_{n_m}} \|y(t) - \eta(t)\|$ as function of the maximum stepsize h_{max} is thereby of only limited practical relevance. Instead, it is desirable to find a sequence of stepsizes such that the global error is bounded by some user-defined tolerance σ_{tol}, and to reach such a sufficiently accurate solution at low computational costs.

Clearly, in order to control and bound the global error, it is first necessary to estimate it. Unfortunately, already the estimation is a computationally hard problem, and numerical methods that guarantee the error to be bounded by σ_{tol} are rarely available, even in the case of ODEs; see Beigel [23] for some recent developments and an overview on existing methods. In the DDE case,

global error estimation is still an unexplored field of research, and none of the popular practical codes known to the author provide an estimate of the global error or control it.

In the following, the presentation is therefore restricted to the simpler and computationally cheaper task of estimating and controlling the discrete and uniform local errors δ_{l+1} and $\bar{\delta}_{l+1}$ (see Definition 5.20), i.e.

$$\delta_{l+1} = \|u_{l+1}(t_{l+1}) - y_{l+1}\| \tag{5.128a}$$

$$\bar{\delta}_{l+1} = \max_{t_l \leqslant t \leqslant t_{l+1}} \|u_{l+1}(t) - \eta_{l+1}(t)\|. \tag{5.128b}$$

The superscript 1, used earlier in Definition 5.20 in order to distinguish u_{l+1}, y_{l+1}, and η_{l+1} from the exact and numerical solutions of a different local problem, is from now on dropped.

The following theorem establishes, under suitable conditions on the local errors, a proportionality of the bound on the global error to the user-defined local tolerance σ_{tol}.

Theorem 5.23 (Proportionality of the Bound of the Global Error to the Tolerance)

Consider the DDE-IVP (5.42) and a continuous one-step method used for its numerical solution realized in the framework of the idealized variant of the modified standard approach. Assume that the conditions of Theorem 5.18 are fulfilled and that in each step the local errors are controlled by

$$\delta_{l+1} \leqslant h_{l+1}\sigma_{tol} \tag{5.129a}$$

$$\bar{\delta}_{l+1} \leqslant \sigma_{tol} \tag{5.129b}$$

for some user-defined tolerance σ_{tol}. Then it holds for the global error that

$$\max_{t_0 \leqslant t \leqslant t_{n_m}} \|y(t) - \eta(t)\| \leqslant K\sigma_{tol} \tag{5.130}$$

for some constant $K < \infty$.

Proof
The proof is given in Bellen and Zennaro, [26], page 188ff, for the case of time-dependent delays and in the context of the standard approach. The generalization of the result to the state-dependent delay case and to the idealized variant of the modified standard approach is straightforward. ■

5.5.2. Estimation of the Local Error

Clearly, the exact solution $u_{l+1}(t)$ of the local problem and the local errors δ_{l+1}, $\bar{\delta}_{l+1}$ are typically unknown. Therefore, estimates $\hat{\delta}_{l+1}$, $\hat{\bar{\delta}}_{l+1}$ of the discrete and uniform local error are used in practice.

For the construction of the estimates $\hat{\delta}_{l+1}$, $\hat{\bar{\delta}}_{l+1}$ a large variety of methods has been developed. A majority of the methods for estimating the discrete local error $\hat{\delta}_{l+1}$ is thereby based on computing two discrete approximations of $u_{l+1}(t_{l+1})$, denoted by y^1_{l+1} and y^2_{l+1}, whose local errors δ^1_{l+1} and δ^2_{l+1} are $\mathcal{O}(h_{l+1}^{p'_1+1})$ and $\mathcal{O}(h_{l+1}^{p'_2+1})$, $p'_1 \neq p'_2$, respectively. Similarly, methods for estimating the uniform local error $\hat{\bar{\delta}}_{l+1}$ are often based on two continuous approximations of $u_{l+1}(t)$ for $t_l \leqslant t \leqslant t_{l+1}$, called $\eta^1_{l+1}(t)$ and $\eta^2_{l+1}(t)$, whose local errors $\bar{\delta}^1_{l+1}$ and $\bar{\delta}^2_{l+1}$ are $\mathcal{O}(h_{l+1}^{q'_1+1})$ and $\mathcal{O}(h_{l+1}^{q'_2+1})$, $q'_1 \neq q'_2$, respectively. The estimates of the local errors are then defined by

$$\hat{\delta}_{l+1} = \|y^1_{l+1} - y^2_{l+1}\| \tag{5.131a}$$

$$\hat{\bar{\delta}}_{l+1} = \max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta^1_{l+1}(t) - \eta^2_{l+1}(t)\|. \tag{5.131b}$$

Note that, in consistency with the notation used in Theorem 5.21, the symbols p'_i and q'_i denote the orders of the local errors, not the discrete local orders and uniform local orders of the methods. How they relate to each other is discussed below in Subsection 5.5.6, but here it is simply assumed that two methods are combined such that their local errors have different orders.

Assume, without loss of generality, that $p'_1 < p'_2$ and that $q'_1 < q'_2$. It then holds for the local

errors δ^1_{l+1} and $\bar{\delta}^1_{l+1}$ of the lower order method that

$$\delta^1_{l+1} \leqslant \hat{\delta}_{l+1} + \delta^2_{l+1} \tag{5.132a}$$

$$\bar{\delta}^1_{l+1} \leqslant \hat{\bar{\delta}}_{l+1} + \bar{\delta}^2_{l+1}, \tag{5.132b}$$

i.e. the error of the lower order method is bounded by its estimate plus the error of the higher order method. This means that the errors of the error estimates are $\mathscr{O}(h_{l+1}^{p'_2+1})$ and $\mathscr{O}(h_{l+1}^{q'_2+1})$, respectively, and hence of higher order than the errors themselves. The error estimation is therefore called *asymptotically correct* (for $h_{l+1} \to 0$) for the lower order method, whose local errors are of the order p'_1 and q'_1.

5.5.3. Selection of Efficient Stepsizes

This finding, together with Theorem 5.23, suggests to accept y^1_{l+1} as an approximation of $y(t_{l+1})$ only if $\hat{\delta}_{l+1} \leqslant h_{l+1}\sigma_{tol}$ and if $\hat{\bar{\delta}}_{l+1} \leqslant \sigma_{tol}$. Assume that these conditions are fulfilled, then a suitable new stepsize h_{l+2} is proposed as follows. According to the asymptotic behavior of the discrete local errors in the current step, δ^1_{l+1}, $\bar{\delta}^1_{l+1}$, and in the next step, δ^1_{l+2}, $\bar{\delta}^1_{l+2}$, it is justified to write

$$\delta^1_{l+1} = A^1_{l+1} h_{l+1}^{p'_1+1}, \qquad \bar{\delta}^1_{l+1} = B^1_{l+1} h_{l+1}^{q'_1+1} \tag{5.133a}$$

$$\delta^1_{l+2} = A^1_{l+2} h_{l+2}^{p_1\prime+1}, \qquad \bar{\delta}^1_{l+2} = B^1_{l+2} h_{l+2}^{q'_1+1}. \tag{5.133b}$$

The crucial assumption is now that from one step to the next the constants should be approximately the same, $A^1_{l+1} \approx A^1_{l+2}$ and $B^1_{l+1} \approx B^1_{l+2}$. In order to obtain an efficient computational method, the next step should be as large as possible without violating the conditions (5.129a) and (5.129b). This suggests to try to achieve $\delta^1_{l+2} = h_{l+2}\sigma_{tol}$ and $\bar{\delta}^1_{l+2} = \sigma_{tol}$. With the asymptotical correctness of the error estimates, this motivates to use

$$h_{l+2} = h_{l+1} \min\left(\sqrt[p'_1]{\frac{\sigma_{tol} h_{l+1}}{\hat{\delta}_{l+1}}}, \quad \sqrt[q'_1+1]{\frac{\sigma_{tol}}{\hat{\bar{\delta}}_{l+1}}} \right) \tag{5.134}$$

as a stepsize for the next step.

However, since the argumentation involves approximations, it is possible that the so-selected h_{l+2} will lead to a discrete error estimate that is slightly larger than $h_{l+2}\sigma_{tol}$ or to a uniform error estimate that is slightly larger than σ_{tol}. This leads to a rejection of the stepsize in the next integration step and therefore to a loss of efficiency. It is thus common to include a safety factor $\rho_{safe} < 1$:

$$h_{l+2} = \rho_{safe} h_{l+1} \min\left(\sqrt[p'_1]{\frac{\sigma_{tol} h_{l+1}}{\hat{\delta}_{l+1}}}, \quad \sqrt[q'_1+1]{\frac{\sigma_{tol}}{\hat{\bar{\delta}}_{l+1}}} \right). \tag{5.135}$$

In the case that the stepsize h_{l+1} is rejected because at least one of the error estimates violates its tolerance condition, the right hand side of equation (5.135) can also be used for the proposition of a new stepsize h^{new}_{l+1} to repeat the current step.

5.5.4. Local Extrapolation

The underlying assumption in the above considerations is that the stepsize is so small that the leading order in h_{l+1} (and h_{l+2}) is dominating. In this domain, it is clear that the higher order results, i.e. y^2_{l+1} and $\eta^2_{l+1}(t)$, are more accurate. This motivates to use these higher order results rather than the lower order results when accepting a step. On the downside, the error estimates $\hat{\delta}_{l+1}$ and $\hat{\bar{\delta}}_{l+1}$ do not have the property to be asymptotically correct for δ^2_{l+1} and $\bar{\delta}^2_{l+1}$.

Nevertheless, advancing with the higher order result, i.e. setting $y_{l+1} = y^2_{l+1}$, $\eta_{l+1}(t) = \eta^2_{l+1}(t)$, is quite common in practice and referred to as *local extrapolation*. For the further analysis of this approach, it is assumed that $p'_2 = p'_1 + 1$ and $q'_2 = q'_1 + 1$, which is typical for the construction of error estimates. The local errors of the higher order result in the step $t_{l+1} \to t_{l+2}$ can then

asymptotically be expressed by

$$\delta^2_{l+2} = A^2_{l+2} h^{p_1'+2}_{l+2}, \qquad \bar{\delta}^2_{l+2} = B^2_{l+2} h^{q_1'+2}_{l+2} \tag{5.136}$$

for some constants A^2_{l+2}, B^2_{l+2}. If the stepsize h_{l+2} is chosen by the strategies described above, i.e.

$$h_{l+2} = \rho_{safe} \min\left(\sqrt[p_1']{\frac{\sigma_{tol}}{A^1_{l+1}}}, \quad \sqrt[q_1'+1]{\frac{\sigma_{tol}}{B^1_{l+1}}} \right) \tag{5.137}$$

then, if the discrete error is dominating, it follows for δ^2_{l+2} that

$$\delta^2_{l+2} = h_{l+2} \cdot \left(\frac{A^2_{l+2}}{\sqrt[p_1']{(A^1_{l+1})^{p_1'+1}}} \rho^{p_1'+1}_{safe} \right) \sqrt[p_1']{\sigma^{p_1'+1}_{tol}} \tag{5.138}$$

and similarly, if the uniform error is dominating, then it follows for $\bar{\delta}^2_{l+2}$ that

$$\bar{\delta}^2_{l+2} = \left(\frac{B^2_{l+2}}{\sqrt[q_1'+1]{(B^1_{l+1})^{q_1'+2}}} \rho^{q_1'+2}_{safe} \right) \sqrt[q_1'+1]{\sigma^{q_1'+2}_{tol}}. \tag{5.139}$$

Asymptotically, for $h_{l+2} \to 0$, the terms in brackets are bounded by constants C_1 and C_2, so that it can be concluded that δ^2_{l+2} fulfills condition (5.129a) with a tolerance $C_1 \sqrt[p_1']{\sigma^{p_1'+1}_{tol}}$ and that $\bar{\delta}^2_{l+2}$ fulfills condition (5.129b) with a tolerance $C_2 \sqrt[q_1'+1]{\sigma^{q_1'+2}_{tol}}$.

There is a possibility to recover the proportionality to σ_{tol} as follows. Instead of requesting that $\delta^1_{l+2} \leqslant h_{l+2}\sigma_{tol}$ and that $\bar{\delta}^1_{l+2} \leqslant \sigma_{tol}$, consider the conditions $\delta^1_{l+2} \leqslant \sigma_{tol}$ and $\bar{\delta}^1_{l+2} \leqslant \sigma_{tol}/h_{l+1}$. This gives, instead of equations (5.138) and (5.139),

$$\delta^2_{l+2} = h_{l+2} \cdot \left(\frac{A^2_{l+2}}{A^1_{l+1}} \rho^{p_1'+1}_{safe} \right) \sigma_{tol} \tag{5.140a}$$

$$\bar{\delta}^2_{l+2} = \left(\frac{B^2_{l+2}}{B^1_{l+1}} \rho^{q_1'+2}_{safe} \right) \sigma_{tol} \tag{5.140b}$$

and hence proportionality to the chosen tolerance.

5.5.5. Error Criteria

For discrete one-step methods, a control of the local error by $\hat{\delta}^1_{l+1} \leqslant \sigma_{tol}$ is commonly termed an *error per step* criterion, whereas $\hat{\delta}^1_{l+1}/h_{l+1} \leqslant \sigma_{tol}$ (see equation (5.129a)) is known as an *error per unit step* criterion. For the control of the error of the continuous representation, these terms are typically not used.

For all conditions that involve the stepsize, it is unsatisfactory from a practical point of view that the error control depends on the unit of the time variable, i.e. whether the time is measured in seconds, hours, or days. In order to remove this dependency, h_{l+1} can be replaced by $h_{l+1}/(t^{fin}(c) - t^{ini}(c))$. Using this expression in the error per unit step criterion in equation (5.129a) is called a *modified error per unit step* criterion.

In the previous subsections, it was discussed that – without local extrapolation – the conditions (5.129a) and (5.129b) are suitable to achive proportionality of a bound on the global error to the chosen tolerance σ_{tol}. With local extrapolation, the conditions $\delta^1_{l+1} \leqslant \sigma_{tol}$ and $\bar{\delta}^1_{l+2} \leqslant \sigma_{tol}/h_{l+1}$ are suitable to obtain this proportionality.

Practical codes do not always follow these guidelines to obtain a proportionality to σ_{tol}. Instead, an error per step criterion might be used even if the lower order result is used for advancing to the next mesh point, or an error per unit step criterion might be used in connection with a higher order result. In some codes the error of the continuous representation is not controlled at all, arguing that, asymptotically, the condition on the discrete local error $\hat{\delta}_{l+1} \leqslant \sigma_{tol} h_{l+1}$ is always more restrictive than the condition on the uniform local error $\hat{\bar{\delta}}_{l+1} \leqslant \sigma_{tol}$.

5.5.6. Choice of Pairs of Discrete Methods and Pairs of Continuous Representations

The basic idea for error estimation in Subsection 5.5.2 was to compute two discrete approximations and two continuous approximations of the solution of the local problem, whose local errors are $\mathscr{O}(h_{l+1}^{p_1'+1})$, $\mathscr{O}(h_{l+1}^{p_2'+1})$, $\mathscr{O}(h_{l+1}^{q_1'+1})$, and $\mathscr{O}(h_{l+1}^{q_2'+1})$ with $p_1' \neq p_2'$ and $q_1' \neq q_2'$. Unfortunately, in the view of Theorem 5.21, the orders p_i' and q_i' of the local errors are not always simply given by the discrete local order p_i and by the uniform local order q_i of the continuous one-step method, i.e. it does in general not hold that $p_i' = p_i$, $q_i' = q_i$ for $i = 1, 2$. Instead, the order of the local errors may be compromised by the order of the error in the approximation of past states.

For the practical choice of pairs for error estimation, consider two discrete one-step methods with discrete local orders p_1 and p_2 and two continuous representations with uniform local orders q_1 and q_2. More precisely, let p_1 and q_1 be the orders of the *advancing methods*, i.e. the discrete and continuous approximations are given by $y_{l+1} \equiv y_{l+1}^1$ and $\eta_{l+1}(t) \equiv \eta_{l+1}^1(t)$, whereas p_2 and q_2 are the orders of the *error-estimating methods*, i.e. y_{l+1}^2 and $\eta_{l+1}^2(t)$ are only computed for the purpose of error estimation.

From Theorem 5.21 it is known that the order of the discrete local error p_i' and the order of the uniform local error q_i' are affected by the uniform global order r of the advancing method, which is determined by the discrete and uniform local order of the advancing method, i.e. $r = \min(p_1, q_1+1)$. The question is: How should p_1, q_1, p_2, and q_2 be chosen such that it follows $p_i' = p_i$, $q_i' = q_i$ (for $i = 1, 2$) in all integration steps?[1]

At first, it follows from Theorem 5.21 that $q_1' = q_1$. Further, if $q_1 = p_1$ or $q_1 = p_1 - 1$, then it also holds that $p_1' = p_1$. Under the same condition, it follows for the uniform error-estimating method that $q_2' = \min(q_2, p_1) = q_2$, because the uniform order of any continuous extension cannot exceed the discrete order. Unfortunately, for the discrete error-estimating method Theorem 5.21 gives that $p_2' = \min(p_2, p_1)$. Hence, it would not be possible to use discrete method for the error estimation that is of a higher order than the discrete local order p_1 of the advancing method. In this context, consider the following remark.

Remark 5.24

If $q_1 = p_1$, and if $p_2 \leqslant p_1 + 1$, then it holds that $p_2' = p_2$.

Hence, if $q_1 = p_1$, then all methods (discrete and continuous, advancing and error-estimating) perform to their order.

The proof for the result in Remark 5.24 works similar to the proof of Theorem 5.21, however, instead of using $\max_{t \leqslant t_l} \|\eta(t) - y(t)\| = \mathscr{O}(h_{l+1}^{r_1})$, the existence of a function $\bar{\eta}(t)$ is postulated such that $\max_{t \leqslant t_l} \|\eta(t) - \bar{\eta}(t)\| = \mathscr{O}(h_{l+1}^{p_1+1})$. With this (and suitable smooth extensions of $\bar{\eta}(t)$), the remark follows. The more technical part of the proof is to show that the function $\bar{\eta}(t)$ exists. For this purpose, it is referred to Bellen and Zennaro [26], page 184ff.

If the conditions of Theorem 5.21 and the extension in Remark 5.24 are respected, any of the following concepts can be used for the practical combination of methods of different orders:

- The most obvious approach is to simply take two continuous one-step methods of different discrete local orders and different uniform local orders and to combine them in such a way that the orders of the local errors are equal to these local orders. However, an intricate issue that has to be considered is which methods should be chosen so that the computational costs are low. An elegant and widely-used approach for estimation of the discrete local error, specifically for explicit Runge-Kutta methods, are so-called *embedded pairs*. The basic idea of embedded pairs is to combine Runge-Kutta methods which share many of their stage values.

  For the estimation of the uniform local error, the same idea can be used. However, in addition it is highly desirable to combine two continuous representations for the error estimation such that the maximum difference between them is located at a known and problem-independent position in $[t_l, t_{l+1}]$. Contrariwise, if the maximum difference is not a problem-independent position, the maximum value has to be determined in every integration step, which may lead to a loss of efficiency.

[1] It should be noted, however, that these relations are convenient for the construction of error estimates, but not necessary.

- As an alternative for the estimation of the uniform local error, it is possible to take a continuous one-step method with $q_1 < p_1$ as a basis, and then apply a method for constructing a higher order continuous extension, e.g. the boot-strapping process described by Enright [99] or the uniform correction procedure by Zennaro [268, 269] and Bellen and Zennaro [25, 26]. As a result, two continuous representations of different uniform local order are obtained, and their maximum difference can be used as an estimate for the uniform local error. In some special situations, it can be shown that the maximum difference is located at $t_l + \theta^\star h_{l+1}$, where $\theta^\star$ is independent of l and independent of the model functions of the DDE-IVP. In Chapter 6.2, some examples are presented where this is the case, and these special examples are exploited in the construction of the new IHDDE-IVP solver Colsol-DDE.

- As an alternative for the estimation of the discrete local error, a continuous representation of order $q_1 = p_1$ can be integrated by a quadrature rule of sufficiently high order, which gives a discrete approximation of order $p_2 = p_1 + 1$. See Zennaro [269] and Section 6.3 for details.

- For the estimation of the discrete local error, *Richardson extrapolation* is an option. This method is based on a single one-step method, which is applied on the interval $[t_l, t_l + h_{l+1}]$ once with a stepsize h_{l+1}, and then with two steps of length $h_{l+1}/2$. Note that for ODEs with smooth right-hand-side functions, this approach can be used to generate, in principle, approximations of arbitrary high order, whereas for DDEs the maximum discrete local order that can be achieved is limited by the order of the error in the approximation of past states.

- Monitoring the so-called *defect*, in the context of ODE-IVPs defined as $\epsilon(t) := \eta_{l+1}(t) - f(t, \eta_{l+1}(t), c)$, can also serve as a basis for error control, see Enright [94]. More precisely, it can be shown that both the discrete local error and the uniform local error are asymptotically bound by $h_{l+1} K \max_{t_l \leqslant t \leqslant t_{l+1}} \|\epsilon(t)\|$ for some $K < \infty$. Hence, controlling the defect implies also a control of the discrete and uniform local errors. The generalization of this *defect control* strategy to DDE-IVP solvers is described in Enright [95].

6. Colsol-DDE: The COLlocation SOLver for DDEs

> *Die praktische Durchführung erfordert einige programmiertechnische Kunstfertigkeit, wenn auf zugleich einfache und narrensichere Weise alle auftauchenden Möglichkeiten erfaßt und bewältigt werden sollen.*
>
> Bulirsch, in the lecture notes of a presentation at the Carl-Cranz-Gesellschaft [51], commenting on the difficulty to implement a practical solver for initial value problems with implicitly defined discontinuities.

In Chapter 5, the mathematical foundation for the use of the modified standard approach has been established by giving a well-posedness and a convergence result. Further, the proportionality of the bound of the global error with respect to the tolerance in a common local error control strategy has been discussed. These results are the basis for the development of the COLlocation SOLver for Delay Differential Equations (Colsol-DDE), a new software package for solving initial value problems (IVPs) in impulsive hybrid discete-continuous delay differential equations (IHDDEs). This chapter gives the details of the algorithms that are used in Colsol-DDE for this purpose.

Survey of Existing Solvers

The number of available solvers that are able to solve IVPs in ordinary differential equations (ODE-IVPs) is very large. Hence, it is not attempted here to give an overview of the available codes for this comparably simple class of differential equations.

Naturally, any ODE-IVP solver can also be applied naively to an HODE-IVP by simply implementing a discontinuous right-hand-side function. However, since the right-hand-side function then violates the smoothness assumptations on which the ODE-IVP solver is based, solvers with fixed stepsizes may be inaccurate. Further, solvers with variable stepsizes may fail to solve the problem, or they may undergo a large number of rejected steps in the vicinity of the discontinuity points and thus become very inefficient. Therefore, when solving HODE-IVPs numerically, it is highly recommended to use special solvers that are designed for the treatment of such problems. In any case, the use of special solvers is indispensable for treating IHODE-IVPs, because the code must be aware of the fact that an impulse has to be applied to the state vector at time points that are in general only defined implicitly.

The following tailored solvers are available for the numerical solution of HODE-IVPs and IHODE-IVPs: DASPKE by Mao and Petzold [184], RKFSWT by Kirches [160], and to the collection of ODE solvers currently incorporated into MATLAB [186], see also Shampine, Gladwell, and Thompson [232].

It is therefore sufficient to restrict the further discussion to solvers for IVPs in differential equations with time delays. The following overview of computer programs gives details about the problem classes that can be solved as well as on the employed numerical methods.

- Archi by Paul [203] can solve DDE-IVPs with multiple state-dependent delays. In addition, IVPs in so-called "DDEs of neutral type" can be solved, and also a limited class of "integro-differential equations". The solver is based on embedded explicit Runge-Kutta methods developed by Dormand and Prince [80] of order 4 and 5. A fifth-order Hermite interpolant suggested by Shampine [229] is used as continuous representation. Tracking of discontinuities is optional. The implementation language is FORTRAN77.

- dde23 by Shampine and Thompson [233] can solve DDE-IVPs with multiple constant delays. In addition, the code is designed for the solution of HDDE-IVPs and IHDDE-IVPs. The solver relies on a pair of explicit Runge-Kutta formulae of order 2 and 3 by Bogacki and

Shampine [46], together with a cubic Hermite interpolant as continuous representation. The solver tracks discontinuities, and since the solver is designed for constant delays only, all propagated discontinuities can be computed in advance in the context of DDE-IVPs. The solver is implemented in MATLAB.

- ddesd by Shampine [230] can solve DDE-IVPs with multiple state-dependent delays. In addition, it can be used for solving HDDE-IVPs and IHDDE-IVPs. It is based on the "classic" explicit Runge-Kutta method of order 4 together with a cubic Hermite interpolant as continuous representation. Defect control as suggested by Enright [94] is used in the variable-stepsize strategy. The code does not track discontinuities but relies on the control of the defect to include discontinuity points approximately in an automatic way. The implementation is in MATLAB.

- HBO414DDE by Yagoub, Nguyen-Ba, and Vaillancourt [264] can solve DDE-IVPs with multiple state-dependent delays. The solver is based on a variable-step variable-order general linear method for ODE-IVPs as described in Nguyen-Ba et al. [194]. A Hermite interpolation procedure is employed to approximate the solution at non-mesh time points. Only those discontinuities are tracked that cause a rejection of a stepsize. The implementation is in C++.

- DDEM by ZivariPiran [271] and ZivariPiran and Enright [272] is a software package for DDE-IVPs with multiple state-dependent delays. In addition, it is designed for "DDE-IVPs of neutral type". It is worth remarking that DDEM is, strictly speaking, not an IVP solver, but rather a software framework. In principle, DDEM can make use of any underlying ODE-IVP solver that provides a continuous representation. The present implementation is based on a continuous explicit Runge-Kutta method of order 6 as developed in Enright and Yan [100], and defect control is used in the error estimation and stepsize selection strategy. The code tracks discontinuities and is implemented in C/C++.

- DDE-STRIDE by Butcher [54], also described in Baker, Butcher, and Paul [12], solves DDE-IVPs with multiple state-dependent delays. It can also be used to solve IVPs in "DDEs of neutral type". The code is a modified version of STRIDE by Burrage, Butcher, and Chipman [53], which is based on so-called "singly implicit Runge-Kutta methods". These methods are non-standard in the sense that their abscissae are outside of the interval $[0, 1]$. The code adapts both stepsize and order during the integration. The error estimation and control relies on embedded formulae. The continuous representation of the method is based on Laguerre polynomials. DDE-STRIDE does not track the discontinuities. The implementation language is FORTRAN77.

- DDE_SOLVER by Thompson and Shampine [246] can solve DDE-IVPs with multiple state-dependent delays. In addition, it is suitable for HDDE-IVPs and IHDDE-IVPs as well as for IVPs in "DDEs of neutral type". DDE_SOLVER is the successor of DKLAG6, see Corwin, Sarafyan, and Thompson [70]. Both codes are based on explicit, embedded continuous Runge-Kutta methods of orders 5 and 6. DDE_SOLVER tracks discontinuities and is implemented in Fortran90/95.

- DDVERK by Hayashi [144] and Enright and Hayashi [96] is a program for solving DDE-IVPs with multiple state-dependent delays. "DDE-IVPs of neutral type" can also be solved. The code is based on continuous explicit Runge-Kutta methods. Defect control is used in the error estimation and variable-stepsize strategy. The code does not track discontinuities, but instead relies on the techniques developed in Enright et al. [98] for the treatment of ODE-IVPs with discontinuous right-hand-side function. The fact that DDVERK features this strategy provides motivation for an application of the code to HDDE-IVPs. The implementation is in FORTRAN77.

- DELSOL by Willé and Baker [255] solves DDE-IVPs with multiple state-dependent delays and is based on an Adams PECE linear multi-step method. The method adapts stepsize and order during the integration. The code tracks discontinuities and is implemented in FORTRAN77.

- DMRODE by Neves [193] solves DDE-IVPs with state-dependent delays, but with the restriction that each component of the state vector is evaluated at one past time point at

most. It is based on an explicit Runge-Kutta method of order 4 equipped with an Hermite interpolant as continuous representation. DMRODE does not feature an automatic tracking of discontinuities.

- PAI4D by Weiner and Strehmel [253] is a program for solving DDE-IVPs with a single constant delay. The solver relies on a class of methods called "adaptive Runge-Kutta method", or "Rosenbrock-type methods", which can be regarded as a linearization of a diagonally implicit Runge-Kutta method (cf. Strehmel and Weiner [243] and Hairer and Wanner [127], page 102ff). The main focus of the code is to automatically detect and address a possible stiffness of the problem. Lagrange and Hermite interpolants of order 2 and 3 are used as continuous representation, and Richardson extrapolation is employed in the error control strategy. PAI4D is implemented in FORTRAN77.

- RADAR5 by Gugliemi and Hairer [122, 123] can solve DDE-IVPs with multiple state-dependent delays. In addition, it can deal with a very general class of "implicit delay differential equations", "delay-differential-algebraic" equations, and "DDEs of neutral type". The solver has been developed as a variant of the ODE-code RADAU5 (Hairer and Wanner [127], page 568) based on the three-stage Radau IIA collocation method, i.e. an implicit Runge-Kutta method. The collocation polynomial induced by the method is used as continuous representation of the solution. By default, RADAR5 employs an algorithm that includes only "significant" discontinuities into the mesh, however, restrictive discontinuity tracking is optional. The code is written in Fortran90.

- A variant of RKFSWT (Kirches [160]) as described in Ernst [101]. This program solves DDE-IVPs with multiple state-dependent delays. It is designed in such a way that the user can choose from various embedded explicit Runge-Kutta methods. Hermite-Birkhoff interpolation is used, or, if this does not provide a continuous representation of sufficiently high order as compared to the discrete local order of the Runge-Kutta method, the "bootstrapping process" of Enright et al. [99] is employed. The code tracks discontinuities and is implemented in C/C++.

- REBUS by Bock and Schlöder [43] solves DDE-IVPs with multiple state-dependent delays. REBUS is a variable-stepsize, variable-order code based on Adams PECE formulae. The code was, to the knowledge of the author, the first to provide an error-controlled continuous representation and to track discontinuities.

- RETARD by Hairer, Nørsett, and Wanner [127] solves DDE-IVPs with multiple state-dependent delays by using the embedded Runge-Kutta methods of orders 4 and 5 suggested by Dormand and Prince [80] with a continuous approximation of order 4. The code does not track discontinuities and is realized in FORTRAN77.

- RKFHB4 by Oppelstrup [198] solves DDE-IVPs with multiple constant or time-dependent delays. The solver is based on a embedded Runge-Kutta methods of order 4 and 5 and uses a 4-th order Hermite-Birkhoff interpolant as continuous representation. The code does not track discontinuities.

- SNDDELM by Jackiewicz and Lo [154] solves DDE-IVPs with multiple state-dependent delays. In addition, it is designed for solving "DDEs of neutral type". SNDDELM is a realization of a variable-order Adams-Bashforth-Moulton multi-step method. The code does not track discontinuities but instead relies on the error control strategy to include them approximately.

- Solv95 by Wood [259] is a code for solving HDDE-IVPs and IHDDE-IVPs with multiple state-dependent delays. "DDEs of neutral type" can also be solved. The solver is based on explicit Runge-Kutta methods of orders 2 and 3 equipped with a cubic Hermite polynomial. The program is written in C. In addition, an interface to R has been made available by the PBSddesolve package of Schnute, Couture-Beil, and Haigh [226].

- SYSDEL by Karoui and Vaillancourt [155, 156] solves DDE-IVPs with multiple state-dependent delays by using explicit Runge-Kutta methods with a 3-point Hermite interpolation as continuous representation. SYSDEL tracks discontinuities and is implemented in FORTRAN77.

A small number of additional legacy codes is mentioned in Bellen and Zennaro [26] and Binder [32].

Features of the New Solver Colsol-DDE

The new solver Colsol-DDE, implemented in Fortran95, contributes to the existing collection of solvers in the following ways.

First of all it is observed that only a relatively small number of the above-listed codes are based on implicit methods. The exceptions are the implicit Runge-Kutta codes DDE-STRIDE and RADAR5, the Adams PECE methods in DELSOL, REBUS, SNDDELM and the Rosenbrock-type method in PAI4D. Colsol-DDE makes use of implicit Runge-Kutta methods of collocation type and uses exclusively implicit strategies in the stepsize-selection mechanism. It is thus among the few existing solvers that are able to solve stiff DDE-IVPs.

Second, Colsol-DDE is able to solve HDDE-IVPs and IHDDE-IVPs with multiple state-dependent delay and switching functions. Among the existing solvers, this feature is only found in dde23, ddesd, DDE_SOLVER, and Solv95. Note that none of these solvers is based on implicit methods.

A third property of Colsol-DDE that is not commonly found in DDE solvers is that it closely follows the definition of the practical variant of the modified standard approach. As discussed in Chapter 5, this incorporates the use of extrapolations if a current trial step is such that a deviating argument crosses a discontinuity point in the past. By using extrapolations, the determination of propagated discontinuity points can be done accurately and efficiently. Even among those DDE solvers that track discontinuities, only a few use extrapolations. To the author's knowledge, this is only the case in REBUS, DDEM, RADAR5, and RKFSWT. Note that, of these three codes, only RADAR5 is based on an implicit method, and none can solve IHDDE-IVPs.

The combination of features mentioned so far provides already sufficient justification for the presentation of a new solver. However, the most important property of Colsol-DDE – and the main reason for its development – is its capability to compute the derivative of the IVP solution with respect to parameters in the model functions. The property to compute these *sensitivities* is unique among IHDDE-IVP solvers. Furthermore, even for the case of DDEs, there is at present only one alternative solver that computes sensitivities, namely DDEM.

Organization of This Chapter

At first, Runge-Kutta methods of collocation type are introduced, which constitute the basis of Colsol-DDE (Section 6.1). After that, the implicit uniform correction procedure is presented for computing a continuous representation, whose uniform local order is equal to the discrete local order of the collocation method (Section 6.2). By application of an implicit quadrature rule, a higher order discrete approximation is obtained (Section 6.3). It is remarked that up to this point in the chapter, only ODEs are considered for notational simplicity.

The extension to DDEs follows in Section 6.4. Since Colsol-DDE relies on implicit methods in order to be suitable for stiff differential equations, nonlinear equation systems need to be solved in every integration step. Section 6.5 discusses the details of a Newton-type method that is used for this purpose. In Section 6.6 it is discussed how the results are used for constructing the error estimates. Investigating the stability properties both of the advancing method and of the error estimates is the subject of Section 6.7.

Section 6.8 gives an overview over the "core algorithm" in Colsol-DDE, i.e. everything that is needed for the solution of DDE-IVPs on intervals where the solution is smooth. The mechanisms for detecting discontinuities and including them into mesh, which are wrapped around this core algorithm, are the subject of the concluding Section 6.9.

For the sake of completeness, it should be mentioned that the presentation of the computational methods that are used in Colsol-DDE for the purpose of sensitivity computation is deferred to Chapter 9.

Notation

As in Chapter 5, the notation $g(\cdot,\cdot) \in \mathscr{C}^p(A_{x_1} \times A_{x_2}, \mathbb{R}^{n_g})$ means that the function g is p-times continuously differentiable with respect to both arguments x_1 and x_2 on the sets A_{x_1}, A_{x_2}. Further, the notation $g(\cdot, x_2) \in \mathscr{C}^p(A_{x_1}, \mathbb{R}^{n_g})$ means that for a given fixed x_2, the function g is p-times continuously differentiable with respect to x_1 on the set A_{x_1}.

Further, also in agreement with Chapter 5, the symbol $\|\cdot\|$ represents any of the norms $\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$ on a finite-dimensional space.

6.1. Runge-Kutta Methods of Collocation Type

The Collocation Solver for DDEs is based on a special class of continuous Runge-Kutta methods (CRK methods) called *Runge-Kutta methods of collocation type*. The methods are introduced here in the context of ODE-IVPs. The generalization to DDE-IVPs is considered later in Section 6.4.

6.1.1. Theoretical Background

For the presentation of this section, and also for the rest of the chapter, some of the notation of Chapter 5 is recalled. That is, for solving the ODE-IVP

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c) \tag{6.1a}$$

$$\mathfrak{y}(t^{ini}(c)) = y^{ini}(c), \tag{6.1b}$$

set $t_0 = t^{ini}(c)$, $y_0 = y^{ini}(c)$. Then apply, in each step $t_l \to t_{l+1} = t_l + h_{l+1}$, a CRK method with abscissae γ_i, weights β_j, continuous weight functions $b_j(\theta)$, and coefficients $a_{i,j}$:

$$y_{l+1} = y_l + h_{l+1} \sum_{j=1}^{\nu} \beta_j g_{l+1}^j \tag{6.2a}$$

$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1} \sum_{j=1}^{\nu} b_j(\theta) g_{l+1}^j \tag{6.2b}$$

$$g_{l+1}^j = f(t_{l+1}^j, y_{l+1}^j, c) \tag{6.2c}$$

$$y_{l+1}^j = y_l + h_{l+1} \sum_{k=1}^{\nu} a_{j,k} g_{l+1}^k. \tag{6.2d}$$

Herein, $t_{l+1}^j = t_l + \gamma_j h_{l+1}^j$. Further, y_{l+1} is the discrete approximation of the exact ODE-IVP solution y at t_{l+1}, and $\eta_{l+1}(t_l + \theta h_{l+1})$ is the continuous representation of the solution, which provides a continuous approximation of $y(t)$ for $t \in [t_l, t_{l+1}]$.

The defining property of *collocation methods* is that the time derivative of the continuous representation and the evaluation of the right-hand-side function coincide at the abscissae.

Definition 6.1 (Runge-Kutta Methods of Collocation Type, Collocation Methods)
Consider a CRK with ν stages and a continuous extension $\eta(t)$ that is given piecewise by polynomial functions $\eta_{l+1}(t)$ of degree ν. Assume that for any continuous ODE right-hand-side function $f(t, y(t), c)$ it holds that

$$\dot{\eta}_{l+1}(t_{l+1}^i) = f(t_{l+1}^i, \eta_{l+1}(t_{l+1}^i), c) \quad \textit{for} \quad 1 \leqslant i \leqslant \nu \quad \textit{and} \quad 0 \leqslant l \leqslant n_m - 1. \tag{6.3}$$

Then the Runge-Kutta method is called a Runge-Kutta method of collocation type*, or, in short, a* collocation method*.*

The weights β_j, continuous weight functions $b_j(\theta)$, and coefficients $a_{i,j}$ of a collocation method depend only on the abscissae γ_j, as the following theorem shows.

Theorem 6.2 (Weights, Continuous Weight Functions, and Coefficients for Collocation Methods)
Let γ_j, $1 \leqslant j \leqslant \nu$, be pairwise distinct abscissae of a collocation method. Then the collocation method has continuous weight functions, coefficients, and weights that are given by

$$b_j(\theta) = \int_0^\theta L_j(\theta')\, d\theta' \tag{6.4a}$$

$$a_{i,j} = b_j(\gamma_i) = \int_0^{\gamma_i} L_j(\theta')\, d\theta' \tag{6.4b}$$

$$\beta_j = b_j(1). \tag{6.4c}$$

Herein, $L_j(\theta)$, $1 \leqslant j \leqslant \nu$, are the Lagrange interpolation polynomials to the abscissae:

$$L_j(\theta) = \prod_{i=1, i\neq j}^{\nu} \frac{\theta - \gamma_i}{\gamma_j - \gamma_i}. \tag{6.5}$$

In turn, if the coefficients of a CRK method are given by equations (6.4), then it is a collocation method, i.e. its continuous extension $\eta_{l+1}(t)$ fulfills equation (6.3) in all integration steps and for all continuous right-hand-side functions f.

Proof (cf. Hairer, Wanner, and Lubich [128])
From the general form (6.2b) of continuous representations of RK methods and from the defining relation (6.3) of a collocation method, it follows that

$$\sum_{j=1}^{\nu} \dot{b}_j(\gamma_i) g_{l+1}^j = f(t_{l+1}^i, \eta_{l+1}(t_{l+1}^i), c) \quad \text{for} \quad 1 \leqslant i \leqslant \nu. \tag{6.6}$$

This equation should hold independent of the right-hand-side function f. Hence, set

$$\dot{b}_j(\gamma_i) = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{else} \end{cases}, \quad \text{for} \quad 1 \leqslant i \leqslant \nu, \quad 1 \leqslant j \leqslant \nu. \tag{6.7}$$

The degree of the interpolation polynomial of a collocation method is ν, hence the functions $\dot{b}_j(\theta)$ are polynomials of degree $\nu - 1$. By the ν conditions in equation (6.7), the functions $\dot{b}_j(\theta)$ are uniquely identified as Lagrange polynomials. From this, and by using the continuity of $\eta_{l+1}(t)$ at t_l, equations (6.4a) follows by integration.

Further, equation (6.4c) follows from the continuity of the continuous representation at t_{l+1}.

For equation (6.4b), recall on the one hand that $g_{l+1}^j = f(t_{l+1}^j, y_{l+1}^j, c)$ by definition, and observe on the other hand that equations (6.6) and (6.7) give $g_{l+1}^j = f(t_{l+1}^j, \eta_{l+1}(t_{l+1}^j), c)$. For arbitrary right-hand-side functions f, this only holds if $y_{l+1}^j = \eta_{l+1}(t_{l+1}^j)$. This directly yields the relations (6.4b).

For the reverse direction, simply insert the equations (6.4) into the general form of CRK methods and verify by standard analysis that the defining relation (6.3) of a collocation method is fulfilled. ■

In general, Theorem 6.2 allows to derive collocation methods for any arbitrary choice of pairwise distinct abscissae. The discrete and uniform local orders of the derived methods can then be determined by verifying the order conditions in Lemma 5.12. In particular, for the uniform local order, the following result is obtained.

Theorem 6.3 (Uniform Local Order of Collocation Methods)

The uniform local order of a collocation method with ν stages is given by $q = \nu$.

Proof
See Hairer, Nørsett, and Wanner [126], page 213f. ■

From Definition 5.5 it is clear that the discrete local order p is greater than or equal to the uniform local order q. In order to obtain a method with $p > q$, i.e. a so-called *superconvergent method*, the abscissae have to be chosen in a special way. For particular choices of the abscissae, the following methods are obtained:

- *Gauss collocation*, which uses for the abscissae γ_i, $1 \leqslant i \leqslant \nu$, the zeros of the function $d^\nu(x^\nu(x-1)^\nu)/dx^\nu$. The resulting methods have discrete local order $p = 2q = 2\nu$.
- *Radau IIA collocation*, which uses for the abscissae γ_i, $1 \leqslant i \leqslant \nu$, the zeros of the function $d^{\nu-1}(x^{\nu-1}(x-1)^\nu)/dx^{\nu-1}$. The last abscissa is always $\gamma_\nu = 1$. The resulting methods have discrete local order $p = 2q - 1 = 2\nu - 1$.
- *Lobatto IIIA collocation*, which uses for the abscissae γ_i, $1 \leqslant i \leqslant \nu$, $\nu \geqslant 2$, the zeros of the function $d^{\nu-2}(x^{\nu-1}(x-1)^{\nu-1})/dx^{\nu-2}$. The first and the last abscissae are always $\gamma_1 = 0$ and $\gamma_\nu = 1$. The resulting methods have discrete local order $p = 2q - 2 = 2\nu - 2$.

It is remarked that the methods that are commonly termed Radau IA, Lobatto IIIB, and Lobatto IIIC are so-called *discontinuous collocation methods*. The functions $\eta_{l+1}(t_l + \theta h_{l+1})$ implied by these methods fulfill the collocation condition (6.3), but violate the continuity condition (5.4) either at $\theta = 0$ or at $\theta = 1$ (or both), see Hairer, Wanner, and Lubich [128], page 36.

For the solution of ODE-IVPs, using a superconvergent method is computationally attractive because a high order approximation of the solution is obtained at the mesh points at comparably low computational costs. However, in the view of Theorem 5.21, superconvergence properties of a method are typically lost when applied to DDE-IVPs. Only for constant or time-dependent delays, and under severe restrictions on the stepsizes that lead to so-called *constrained meshes*, it is possible to obtain a superconvergent method for DDEs, see Bellen and Zennaro [26], page 159f. On the downside, the use of constrained meshes does not allow to select the stepsizes efficiently based on local error estimates (see Subsection 5.5.3), so that this approach is rarely used in practice.

6.1.2. Collocation Methods in Colsol-DDE

The basis of Colsol-DDE are the Gauss collocation method with one stage (i.e. the implicit midpoint rule) the Radau IIA collocation method with two stages, and the Lobatto IIIA collocation method with three stages. They have the Butcher tableaus given in Table 6.1 and the following continuous weight functions: $b_1(\theta) = \theta$ for the one-stage Gauss collocation method, $b_1(\theta) = -\frac{3}{4}\theta(\theta - 2)$ and $b_2(\theta) = \frac{3}{4}\theta(\theta - \frac{2}{3})$ for the two-stage Radau IIA collocation method, and $b_1(\theta) = 2\theta\left(\frac{1}{3}\theta^2 - \frac{3}{4}\theta + \frac{1}{2}\right)$, $b_2(\theta) = -4\theta^2\left(\frac{1}{3}\theta - \frac{1}{2}\right)$ and $b_3(\theta) = 2\theta^2\left(\frac{1}{3}\theta - \frac{1}{4}\right)$ for the three-stage Lobatto IIIA collocation method.

According to the general results for the discrete and uniform local orders of collocation methods, it follows that $(p, q) = (2, 1)$ for one-stage Gauss collocation, $(p, q) = (3, 2)$ for two-stage Radau IIA collocation, and $(p, q) = (4, 3)$ for three-stage Lobatto IIIA collocation. Hence, the employed methods are superconvergent, but the uniform local order is only one order lower than the discrete local order. This is important in order to ensure that the order of the discrete local error equals the discrete local order in the context of DDE-IVPs, in particular if overlapping occurs, see Section 6.4.

$\frac{1}{2}$	$\frac{1}{2}$
	1

(a)

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

(b)

0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$

(c)

Table 6.1.: Butcher Tableaus of (a) the one-stage Gauss collocation method, (b) the two-stage Radau IIA collocation method and (c) the three-stage Lobatto IIIA collocation method

6.2. The Uniform Correction Procedure

For the three collocation methods presented in Subsection 6.1.2 the discrete local order p and the uniform local order q are related in such a way that $q = p - 1$. Colsol-DDE uses the uniform correction procedure, suggested by Zennaro [268, 269] and Bellen and Zennaro [25, 26], for the computation of another continuous representation that has uniform local order p. For notational convenience, the application to ODE-IVPs is regarded first and the generalization to DDE-IVPs is deferred to Section 6.4.

6.2.1. Theoretical Background

Before formulating the algorithm for the uniform order corrections, some theory is established. For this purpose, the notion of the local problem is recalled from Chapter 5. The local problem in step

$t_l \to t_{l+1}$ is defined by

$$\dot{\mathbf{u}}_{l+1}(t) = f(t, \mathbf{u}_{l+1}(t), c) \tag{6.8a}$$
$$\mathbf{u}_{l+1}(t_l) = y_l. \tag{6.8b}$$

In the following, let $u_{l+1}(t)$ be the exact solution of the local problem.

Definition 6.4 (Natural Continuous Extensions, Asymptotic Orthogonality Condition)
If, for any right-hand-side function $f(\cdot,\cdot,c) \in \mathscr{C}^p(\mathbb{R}\times\mathbb{R}^{n_y}, \mathbb{R}^{n_y})$, the continuous representation $\eta(t)$ satifies the asymptotic orthogonality condition

$$\int_{t_l}^{t_{l+1}} G(t)[\dot{u}_{l+1}(t) - \dot{\eta}(t)]dt = \mathscr{O}(h_{l+1}^{p+1}) \tag{6.9}$$

uniformly on all intervals $[t_l, t_{l+1}]$ and for all matrix-valued functions $G(t) \in \mathscr{C}^p(\mathbb{R}, \mathbb{R}^{n_G \times n_y})$, then the continuous representation $\eta(t)$ of the CRK method is called a natural continuous extension.

Concerning the existence of natural continuous extensions, regard the following theorem.

Theorem 6.5 (Existence of Natural Continuous Extensions)
For any CRK method of discrete local order p, there exists a natural continuous extension that is at least of uniform local order $q = \lfloor \frac{p+1}{2} \rfloor$.

Proof
See Bellen and Zennaro [26], page 124. ■

For reasons of efficiency, the uniform correction procedure that is presented later in this section makes use of *time points of inner superconvergence*, defined as follows.

Definition 6.6 (Time Points of Inner Superconvergence)
Assume that the continuous extension of a CRK method is of uniform local order q, i.e. it holds

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|u_{l+1}(t) - \eta_{l+1}(t)\| = \mathscr{O}(h_{l+1}^{q+1}) \tag{6.10a}$$
$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|\dot{u}_{l+1}(t) - \dot{\eta}_{l+1}(t)\| = \mathscr{O}(h_{l+1}^{q}) \tag{6.10b}$$

for all intervals and all sufficiently smooth right-hand-side functions f. If there exist inner time points $t_l + \bar{\theta}h_{l+1}$, $\bar{\theta} \in (0,1)$, or $t_l + \bar{\bar{\theta}}h_{l+1}$, $\bar{\bar{\theta}} \in (0,1)$, such that

$$\|u_{l+1}(t_l + \bar{\theta}h_{l+1}) - \eta_{l+1}(t_l + \bar{\theta}h_{l+1})\| = \mathscr{O}(h_{l+1}^{q+2}) \tag{6.11a}$$
$$\|\dot{u}_{l+1}(t_l + \bar{\bar{\theta}}h_{l+1}) - \dot{\eta}_{l+1}(t_l + \bar{\bar{\theta}}h_{l+1})\| = \mathscr{O}(h_{l+1}^{q+1}) \tag{6.11b}$$

for all intervals and all sufficiently smooth right-hand-side functions f, then the times $\bar{\theta}$ and $\bar{\bar{\theta}}$ are called time points of inner superconvergence for $u_{l+1}(t)$ and $\dot{u}_{l+1}(t)$.

In the context of Colsol-DDE, the interior abscissae of the collocation methods are time points of inner superconvergence for $\dot{u}_{l+1}(t)$.

Lemma 6.7 (Time Points of Inner Superconvergence for Collocation Methods)
Let $\gamma_j \in (0,1)$ be an interior abscissa of a collcation method and assume that the right-hand-side function $f(t,y,c)$ is continuous and Lipschitz continuous with respect to y with Lipschitz constant L_f. Then γ_j is a time point of inner superconvergence for $\dot{u}_{l+1}(t)$.

Proof
At the abscissae of the collocation method it holds that $\dot{\eta}_{l+1}(t_l + \gamma_j h_{l+1}) = f(t_l + \gamma_j h_{l+1}, \eta_{l+1}(t_l + \gamma_j h_{l+1}), c)$. Therefore, it follows that

$$\begin{aligned}
&\|\dot{u}_{l+1}(t_l + \gamma_j h_{l+1}) - \dot{\eta}_{l+1}(t_l + \gamma_j h_{l+1})\| \\
=&\|f(t_l + \gamma_j h_{l+1}, u_{l+1}(t_l + \gamma_j h_{l+1}), c) - f(t_l + \gamma_j h_{l+1}, \eta_{l+1}(t_l + \gamma_j h_{l+1}), c)\| \\
\leqslant& L_f \|u_{l+1}(t_l + \gamma_j h_{l+1}) - \eta_{l+1}(t_l + \gamma_j h_{l+1})\|.
\end{aligned} \tag{6.12}$$

By the uniform local order q of the CRK method, it follows that this term is $\mathscr{O}(h_{l+1}^{q+1})$. ■

In the context of Gauss collocation methods, Zennaro [268] has presented an algorithm that starts with the collocation polynomial (which has uniform local order q) and constructs, successively, continuous representations of higher order up to a uniform local order $\tilde{q}$, $q < \tilde{q} \leqslant p$. The concept of natural continuous extensions made it possible to generalize the approach such that it can be applied to all Runge-Kutta methods, see Zennaro [269]. For stiff ODE-IVPs, an implicit variant is suggested in Bellen and Zennaro [25] that exhibits better stability properties, see also Section 6.7 below. Bellen and Zennaro [26] contains a modification of the implicit variant, which is replicated here:

Algorithm 6.8 (Implicit Uniform Correction Procedure)

1. *Start with a CRK method of discrete local order p, which employs a natural continuous representation $\eta_{l+1}(t)$ of uniform local order q. Set $r = q$. Further, for a shorter notation, define $\varphi_r(t) := \eta_{l+1}(t)$, thereby dropping the subscript $l+1$ of the current step and including, instead, the uniform local order r. The goal is to determine a polynomial continuous representation $\varphi_{r+1}(t)$ of degree $r+1$ by formulating $r+2$ linearly independent conditions on its coefficients.*

2. *Identify the K_1' and K_2' time points of inner superconvergence of $\varphi_r(t)$ for $u_{l+1}(t)$ and $\dot{u}_{l+1}(t)$ (possibly $K_1' = 0$ and/or $K_2' = 0$).*

3. *Remove from the so-defined set of time points of inner superconvergence as many points as necessary such that the following $K_1 + K_2 + p - r + 1$ conditions on φ_{r+1} are linearly independent:*

   (i) $\varphi_{r+1}(t_l + \bar{\theta}_k h_{l+1}) = \varphi_r(t_l + \bar{\theta}_k h_{l+1})$ *for* $1 \leqslant k \leqslant K_1$, $(K_1 \leqslant K_1')$.

   (ii) $\dot{\varphi}_{r+1}(t_l + \bar{\bar{\theta}}_k h_{l+1}) = \dot{\varphi}_r(t_l + \bar{\bar{\theta}}_k h_{l+1})$ *for* $1 \leqslant k \leqslant K_2$, $(K_2 \leqslant K_2')$.

   (iii) $\int_{t_l}^{t_{l+1}} (t - t_l)^i \dot{\varphi}_{r+1}(t) dt = \int_{t_l}^{t_{l+1}} (t - t_l)^i \dot{\varphi}_r(t) dt$ *for* $1 \leqslant i \leqslant p - r - 1$.

   (iv) $\varphi_{r+1}(t_l) = y_l$, *and* $\varphi_{r+1}(t_{l+1}) = y_{l+1}$.

4. *Choose θ_k^*, $1 \leqslant k \leqslant K_3$, $K_3 = 2r+1-p-K_1-K_2 \geqslant 0$ such that the following collocation-like conditions are linearly independent of the $K_1 + K_2 + p - r + 1$ conditions given in step 3:*

   (v) $\dot{\varphi}_{r+1}(t_l + \theta_k^* h_{l+1}) = f(t_l + \theta_k^* h_{l+1}, \varphi_{r+1}(t_l + \theta_k^* h_{l+1}), c)$.

   Thereby $\theta_k^* \neq 0$ and $\theta_k^* \neq 1$, unless 0 or 1 are abscissae of the basic CRK method.

5. *Solve the $r+2$-dimensional equation system of the conditions (i)-(v), which uniquely defines the polynomial continuous representation $\varphi_{r+1}(t)$ of degree $r+1$.*

6. *The steps 2 to 5 can be repeated with $r \to r+1$ until a polynomial continuous representation φ_m of the desired degree $q < m \leqslant p$ is obtained.*

The following theorem constitutes the main result for the implicit uniform correction procedure.

Theorem 6.9 (Properties of the Polynomial Continuous Representations obtained by Algorithm 6.8)

The polynomials φ_m, for $q+1 \leqslant m \leqslant p$, generated by Algorithm 6.8, have uniform local order m, and they fulfill the asymptotic orthogonality condition (6.9).

Proof
Zennaro [268] contains a proof of the original uniform order correction method. This can be generalized to the implicit variant quoted here. ■

It is remarked that a violation of the requirement $\theta_k^* \neq 0$ and $\theta_k^* \neq 1$ on the abscissae of the additional collocation-like conditions 4 does not affect the conclusions of Theorem 6.9. These requirements are imposed for stability reasons only.

6.2.2. Application to the Collocation Methods in Colsol-DDE

Gauss Collocation

Regard first the Gauss collocation method with one stage, which has discrete local order 2 and which provides, on each interval $[t_l, t_{l+1}]$, a continuous representation $\varphi_1(t) := \eta_{l+1}(t)$ of uniform local order 1. By application of Algorithm 6.8, a polynomial continuous representation $\varphi_2(t)$ should be determined that has degree and order 2. It is convenient to make the following ansatz:

$$\varphi_2(t) = \varphi_1(t) + \delta(t), \tag{6.13}$$

where

$$\delta(t_l + \theta h_{l+1}) = \delta_0 + \delta_1\theta + \delta_2\theta^2. \tag{6.14}$$

Three linear independent conditions are needed to uniquely determine the coefficients δ_0, δ_1, and δ_2. From the continuity conditions (iv) in step 3 of Algorithm 6.8, the two equations

$$\delta_0 = 0 \quad \text{and} \quad \delta_0 + \delta_1 + \delta_2 = 0 \tag{6.15}$$

are obtained. Due to the collocation condition, the continuous representation $\varphi_1(t)$ has an inner superconvergence point at $\theta = 1/2$, hence $K_2' = 1$. However, condition (ii) in Algorithm 6.8 on $\dot{\delta}(t_l + \theta h_{l+1})$ at $\theta = 1/2$ yields

$$\delta_1 + \delta_2 = 0, \tag{6.16}$$

which is linearly dependent on the two previous conditions and can therefore not be exploited. Therefore, $K_2 = 0$ and that $K_3 = 1$. Using an arbitrary point θ_1^* in the additional collocation-like condition (v), the following relation is obtained:

$$\begin{aligned}\dot{\delta}(t_l + \theta_1^* h_{l+1}) =& \frac{\delta_1 + 2\delta_2\theta_1^*}{h_{l+1}} \\ =& f(t_l + \theta_1^* h_{l+1}, \varphi_1(t_l + \theta_1^* h_{l+1}) + \delta(t_l + \theta_1^* h_{l+1}), c) - \dot{\varphi}_1(t_l + \theta_1^* h_{l+1}) \\ =:& g^*(\theta_1^*).\end{aligned} \tag{6.17}$$

Together, the equations (6.15) and (6.17) give

$$\delta_0 = 0, \quad \delta_1 = h_{l+1}\frac{g^*(\theta_1^*)}{1 - 2\theta_1^*}, \quad \delta_2 = -h_{l+1}\frac{g^*(\theta_1^*)}{1 - 2\theta_1^*} \tag{6.18}$$

provided that $\theta_1^* \neq 1/2$. In Colsol-DDE, the choice is $\theta_1^* = 1/3$. It therefore follows that

$$\delta(t_l + \theta h_{l+1}) = -3(\theta^2 - \theta)h_{l+1}g^*\left(\frac{1}{3}\right). \tag{6.19}$$

By defining $g_{l+1}^* := g^*(1/3)$ and $b_*(\theta) = -3(\theta^2 - \theta)$, the continuous representation $\varphi_2(t)$ of order 2 can be expressed as

$$\varphi_2(t) = \underbrace{y_l + h_{l+1}b_1(\theta)g_{l+1}^1}_{\varphi_1(t)} + h_{l+1}b_*(\theta)g_{l+1}^*. \tag{6.20}$$

Radau IIA Collocation

Consider next the Radau IIA collocation method with two stages, which has discrete local order 3 and which provides, on each interval $[t_l, t_{l+1}]$, a continuous representation $\varphi_2(t)$ of uniform local order 2. The goal is to find a continuous representation $\varphi_3(t)$ of degree and uniform local order 3. It is again convenient to determine the difference $\delta(t)$ of the two polynomial continuous representations such that

$$\varphi_3(t) = \varphi_2(t) + \delta(t). \tag{6.21}$$

The goal is to find four linear independent conditions for the four coefficients of the polynomial function $\delta(t)$:

$$\delta(t_l + \theta h_{l+1}) = \delta_0 + \delta_1\theta + \delta_2\theta^2 + \delta_3\theta^3. \tag{6.22}$$

From the continuity conditions (iv) it follows that

$$\delta_0 = 0 \quad \text{and} \quad \delta_0 + \delta_1 + \delta_2 + \delta_3 = 0. \tag{6.23}$$

The Radau IIA method with two stages has one interior collocation stage at 1/3, i.e. $K_2' = 1$. The corresponding condition (ii) on $\dot{\delta}(t_l + \theta h_{l+1})$ at $\theta = 1/3$ reads

$$\delta_1 + \frac{2}{3}\delta_2 + \frac{1}{3}\delta_3 = 0, \tag{6.24}$$

which is linearly independent of the two equations obtained from the continuity conditions (iv). Hence, it follows that $K_2 = 1$ and that $K_3 = 1$, i.e. one additional collocation-like condition (v) needs to be formulated in order to obtain a continuous representation of uniform local order 3. This yields

$$\begin{aligned}\dot{\delta}(t_l + \theta_1^* h_{l+1}) &= \frac{\delta_1 + 2\delta_2\theta_1^* + 3\delta_3(\theta_1^*)^2}{h_{l+1}} \\ &= f(t_l + \theta_1^* h_{l+1}, \varphi_2(t_l + \theta_1^* h_{l+1}) + \delta(t_l + \theta_1^* h_{l+1}), c) - \dot{\varphi_2}(t_l + \theta_1^* h_{l+1}). \\ &=: g^*(\theta_1^*)\end{aligned} \tag{6.25}$$

Together with the equations (6.23) and (6.24), this gives

$$\delta_0 = 0, \quad \delta_1 = \delta_3 = h_{l+1}\frac{g^*(\theta_1^*)}{3(\theta_1^*)^2 - 4\theta_1^* + 1}, \quad \delta_2 = -2h_{l+1}\frac{g^*(\theta_1^*)}{3(\theta_1^*)^2 - 4\theta_1^* + 1}, \tag{6.26}$$

provided that $\theta_1^* \neq 1/3$ and $\theta_1^* \neq 1$. In Colsol-DDE, $\theta_1^* = 1/6$ is used, and therefore

$$\delta(t_l + \theta h_{l+1}) = \frac{12}{5}(\theta^3 - 2\theta^2 + \theta)h_{l+1}g^*\left(\frac{1}{6}\right). \tag{6.27}$$

By defining $g_{l+1}^* := g^*(1/6)$ and $b_*(\theta) = \frac{12}{5}(\theta^3 - 2\theta^2 + \theta)$, the polynomial continuous representation $\varphi_3(t)$ of order 3 can be expressed as

$$\varphi_3(t_l + \theta h_{l+1}) = \underbrace{y_l + h_{l+1}\sum_{i=1}^{2} b_i(\theta)g_{l+1}^i}_{\varphi_2(t)} + h_{l+1}b_*(\theta)g_{l+1}^*. \tag{6.28}$$

Lobatto IIIA Collocation

The Lobatto IIIA collocation method with 3 stages has discrete local order 4 and implies a continuous representation $\varphi_3(t)$ of uniform local order 3. By applying Algorithm 6.8, a polynomial continuous representation $\varphi_4(t)$ of uniform local order 4 can be determined as follows. As usual, the ansatz is

$$\varphi_4(t) = \varphi_3(t) + \delta(t) \tag{6.29}$$

with

$$\delta(t_l + \theta h_{l+1}) = \delta_0 + \delta_1\theta + \delta_2\theta^2 + \delta_3\theta^3 + \delta_4\theta^4. \tag{6.30}$$

Five linear independent conditions are needed in order to determine the coefficients δ_i. From the continuity conditions, the equations

$$\delta_0 = 0 \quad \text{and} \quad \delta_0 + \delta_1 + \delta_2 + \delta_3 + \delta_4 = 0 \tag{6.31}$$

follow. The Lobatto IIIA method with 3 stages has one interior collocation stage at $\theta = 1/2$, i.e. $K_2' = 1$. Condition (ii) on $\dot{\delta}(t_l + \theta h_{l+1})$ at $\theta = 1/2$ reads

$$\delta_1 + \delta_2 + \frac{3}{4}\delta_3 + \frac{1}{2}\delta_4 = 0, \tag{6.32}$$

which is linearly independent of the equations (6.31).

It follows that $K_2 = 1$ and that $K_3 = 2$, i.e. two additional collocation-like conditions (v) need to be formulated in order to obtain a continuous representation of order 4. For reasons of efficiency, one of the additional stages is chosen to be $\theta_1^* = 0$, which is acceptable because the Lobatto IIIA collocation method already employs this abscissa, and which requires no extra computation. This yields

$$\delta_1 = 0. \tag{6.33}$$

For an arbitrary second additional stage θ_2^* the relation

$$\begin{aligned}\dot{\delta}(t_l + \theta_2^* h_{l+1}) =& \frac{\delta_1 + 2\theta_2^*\delta_2 + 3(\theta_2^*)^2\delta_3 + 4(\theta_2^*)^3\delta_4}{h_{l+1}} \\ =& f(t_l + \theta_2^* h_{l+1}, \varphi_3(t_l + \theta_2^* h_{l+1}) + \delta(t_l + \theta_2^* h_{l+1}), c) - \dot{\varphi}_3(t_l + \theta_2^* h_{l+1}) \\ =:& g^*(\theta_2^*)\end{aligned} \tag{6.34}$$

is obtained. Together with the conditions (6.31), (6.32), and (6.33), this gives

$$\delta_0 = \delta_1 = 0, \quad \delta_2 = \delta_4 = h_{l+1}\frac{g^*(\theta_2^*)}{4(\theta_2^*)^3 - 6(\theta_2^*)^2 + 2\theta_2^*}, \quad \delta_3 = -2h_{l+1}\frac{g^*(\theta_2^*)}{4(\theta_2^*)^3 - 6(\theta_2^*)^2 + 2\theta_2^*} \tag{6.35}$$

if $\theta_2^* \neq 0$, $\theta_2^* \neq 1/2$, and $\theta_2^* \neq 1$. Colsol-DDE uses $\theta_2^* = 1/4$, so that the correction polynomial becomes

$$\delta(t_l + \theta h_{l+1}) = \frac{16}{3}(\theta^4 - 2\theta^3 + \theta^2)h_{l+1}g_{l+1}^*, \quad g_{l+1}^* := g^*\left(\frac{1}{4}\right). \tag{6.36}$$

By defining $g_{l+1}^* := g^*(1/4)$ and $b_*(\theta) = \frac{16}{3}(\theta^4 - 2\theta^3 + \theta^2)$, the continuous representation $\varphi_4(t)$ of order 4 can be expressed as

$$\varphi_4(t) = \underbrace{y_l + h_{l+1}\sum_{i=1}^{3} b_i(\theta)g_{l+1}^i}_{\varphi_3(t)} + h_{l+1}b_*(\theta)g_{l+1}^*. \tag{6.37}$$

The "Augmented" CRK Methods

For all three collocation methods implemented in Colsol-DDE, one additional stage is sufficient to obtain a continuous representation whose uniform local order is p. Formally the introduction of this extra stage defines "augmented" CRK methods with Butcher tableaus given in Table 6.2.

The continuous representation associated with the augmented CRK methods are denoted by $\eta_{l+1,p}(t_l + \theta h_{l+1}) := \varphi_p(t_l + \theta h_{l+1})$, where the subscripts now refers to both the index of the current step and to the uniform local order. Denoting, in analogy, the continuous representation of the basic collocation method as $\eta_{l+1,q}(t_l + \theta h_{l+1})$, yields

$$\eta_{l+1,p}(t_l + \theta h_{l+1}) = \eta_{l+1,q}(t_l + \theta h_{l+1}) + h_{l+1}b_*(\theta)g_{l+1}^*. \tag{6.38}$$

Thereby, $b_*(\theta) = -3(\theta^2 - \theta)$ for the uniform correction to the one-stage Gauss collocation method, $b_*(\theta) = \frac{12}{5}(\theta^3 - 2\theta^2 + \theta)$ for the uniform correction to the two-stage Radau IIA collocation method, and $b_*(\theta) = \frac{16}{3}(\theta^4 - 2\theta^3 + \theta^2)$ for the uniform correction to the three-stage Lobatto IIIA collocation method.

Later, e.g. in Section 6.4, it will become necessary to denote explicitly both the index of the step and the uniform local order of the continuous representation. However, in the next section, the short-hand notation $\varphi_q(t)$ and $\varphi_p(t)$ will still be used.

$\frac{1}{2}$	$\frac{1}{2}$	0
$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
	1	0

(a)

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$	0
1	$\frac{3}{4}$	$\frac{1}{4}$	0
$\frac{1}{6}$	$\frac{11}{48}$	$-\frac{3}{48}$	$\frac{5}{18}$
	$\frac{3}{4}$	$\frac{1}{4}$	0

(b)

0	0	0	0	0
$\frac{1}{2}$	$\frac{5}{24}$	$\frac{1}{3}$	$-\frac{1}{24}$	0
1	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	0
$\frac{1}{4}$	$\frac{1}{6}$	$\frac{5}{48}$	$-\frac{1}{48}$	$\frac{3}{16}$
	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	0

(c)

Table 6.2.: Butcher Tableaus of the augmented CRK methods, taking into account the extra stages from the uniform correction procedure.

6.3. A Quadrature Rule Applied to Polynomial Continuous Representations

The basic collocation method provides, at each mesh point t_{l+1}, an approximation y_{l+1} of discrete local order p of the exact solution of the local problem at the new mesh point, i.e. of $u_{l+1}(t_{l+1})$. From now on, this approximation is more precisely called $y_{l+1,p}$. In this section it is discussed how a continuous representation of maximum uniform local order p, i.e. $\varphi_p(t)$, can be used to obtain an approximation $y_{l+1,p+1}$ that has order $p+1$. The method is introduced here in the context of ODE-IVPs, and the generalization to DDE-IVPs is done later in Section 6.4.

6.3.1. Theoretical Background

Theorem 6.10 (A Higher Order Discrete Method by Integration)

Consider a CRK method for which the uniform local order of the continuous representation $\varphi_p(t)$ equals the discrete local order p. The right-hand-side function $f(t,y,c)$ of the ODE-IVP shall be Lipschitz continuous with respect to y. Then

$$y_{int} := y_l + \int_{t_l}^{t_{l+1}} f(t, \varphi_p(t), c)dt \tag{6.39}$$

approximates the exact solution of the local problem at the new mesh point, i.e. $u_{l+1}(t_{l+1})$, with discrete local order $p+1$:

$$\|u_{l+1}(t_{l+1}) - y_{int}\| = \mathscr{O}(h_{l+1}^{p+2}). \tag{6.40}$$

Proof
The result is obtained as follows:

$$\begin{aligned}\|u_{l+1}(t_{l+1}) - y_{int}\| &= \left\| \int_{t_l}^{t_{l+1}} f(t, u_{l+1}(t), c) - f(t, \varphi_p(t), c)dt \right\| \\ &\leqslant L_f \int_{t_l}^{t_{l+1}} \|u_{l+1}(t) - \varphi_p(t)\|\, dt \\ &\leqslant L_f h_{l+1} \max_{t_l \leqslant t \leqslant t_{l+1}} \|u_{l+1}(t) - \varphi_p(t)\|. \end{aligned} \tag{6.41}$$

Since $\varphi_p(t)$ has uniform local order p, the expression in the last row is obviously $\mathscr{O}(h_{l+1}^{p+2})$. ■

Note that $f(t, \varphi_p(t), c)$ is a general nonlinear function, hence y_{int} can in general not be computed directly. Fortunately, the discrete local order $p+2$ is also obtained for any quadrature rule that is accurate for polynomials of order p or higher, as seen in the following theorem.

Theorem 6.11 (Order Result for an Approximation by a Quadrature Rule)

Consider a μ-stage quadrature rule applied to the integral $\int_{t_l}^{t_{l+1}} f(t, \varphi_p(t), c)$, such that

$$y_{l+1,p+1} := y_l + h_{l+1} \sum_{i=1}^{\mu} B_i f(t_l + \Gamma_i h_{l+1}, \varphi_p(t_l + \Gamma_i h_{l+1}), c), \tag{6.42}$$

where Γ_i and B_i are the abscissae and weights of the quadrature rule, and where $\varphi_p(t)$ is a polynomial continuous representation of order p. If the quadrature rule is exact for polynomials of order p (or higher) and the right-hand-side function is $f(\cdot, \cdot, c) \in \mathscr{C}^{p+1}(\mathbb{R} \times \mathbb{R}^{n_y}, \mathbb{R}^{n_y})$, then it holds that

$$\|u_{l+1}(t_{l+1}) - y_{l+1,p+1}\| = \mathscr{O}(h_{l+1}^{p+2}). \tag{6.43}$$

Proof
Consider an interpolation polynomial $\rho(t)$ of $f(t, \varphi_p(t), c)$ of degree and order p. Since the right-hand-side function is $p+1$-times continuously differentiable, standard results for the interpolation error yield

$$f_i(t, \varphi_p(t), c) = \rho_i(t) + \mathscr{O}(h_{l+1}^{p+1}) \tag{6.44}$$

for $t_l \leqslant t \leqslant t_{l+1}$ and all components $i = 1, \ldots, n_y$. Accordingly, it holds that

$$y_{int} = y_l + \int_{t_l}^{t_{l+1}} \rho(t) dt + \mathscr{O}(h_{l+1}^{p+2}). \tag{6.45}$$

Consider then

$$\|u_{l+1}(t_{l+1}) - y_{l+1,p+1}\| \leqslant \|u_{l+1}(t_{l+1}) - y_{int}\| + \|y_{int} - y_{l+1,p+1}\|, \tag{6.46}$$

where the first term in the right hand side is $\mathscr{O}(h_{l+1}^{p+2})$ by Theorem 6.10. Further, the second term is also $\mathscr{O}(h_{l+1}^{p+2})$ because of equation (6.45) and because the employed quadrature rule is exact for polynomials up to order p.

The use of quadrature formulas that are exact for polynomials of higher degree than p does not affect the result, because the error in the first term on the right hand side of equation (6.46) remains $\mathscr{O}(h_{l+1}^{p+2})$. ■

The quadrature rule in equation (6.42) yields an approximation of $u_{l+1}(t_{l+1})$ of the desired order $p+1$. Unfortunately, good stability properties of the continuous representation $\varphi_p(t)$ are typically not transferred to the discrete approximation $y_{l+1,p+1}$ because of the explicit nature of the quadrature formula (6.42). Therefore, Bellen and Zennaro [25, 26] suggest an implicit variant:

Theorem 6.12 (Order Result for an Approximation by an Implicit Quadrature Rule)

Let the assumptions of Theorem 6.11 be fulfilled and let the last abscissa be $\Gamma_\mu = 1$. Assume that the stepsize is $h_{l+1} \leqslant h_0 := B_\mu L_f / 2$. Then the implicit variant of the quadrature rule defined by

$$\begin{aligned} y_{l+1,p+1}^{impl} :=& y_l + h_{l+1} \sum_{i=1}^{\mu-1} B_i f(t_l + \Gamma_i h_{l+1}, \varphi_p(t_l + \Gamma_i h_{l+1}), c) \\ &+ h_{l+1} B_\mu f(t_l + h_{l+1}, y_{l+1,p+1}^{impl}, c) \end{aligned} \tag{6.47}$$

also provides an approximation such that

$$\|u_{l+1}(t_{l+1}) - y_{l+1,p+1}^{impl}\| = \mathscr{O}(h_{l+1}^{p+2}). \tag{6.48}$$

Proof
Because of Theorem 6.11, the assertion apparently follows if the results of the explicit and of the implicit quadrature rule differ by $\|y_{l+1,p+1} - y_{l+1,p+1}^{impl}\| = \mathscr{O}(h_{l+1}^{p+2})$.

Subtraction of equation (6.42) from (6.47) yields

$$\begin{aligned}\|y_{l+1,p+1} - y^{impl}_{l+1,p+1}\| &= h_{l+1}B_\mu \Big\| \Big(f(t_{l+1}, \varphi_p(t_{l+1}), c) - f(t_{l+1}, y^{impl}_{l+1,p+1}, c) \Big) \Big\| \\ &\leqslant h_{l+1}B_\mu L_f \Big(\|\varphi_p(t_{l+1}) - u_{l+1}(t_{l+1})\| + \|u_{l+1}(t_{l+1}) - y_{l+1,p+1}\| \\ &\qquad + \|y_{l+1,p+1} - y^{impl}_{l+1,p+1}\| \Big). \end{aligned} \tag{6.49}$$

Since $\varphi_p(t_l + h_{l+1})$ is identical to the discrete approximation $y_{l+1,p}$ obtained with the collocation method, the first term in brackets is clearly $\mathscr{O}(h_{l+1}^{p+1})$. Further, the second term is also $\mathscr{O}(h_{l+1}^{p+1})$ because of Theorem 6.11. Hence, the expression

$$(1 - h_{l+1}B_\mu L_f)\|y_{l+1,p+1} - y^{impl}_{l+1,p+1}\| = \mathscr{O}(h_{l+1}^{p+2}) \tag{6.50}$$

is obtained. Clearly, for $h_{l+1} \leqslant h_0$, it holds that $(1 - h_{l+1}B_\mu L_f) \in [1/2, 1]$. This completes the proof. ■

For the remainder of this chapter, only the implicit version of the quadrature rule is considered for obtaining a higher order discrete approximation. Therefore, the superscript *impl* is from now on dropped.

It is remarked that solving the equation system (6.47) for $y_{l+1,p+1}$ is equivalent to solving the following equation system for its unknowns $g^{\diamondsuit}_{l+1}$:

$$g^{\diamondsuit}_{l+1} = f\Big(t_{l+1}, y_l + h_{l+1} \sum_{i=1}^{\mu-1} B_i f(t_l + \Gamma_i h_{l+1}, \varphi_p(t_l + \Gamma_i h_{l+1}), c) + h_{l+1}B_\mu g^{\diamondsuit}_{l+1}, c\Big) \tag{6.51}$$

The systems are transformed into each other by the definition

$$g^{\diamondsuit}_{l+1} = f(t_{l+1}, y_{l+1,p+1}, c). \tag{6.52}$$

6.3.2. Implementation in Colsol-DDE

In order to obtain approximations of discrete local order $p + 1$, quadrature formulas are needed that are accurate for polynomials of order 2, 3, and 4 for Gauss, Radau IIA, and Lobatto IIIA collocation respectively.

Colsol-DDE uses superconvergent quadrature formulas in order to minimize the number of extra function evaluations.

In particular, for the one-stage Gauss collocation method the two-stage Radau quadrature formula with abscissa at 1 is employed, which is exact for polynomials up to degree 2. For the two-stage Radau IIA method, the three-stage Lobatto quadrature formula with abscissa at 1 is used, which is exact for polynomials up to order 3. Eventually, for the three-stage Lobatto IIIA method, the three-stage Radau quadrature formula with abscissa at 1 is used, which is exact for polynomials up to order 4.

6.4. Extension to DDE-IVPs, Computation of Past States

In the Sections 6.1, 6.2, and 6.3 the basic integration methods that are implemented in Colsol-DDE were presented in the context of ODE-IVPs. This section is concerned with the generalization of the methods to the solution of DDE-IVPs. The focus lies on the realization of the computation of past states, which is done in such a way that all employed methods (collocation method, uniform correction, implicit quadrature rule) always attain their discrete and uniform local order.

In this section, only a single integration step $t_l \to t_{l+1}$ is considered, and for the purpose of computing the past states it is assumed that the discontinuity interval indicators are available. How the discontinuity interval indicators are obtained in practice in Colsol-DDE is the subject of Section 6.9.

It is remarked that the solution of IHDDE-IVPs requires, in addition, to monitor the signs of the switching functions, to locate the zeros of the switching functions, and to apply the impulses. However, on the level of a single integration step $t_l \to t_{l+1}$, the presence of switching functions and

impulses in the problem formulation does not impose additional difficulties and therefore does not need to be considered here.

For notational simplicity the presentation is restricted to the case of a single delay τ_1. The extension to the case of multiple delays is straightforward.

As introduced in the previous sections, let $y_{l+1,p}$ and $\eta_{l+1,q}(t_l + \theta h_{l+1})$ denote the lower order discrete and continuous approximations of the exact solution of the local problem, i.e. $u_{l+1}(t_{l+1})$ and $u_{l+1}(t_l + \theta h_{l+1})$, which are obtained by the collocation methods. Further, let $y_{l+1,p+1}$ and $\eta_{l+1,p}(t_l + \theta h_{l+1})$ denote the higher order discrete and continuous approximations, which are obtained by the implicit quadrature rule and by the implicit uniform correction procedure, respectively.

6.4.1. Collocation Method

Consider first the application of one of the collocation methods in Colsol-DDE for the solution of a DDE-IVP. In every step $t_l \to t_{l+1} = t_l + h_{l+1}$, the following $\nu \cdot n_y$-dimensional equation system is solved:

$$g^j_{l+1} = f(t^j_{l+1},\ y^j_{l+1},\ c,\ v^j_{l+1}) \quad \text{for} \quad 1 \leqslant j \leqslant \nu. \tag{6.53}$$

For the Lobatto IIIA collocation method, the first stage is explicit, and hence the dimension of the system is only $(\nu - 1) \cdot n_y$.

In equation (6.53), y^j_{l+1} is defined by

$$y^j_{l+1} = y_l + h_{l+1} \sum_{i=1}^{\nu} a_{j,i} g^i_{l+1}. \tag{6.54}$$

Further, v^j_{l+1} represents the numerical approximation of the past state. Since the implementation of Colsol-DDE follows the ideas of the practical variant of the modified standard approach, the computation is primarily determined by the value of the discontinuity interval indicator ξ^α_1 for the sole deviating argument $\alpha_1(t, y, c) = t - \tau_1(t, y, c)$.

In the following, it is described how the use of extrapolations is realized for the computation of trial steps in Colsol-DDE. Therefore, as in Definition 5.22, denote the time points of the initial discontinuities up to order p by $\hat{s}_{-n^\phi_s}, \ldots, \hat{s}_{-1}$ and let, by convention, $\hat{s}_0 = t^{ini}(c)$, regardless of whether or not the initial time is a point of discontinuity of order less than p (typically, this will be the case). Further, let $\hat{s}_j$, $1 \leqslant j \leqslant n_s$, be the time points of the n_s propagated discontinuities that are detected by the methods presented in Section 6.9 until the mesh point t_l. This means that each time point $\hat{s}_k$ is an (approximate) zero of a propagation switching function

$$\sigma^\alpha_{1,\hat{s}_j}(t, \eta(t), c) = t - \alpha_1(t, \eta(t), c) - \hat{s}_j \tag{6.55}$$

for at least one $j < k$. The mechanisms that are implemented in Colsol-DDE to check for uniqueness of the discontinuity interval indicator are described in Section 6.9.

Computation of Past States from the Initial Function

The computation of past states is carried out depending on the value of the discontinuity interval indicator ξ^α_1 as follows. If $-n^\phi_s \leqslant \xi^\alpha_1 \leqslant 0$, then a smooth branch of the initial function (or smooth extension thereof) is used for computing past states, i.e.

$$v^j_{l+1} = \phi_{\xi^\alpha_1}(t^j_{l+1} - \tau_1(t^j_{l+1}, y^j_{l+1}, c)). \tag{6.56}$$

Computation of Past States from a Past Discontinuity Interval

If $1 \leqslant \xi^\alpha_1 \leqslant n_s$, let $\underline{l}$ and $\bar{l}$ be the indices such that $s_{\xi^\alpha_1 - 1} = t_{\underline{l}}$ and $s_{\xi^\alpha_1} = t_{\bar{l}}$, i.e. $\underline{l}$ and $\bar{l}$ are the first and the last mesh point within the past discontinuity interval. The computation of past states is

then done as follows:

$$\begin{aligned} v_{l+1}^j &= \eta_{l'+1,p}(t_{l'} + \theta_{l,j} h_{l'+1}) \\ &= y_{l'} + h_{l'+1} \sum_{i=1}^{\nu} b_i(\theta_{l,j}) g_{l'+1}^i + h_{l'+1} b_*(\theta_{l,j}) g_{l'+1}^*. \end{aligned} \tag{6.57}$$

The expression in the right hand side is the higher order continuous approximation on the interval $[t_{l'}, t_{l'+1}]$. The index l' is thereby given as

(i) $l' = \underline{l}$ if it holds that $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) < t_{\underline{l}}$ (i.e. extrapolation to the left),

(ii) $\underline{l} \leqslant l' \leqslant \bar{l} - 1$ if it holds $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) \in [t_{\underline{l}}, t_{\bar{l}}]$, and l' is the index such that $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) \in [t_{l'}, t_{l'+1}]$,

(iii) $l' = \bar{l} - 1$ if it holds that $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) > t_{\bar{l}}$ (i.e. extrapolation to the right).

It is clear that the index l' depends both on the step l and on the stage j, i.e. $l'(l, j)$ would be a more accurate notation. However, for the sake of brevity, these dependencies are usually not written explicitly.

The evaluation point $\theta_{l,j}$ in equation (6.57) is the relative position of the past time point in the interval $[t_{l'}, t_{l'+1}]$:

$$\theta_{l,j} = \frac{t_{l+1}^j - \tau_1(t_{l+1}^j,\ y_{l+1}^j,\ c) - t_{l'}}{h_{l'+1}}. \tag{6.58}$$

Whenever the case (ii) occurs, i.e. the past time point is located within the discontinuity interval indicated by ξ_1^α, then it holds that $\theta_{l,j} \in [0, 1]$. However, due to the use of extrapolations in the modified standard approach, it may happen for a trial stepsize that $\theta_{l,j}$ is considerably smaller than 0 or considerably larger than 1 (cases (i) ans (iii)). However, the mechanisms for including discontinuity points in the mesh that are implemented in Colsol-DDE, presented in Section 6.9, ensure that for the eventually accepted stepsizes it holds that $0 \lesssim \theta_{l,j} \lesssim 1$.

Computation of Past States from the Current Discontinuity Interval

Eventually, consider the case that $\xi_1^\alpha = n_s + 1$, i.e. the past states have to be obtained from the current discontinuity interval whose left border is the currently last detected discontinuity point s_{n_s}. Let $\underline{l}$ be such that $t_{\underline{l}} = s_{n_s}$. Then the computation of past states is done in the following way:

$$v_{l+1}^j = y_{l'} + h_{l'+1} \sum_{i=1}^{\nu} b_i(\theta_{l,j}) g_{l'+1}^i + h_{l'+1} \Theta_{l,j} b_*(\theta_{l,j}) g_{l'+1}^*. \tag{6.59}$$

Herein, the index l' is given as follows:

(i) $l' = \underline{l}$ if it holds that $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) < t_{\underline{l}}$ (i.e. extrapolation to the left),

(ii) $\underline{l} \leqslant l' \leqslant \bar{l} - 1$ if it holds that $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) \in [t_{\underline{l}}, t_l]$, and l' is the index such that $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) \in [t_{l'}, t_{l'+1}]$,

(iii) $l' = l$ if it holds that $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) > t_l$, i.e. if overlapping occurs.

The evaluation point $\theta_{l,j}$ in equation (6.59) is given by equation (6.58), i.e. the expression is formally the same as for the computation of past states from past discontinuity intervals. The additional factor $\Theta_{l,j}$ is 1 if $l' < l$ (non-overlapping case), and it is 0 if $l' = l$ (overlapping case). This means that in the overlapping case, the lower order approximation $\eta_{l+1,q}(t_l + \theta h_{l+1})$ on the current interval is used, because the higher order approximation is not yet available.

Importantly, the uniform local order q of any of the three employed collocation methods is only one below the corresponding discrete local order p. Hence, in the view of Theorem 5.21, this is still sufficient so that the discrete and uniform local errors are $\mathscr{O}(h_{l+1}^{p'+1})$ and $\mathscr{O}(h_{l+1}^{q'+1})$ with $p' = p$ and $q' = q$, respectively.

It is worth mentioning that even in the case of overlapping the stage values v_{l+1}^j in equation (6.53) can, like y_{l+1}^j, be eliminated and expressed in terms of g_{l+1}^i, $1 \leqslant i \leqslant \nu$. Thus, the equation system can be formulated in such a way that only the quantities g_{l+1}^j occur as unknowns. Contrariwise, it is not possible to formulate, in the overlapping case, an equation system in which only y_{l+1}^j, $1 \leqslant j \leqslant \nu$ occur as unknowns. This is the reason why the equation system is, in Colsol-DDE and other practical DDE-IVP solvers, usually formulated for g_{l+1}^j rather than for y_{l+1}^j.

6.4.2. Implicit Uniform Correction Procedure

For the three considered methods, the implicit uniform correction procedure consists in solving the following n_y-dimensional equation system in each step $t_l \to t_{l+1} = t_l + h_{l+1}$:

$$g_{l+1}^* = f(t_{l+1}^*, y_{l+1}^*, c, v_{l+1}^*) - \dot{\eta}_{l+1,q}(t_{l+1}^*). \tag{6.60}$$

Therein, $t_{l+1}^* = t_l + \theta^* h_{l+1}$ is the additional abscissa of the augmented CRK method, and

$$\begin{aligned} y_{l+1}^* =& \eta_{l+1,p}(t_{l+1}^*) \\ =& y_l + h_{l+1} \left(\sum_{i=1}^{\nu} b_i(\theta^*) g_{l+1}^i + b_*(\theta^*) g_{l+1}^* \right) \end{aligned} \tag{6.61}$$

is the higher order continuous approximation of the exact solution u_{l+1} of the local problem at t_{l+1}^*. Further, v_{l+1}^* is the approximation of the past state, whose computation depends on the discontinuity interval indicator ξ_1^α. If $-n_s^\phi \leqslant \xi_1^\alpha \leqslant 0$, i.e. if the indicator points to a smooth branch of the initial function, then

$$v_{l+1}^* = \phi_{\xi_1^\alpha}(t_{l+1}^* - \tau_1(t_{l+1}^*, y_{l+1}^*, c)). \tag{6.62}$$

If $1 \leqslant \xi_1^\alpha \leqslant n_s + 1$, i.e. if the indicator points to a discontinuity interval to the right of $t^{ini}(c)$, then v_{l+1}^* is given by:

$$\begin{aligned} v_{l+1}^* &= \eta_{l'+1,p}(t_{l'} + \theta_{l,*} h_{l'+1}) \\ &= y_{l'} + h_{l'+1} \sum_{i=1}^{\nu} b_i(\theta_{l,*}) g_{l'+1}^i + h_{l'+1} b_*(\theta_{l,*}) g_{l'+1}^*. \end{aligned} \tag{6.63}$$

In analogy to the collocation method, $l'+1$ denotes the index of the step $t_{l'} \to t_{l'+1}$ from which the continuous representation is used. If the deviating argument is located outside of the discontinuity interval indicated by ξ_1^α, extrapolations are used in the same way as discussed in Subsection 6.4.1. The symbol $\theta_{l,*}$ denotes the relative position of the value of the past time point on the interval $[t_{l'}, t_{l'+1}]$:

$$\theta_{l,*} = \frac{t_{l+1}^* - \tau_1(t_{l+1}^*,\ y_{l+1}^*,\ c) - t_{l'}}{h_{l'+1}}. \tag{6.64}$$

Due to the use of extrapolations, it may happen in trial steps that $\theta_{l,*} \notin [0,1]$, and only for eventually accepted stepsizes it holds that $0 \lesssim \theta_{l,*} \lesssim 1$.

Since the uniform correction procedure itself aims at the determination of g_{l+1}^*, no distinction needs to be made between the overlapping and the non-overlapping case. This means that the higher order approximation in equation (6.63) is also used if $l = l'$.

However, since the polynomial continuous representation $\eta_{l+1,p}(t_l + \theta h_{l+1})$ is given by

$$\eta_{l+1,p}(t_l + \theta h_{l+1}) = y_l + h_{l+1} \left(\sum_{i=1}^{\nu} b_i(\theta) g_{l+1}^i + b_*(\theta) g_{l+1}^* \right), \tag{6.65}$$

it still consists of the (unchanged) stage values g_{l+1}^i of the collocation method. If overlapping occurs, the computation of these stage values was done by the lower order continuous approximation. Nevertheless, in view of Theorem 5.21, this is sufficient so that a CRK method can reach uniform local order p in the current step. Accordingly, the implicit uniform correction procedure will indeed

provide a continuous representation of degree and order p also in the overlapping case.

6.4.3. Implicit Quadrature Rule

In the context of DDE-IVPs, the defining equation for $g^{\diamond}_{l+1}$, i.e. equation (6.47), is generalized to

$$g^{\diamond}_{l+1} = f(t_{l+1}, y_{l+1,p+1}, c, v^{\diamond}_{l+1}). \tag{6.66}$$

Herein, $y_{l+1,p+1}$ is given by

$$y_{l+1,p+1} = y_l + h_{l+1} \sum_{i=1}^{\mu-1} B_i f(\bar{t}^i_{l+1}, \bar{y}^i_{l+1}, c, \bar{v}^i_{l+1}) + h_{l+1} B_\mu g^{\diamond}_{l+1}, \tag{6.67}$$

where $\bar{t}^i_{l+1} = t_l + \Gamma_i h_{l+1}$ are the abscissae of the quadrature rule and $\bar{y}^i_{l+1}$ represents the evaluations of the higher order continuous representation at these abscissae, i.e.

$$\bar{y}^i_{l+1} := \eta_{l+1,p}(\bar{t}^i_{l+1}). \tag{6.68}$$

It remains to discuss the computation of the past states $v^{\diamond}_{l+1}$ and $\bar{v}^i_{l+1}$, $1 \leqslant i \leqslant \mu - 1$, which depends on the discontinuity interval indicator ξ^α_1. If $-n^\phi_s \leqslant \xi^\alpha_1 \leqslant 0$, i.e. if the indicator points to a smooth branch of the initial function, then

$$v^{\diamond}_{l+1} = \phi_{\xi^\alpha_1}(t_{l+1} - \tau_1(t_{l+1}, y_{l+1,p+1}, c)) \tag{6.69a}$$

$$\bar{v}^i_{l+1} = \phi_{\xi^\alpha_1}(\bar{t}^i_{l+1} - \tau_1(\bar{t}^i_{l+1}, \bar{y}^i_{l+1}, c)). \tag{6.69b}$$

If $1 \leqslant \xi^\alpha_1 \leqslant n_s + 1$, i.e. the indicator points to a discontinuity interval to the right of $t^{ini}(c)$, then

$$\begin{aligned} v^{\diamond}_{l+1} &= \eta_{l'+1,p}\left(\frac{t_{l+1} - \tau_1(t_{l+1}, y_{l+1,p+1}, c) - t_{l'}}{h_{l'+1}}\right) \\ &= y_{l'} + h_{l'+1} \sum_{i=1}^{\nu} b_i(\theta_{l,\diamond}) g^i_{l'+1} + b_*(\theta_{l,\diamond}) g^*_{l'+1} \end{aligned} \tag{6.70a}$$

$$\bar{v}^i_{l+1} = \eta_{l'+1,p}\left(\frac{\bar{t}^i_{l+1} - \tau_1(\bar{t}^i_{l+1}, \bar{y}^i_{l+1}, c) - t_{l'}}{h_{l'+1}}\right) \tag{6.70b}$$

Herein, $l' + 1$ is the index of the step $t_{l'} \to t_{l'+1}$ from which the continuous representation is used. Thereby l' is different for each stage i in equation (6.70b) and for $v^{\diamond}_{l+1}$, but for the sake of brevity these dependency is not given. The symbol $\theta_{l,\diamond}$ denotes the relative position on the interval $[t_{l'}, t_{l'+1}]$:

$$\theta_{l,\diamond} = \frac{t^{\diamond}_{l+1} - \tau_1(t^{\diamond}_{l+1},\ y_{l+1,p+1},\ c) - t_{l'}}{h_{l'+1}}. \tag{6.71}$$

Since extrapolations beyond past discontinuity points are used it may happen for trial steps that $\theta_{l,\diamond} \notin [0, 1]$.

Note that for state-dependent delays, the value $y_{l+1,p+1}$ is used for the evaluation of the delay function in equation (6.71).

Both $v^{\diamond}_{l+1}$ and $\bar{v}^i_{l+1}$ are computed from the higher order continuous representation regardless of whether or not overlapping occurs. This guarantees that the implicit quadrature rule provides an approximation of $u_{l+1}(t_{l+1})$ that is of discrete local order $p + 1$.

6.5. Practical Solution of Equation Systems

6.5.1. Formulation of the Equation Systems

The collocation method, the uniform correction procedure, and the implicit quadrature rule lead to three nonlinear equation systems. As a first step, all dependencies of the three equation systems on the respective unknowns are collected. For example, for the collocation method, the system to

be solved reads

$$F_{col}(g_{l+1}^1,\dots,g_{l+1}^\nu) := \begin{pmatrix} g_{l+1}^1 - f(t_{l+1}^1, y_{l+1}^1, c, v_{l+1}^1) \\ \vdots \\ g_{l+1}^\nu - f(t_{l+1}^\nu, y_{l+1}^\nu, c, v_{l+1}^\nu) \end{pmatrix} = 0, \tag{6.72}$$

where

$$y_{l+1}^j = y_l + h_{l+1}\sum_{i=1}^{\nu} a_{j,i} g_{l+1}^i \tag{6.73a}$$

$$v_{l+1}^j = y_{l'} + h_{l'+1}\sum_{i=1}^{\nu} b_i(\theta_{l,j}) g_{l'+1}^i + h_{l'+1} b_*(\theta_{l,j})\Theta_{l,j} g_{l'+1}^* \tag{6.73b}$$

$$\theta_{l,j} = \frac{t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) - t_{l'}}{h_{l'+1}}. \tag{6.73c}$$

It has thereby been assumed that $1 \leqslant \xi_1^\alpha \leqslant n_s + 1$, i.e. the discontinuity interval indicator points to a discontinuity interval to the right of the initial time $t^{ini}(c)$; the necessary modifications for the case $-n_s^\phi \leqslant \xi_1^\alpha \leqslant 0$ are obvious.

It is recalled that, in a strict notation, $l'+1 = l'(l,j)+1$, and $l'+1$ gives the index of the step $t_{l'} \to t_{l'+1}$ from which the continuous representation is used for the computation of the state at the time $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c)$. Further, $\Theta_{l,j} = 0$ if overlapping occurs (i.e. if $l' = l'(l,j) = l$) and $\Theta_{l,j} = 1$ otherwise.

For the implicit uniform correction procedure, the system to be solved is

$$F_{ucp}(g_{l+1}^*) := g_{l+1}^* - f\left(t_{l+1}^*,\ y_{l+1}^*,\ c,\ v_{l+1}^*\right) - \dot\eta_{l+1,q}(t_l + \theta^* h_{l+1}) = 0. \tag{6.74}$$

Thereby,

$$y_{l+1}^* = y_l + h_{l+1}\sum_{i=1}^{\nu} b_i(\theta^*) g_{l+1}^i + h_{l+1} b_*(\theta^*) g_{l+1}^* \tag{6.75a}$$

$$v_{l+1}^* = y_{l'} + h_{l'+1}\sum_{i=1}^{\nu} b_i(\theta_{l,*}) g_{l'+1}^i + h_{l'+1} b_*(\theta_{l,*}) g_{l'+1}^* \tag{6.75b}$$

$$\theta_{l,*} = \frac{t_{l+1}^* - \tau_1(t_{l+1}^*, y_{l+1}^*, c) - t_{l'}}{h_{l'+1}}. \tag{6.75c}$$

In the context of the implicit uniform correction procedure, $l'+1$ denotes the index of the step $t_{l'} \to t_{l'+1}$ from which the continuous representation is used for the computation of the state at the time $t_{l+1}^* - \tau_1(t_{l+1}^*, y_{l+1}^*, c)$.

Eventually, consider the implicit quadrature rule. Here, the following system has to be solved:

$$F_{iqr}(g_{l+1}^\diamond) := g_{l+1}^\diamond - f(t_{l+1},\ y_{l+1,p+1},\ c, v_{l+1}^\diamond) = 0. \tag{6.76}$$

Thereby,

$$y_{l+1,p+1} = y_l + h_{l+1}\sum_{i=1}^{\mu-1} B_i f(\bar t_{l+1}^i, \eta_{l+1,p}(\bar t_{l+1}^i), c, \bar v_{l+1}^i) + h_{l+1} B_\mu g_{l+1}^\diamond \tag{6.77a}$$

$$v_{l+1}^\diamond = y_{l'} + h_{l'+1}\sum_{i=1}^{\nu} b_i(\theta_{l,\diamond}) g_{l'+1}^i + h_{l'+1} b_*(\theta_{l,\diamond}) g_{l'+1}^* \tag{6.77b}$$

$$\theta_{l,\diamond} = \frac{t_{l+1}^\diamond - \tau_1(t_{l+1}^\diamond,\ y_{l+1,p+1},\ c) - t_{l'}}{h_{l'+1}} \tag{6.77c}$$

$$\bar v_{l+1}^i = \eta_{l'+1,p}(\bar t_{l+1}^i - \tau_1(\bar t_{l+1}^i, \eta_{l+1,p}(\bar t_{l+1}^i), c)) \tag{6.77d}$$

Therein $l'+1$ in the second equation indicates the index of the interval which is used for the computation of the state at the past time point $t_{l+1} - \tau_1(t_{l+1}, y_{l+1,p+1}, c)$. In the third line $l'+1$ depends on i and denotes the index of the interval which is used for the computation of the states

at the past time points $\bar{t}^i_{l+1} - \tau_1(\bar{t}^i_{l+1}, \eta_{l+1,p}(\bar{t}^i_{l+1}), c)$. In both cases, extrapolations beyond past discontinuity points are used if the past time points are located outside of the discontinuity interval indicated by ξ^α_1.

6.5.2. Newton's Method

By inserting the equations into each other (e.g., equations (6.73) into equation (6.72) in case of the collocation method), the systems can shortly be expressed as

$$F(x) = 0 \tag{6.78}$$

for all three cases. Thereby, F represents the nonlinear functions F_{col}, F_{ucp}, and F_{iqr} for the collocation method, the implicit uniform correction procedure, and the implicit quadrature rule, and x represents the respective unknowns g^i_{l+1}, $1 \leqslant i \leqslant \nu$, g^*_{l+1}, and $g^\diamondsuit_{l+1}$.

For the solution of an equation system $F(x) = 0$, *Newton's method* or a *Newton-type method* can be used. For this purpose, denote the derivative of the function F with respect to x, the *Jacobian matrix*, by

$$\mathbf{J}(x) = \left.\frac{\partial F(x')}{\partial x'}\right|_{x'=x}. \tag{6.79}$$

Then *Newton's method* and a *Newton-type method* are defined as follows:

Definition 6.13 (Newton's Method, Newton-Type Method)

Let x^0 be an initial guess *of the unknowns x. Then* Newton's Method *finds new iterates x^k, $k \geqslant 1$, by*

$$x^{k+1} = x^k + \Delta x^k, \tag{6.80}$$

where the increment Δx^k *is determined by*

$$\Delta x^k = -\mathbf{J}^{-1}(x^k)F(x^k). \tag{6.81}$$

If an approximation $\mathbf{M}(x)$ of the inverse of the Jacobian matrix $\mathbf{J}^{-1}(x)$ is employed for defining the increments, i.e.

$$\Delta x^k = -\mathbf{M}(x^k)F(x^k), \tag{6.82}$$

then the method is called a Newton-type method*.*

Under certain conditions, Newton's Method and Newton-type methods converge to a solution of the equation system.

Theorem 6.14 (Local Contraction (of Newton-Type Methods))

Let $F(x)$ be a continuously differentiable function i.e. $F \in \mathscr{C}^1(\mathcal{D}^x, \mathbb{R}^{n_x})$, where $\mathcal{D}^x \subset \mathbb{R}^{n_x}$. Consider $y \in \mathcal{D}^x$, $z \in \mathcal{D}^x$, $z = y + \Delta y$, where $\Delta y = -\mathbf{M}(y)F(y)$ is the increment of a Newton-type method. Let further $\tilde{z} \in \mathcal{D}^x$. Assume that the following conditions are fulfilled for all y, z, $\tilde{z}$, and $\vartheta \in [0,1]$:

1. $\|\mathbf{M}(z)\left(\mathbf{J}(y + \vartheta\Delta y) - \mathbf{J}(y)\right)\Delta y\| \leqslant \omega\vartheta\|\Delta y\|^2$ *with* $\omega < \infty$,
2. $\|\mathbf{M}(\tilde{z})R(y)\| \leqslant \kappa\|\tilde{z} - y\|$, *where* $R(y) := F(y) + \mathbf{J}(y)\Delta y$ *and* $\kappa < 1$.

Let $x^0 \in \mathcal{D}^x$ be an initial guess such that

3. $\delta_0 := \kappa + \frac{\omega}{2}\|\Delta x^0\| < 1$, *where* $\|\Delta x^0\| = \|\mathbf{M}(x^0)F(x^0)\|$,
4. *the ball centered at x^0 defined by* $\mathcal{B}_{x_0} := \left\{x \mid \|x - x_0\| \leqslant \frac{\|\Delta x_0\|}{1-\delta_0}\right\}$ *is contained in $\mathcal{D}^x$.*

Then it holds that

(I) the iterates x^k are within $\mathcal{B}_{x_0}$,

(II) there exists $x^\star \in \mathcal{B}_{x_0}$ such that $x^k \to x^\star$ and $\|\Delta x^k\| \to 0$ for $k \to \infty$,

(III) $\|\Delta x^{k+1}\| \leqslant \delta_k \|\Delta x^k\|$, *with* $\delta_k := \kappa + \frac{\omega}{2}\|\Delta x^k\|$.

(IV) $\|x^k - x^\star\| \leqslant \delta_k \frac{\|\Delta x^k\|}{1-\delta_k}$.

Further, if $\|\mathbf{M}^{-1}(x)\| \leqslant M < \infty$ *for all* $x \in \mathcal{D}^x$, *then*

(V) $F(x^\star) = 0$.

Proof
See Bock [39] for the proof of a theorem that establishes local convergence in the more general setting of constrained least-squares problems. The proof for the theorem given here follows as a special case. ■

In practice, convergence is assumed if a *termination criterion* is fulfilled.

Note that $R(y) = F(y) + \mathbf{J}(y)\Delta y = F(y) - \mathbf{J}(y)\mathbf{M}(y)F(y) = 0$, if $\mathbf{M}(y)$ is the exact inverse of $\mathbf{J}(y)$. Hence, it follows that $\kappa = 0$ for the exact Newton method, and consequently the convergence is quadratic because $\|\Delta x^{k+1}\| \leqslant \frac{\omega}{2}\|\Delta x^k\|^2$.

By using an approximation instead of the exact inverse, the convergence is only linear and typically more iterations are needed until the termination criterion is fulfilled. Nevertheless, the resulting algorithm is often more efficient with respect to floating point operations and runtimes, because the Jacobian does not need to be computed and numerically inverted in each iteration.

In the following, the details of a Newton-type method as it is implemented in Colsol-DDE are presented. In particular, the following issues are addressed: the generation of initial guesses, the iterative solution with approximate inverses, the recomputation strategies for the inverse of the Jacobian, the structure of the exact Jacobian matrices, and the termination criterion.

6.5.3. Initial Guesses

This subsection is devoted to the question how to choose the initial guess x^0 for Newton's method.

Collocation Method

For the collocation method, initial guesses are needed for g^j_{l+1}, $1 \leqslant j \leqslant \nu$. These initial guesses are called $(g^j_{l+1})^0$ and are obtained by

$$(g^j_{l+1})^0 = f(t^j_{l+1}, (y^j_{l+1})^0, c, (v^j_{l+1})^0). \tag{6.83}$$

Herein, $(y^j_{l+1})^0$ and $(v^j_{l+1})^0$ are initial approximations of the y^j_{l+1} and v^j_{l+1}.

The initial guesses $(y^j_{l+1})^0$ are, in most integration steps, obtained from

$$(y^j_{l+1})^0 = \eta_{l,p}(t^j_{l+1}), \tag{6.84}$$

that is from an extrapolation of the higher order continuous representation from the previous step to the abscissae of the current step. Exceptions from this initialization strategy are made in the first integration step $t_0 \to t_1$, and in those integration steps that follow a discontinuity for which the *practically determined order* is 0 or 1. In the former case, there exists no previous integration step l, and in the latter case extrapolation from the previous step is unsuitable because of the presence of a low order discontinuity. Therefore, in these cases, a linear approximation is used for $(y^j_{l+1})^0$:

$$(y^j_{l+1})^0 = y_l + c_j h_{l+1} f(t_l, y_l, c, v_l). \tag{6.85}$$

Which order is, in practice, attributed to a newly found discontinuity is discussed in Section 6.9.

The computation of the initial guesses $(v^j_{l+1})^0$ depends primarily on the value of the discontinuity interval indicator ξ^α_1. If $-n^\phi_s \leqslant \xi^\alpha_1 \leqslant 0$, then an evaluation of a smooth branch of the initial function is used. For $1 \leqslant \xi^\alpha_1 \leqslant n_s$, and if overlapping does not occur, $(v^j_{l+1})^0$ is computed from

$$(v^j_{l+1})^0 = \eta_{l'+1,p}(t^j_{l+1} - \tau_1(t^j_{l+1}, (y^j_{l+1})^0, c)), \tag{6.86}$$

i.e. from an evaluation of the higher order continuous representation in the "correct" step $t_{l'} \to t_{l'+1}$. If overlapping occurs, an extrapolation from the previous step is used ($l'+1=l$ in equation (6.86)). The computation of $(v_{l+1}^j)^0$ then resembles that for $(y_{l+1}^j)^0$ in equation (6.84).

Implicit Uniform Correction

For the implicit uniform correction, an initial guess is needed for g_{l+1}^*, which is denoted by $(g_{l+1}^*)^0$. This initial guess is given by

$$(g_{l+1}^*)^0 = f(t_{l+1}^*, (y_{l+1}^*)^0, c, (v_{l+1}^*)^0) - \dot{\eta}_{l+1,q}(t_{l+1}^*) \tag{6.87}$$

where $(y_{l+1}^*)^0$ and $(v_{l+1}^*)^0$ are the initial approximations of y_{l+1}^* and v_{l+1}^*.

The initial guess $(y_{l+1}^*)^0$ is obtained from

$$(y_{l+1}^*)^0 = \eta_{l+1,q}(t_{l+1}^*), \tag{6.88}$$

i.e. from an evaluation of the continuous representation implied by the collocation method in the current step.

The computation of the initial guess $(v_{l+1}^*)^0$ is determined by the discontinuity interval indicator ξ_1^α. If it holds that $-n_s^\phi \leqslant \xi_1^\alpha \leqslant 0$, then $(v_{l+1}^*)^0$ is obtained from an evaluation of a smooth branch of the initial function. If $1 \leqslant \xi_1^\alpha \leqslant n_s+1$, then $(v_{l+1}^*)^0$ is obtained, in the non-overlapping case, from

$$(v_{l+1}^*)^0 = \eta_{l'+1,p}(t_{l+1}^* - \tau_1(t_{l+1}^*, (y_{l+1}^*)^0, c)). \tag{6.89}$$

In the case of overlapping, $(v_{l+1}^*)^0$ cannot be obtained from the higher order continuous representation, which is yet to be determined. Instead, the same approach as for $(y_{l+1}^*)^0$ is used, i.e. the polynomial continuous representation of lower order (implied by the collocation method itself) is employed.

Implicit Quadrature Rule

For the implicit quadrature rule, an initial guess $(g_{l+1}^\diamondsuit)^0$ is needed for $g_{l+1}^\diamondsuit$, which is obtained from

$$(g_{l+1}^\diamondsuit)^0 = f(t_{l+1},\ (y_{l+1,p+1})^0,\ c, (v_{l+1}^\diamondsuit)^0). \tag{6.90}$$

Colsol-DDE uses the discrete approximation of the collocation method for the initialization of $y_{l+1,p+1}$, i.e.

$$(y_{l+1,p+1})^0 = y_{l+1,p}. \tag{6.91}$$

Further, $(v_{l+1}^\diamondsuit)^0$ is an initial guess for $v_{l+1}^\diamondsuit$. If the discontinuity interval indicator ξ_1^α is less than or equal to 0, then this initial guess is obtained from an evaluation of a smooth branch of the initial function. Otherwise it is computed from

$$(v_{l+1}^\diamondsuit)^0 = \eta_{l'+1,p}(t_{l+1} - \tau_1(t_{l+1}, (y_{l+1,p+1})^0, c)), \tag{6.92}$$

i.e. from an evaluation of the higher order continuous representation. The overlapping case ($l'=l$) does not need to be treated differently, because the uniform correction procedure has been applied before and thus $\eta_{l+1,p}$ is already available.

6.5.4. Structure of the Exact Jacobian Matrices

Colsol-DDE uses, for the solution of the equation systems, approximate inverses of the Jacobians of the three equation systems (6.72), (6.74), (6.76). This subsection is concerned with the structure of these Jacobian matrices.

Collocation Method

The equation system (6.72) has dimension $\nu \times n_y$, where, as usual, ν is the number of stages of the collocation method and n_y is the dimension of the state vector y. For the Lobatto method, the

first stage is explicit, therefore the dimension reduces to $(\nu - 1) \times n_y$. For notational convenience, define the possibly reduced number of stages such that $\nu_r := 2$ for the two-stage Radau IIA collocation method and the three-stage Lobatto IIIA collocation method, and such that $\nu_r := 1$ for the one-stage Gauss collocation method.

When taking the derivative of the function F_{col} in equation (6.72) with respect to g_{l+1}^j, derivatives of the right-hand-side function f and the delay function τ_1 occur. In Colsol-DDE, the Jacobian can be recomputed once per integration step before the iterations are started. Accordingly, the partial derivatives of f are evaluated at the initial guesses of the unknowns, i.e. $(t_{l+1}^j, (y_{l+1}^j)^0, c, (v_{l+1}^j)^0)$. Therefore, define

$$\left(\frac{\partial f}{\partial y}\right)_{l+1}^{j} := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(t_{l+1}^j,(y_{l+1}^j)^0,c,(v_{l+1}^j)^0)} \tag{6.93a}$$

$$\left(\frac{\partial f}{\partial v}\right)_{l+1}^{j} := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(t_{l+1}^j,(y_{l+1}^j)^0,c,(v_{l+1}^j)^0)} \tag{6.93b}$$

$$\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{j} := \left.\frac{\partial \tau_1(t,y,c)}{\partial y}\right|_{(t_{l+1}^j,(y_{l+1}^j)^0,c)}. \tag{6.93c}$$

The derivative of the function F_{col} in equation (6.72) with respect to its unknowns is a $(\nu_r \times n_y, \nu_r \times n_y)$ square matrix. With the notation introduced above, the (n_y, n_y)-dimensional block that represents the derivative of g_{l+1}^j with respect to g_{l+1}^k becomes

$$\begin{aligned}\frac{\partial g_{l+1}^j}{\partial g_{l+1}^k} =& \delta_{j,k}\mathbf{1}_{n_y,n_y} - \left(\frac{\partial f}{\partial y}\right)_{l+1}^{j} h_{l+1}a_{j,k} - \left(\frac{\partial f}{\partial v}\right)_{l+1}^{j} h_{l+1}b_k(\theta_{l,j})(1-\Theta_{l,j}) \\ &+ \left(\frac{\partial f}{\partial v}\right)_{l+1}^{j} h_{l+1}\left[\sum_{i=1}^{\nu} \dot{b}_i(\theta_{l,j})g_{l'+1}^j + \dot{b}_*(\theta_{l,j})\Theta_{l,j}g_{l'+1}^*\right]\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{j} a_{j,k}.\end{aligned} \tag{6.94}$$

In the first term, $\delta_{j,k}$ represents the Kronecker-δ, which is 1 for $j = k$ and 0 otherwise, and $\mathbf{1}_{n_y,n_y}$ represents the (n_y, n_y)-dimensional identity matrix. Both the first and the second term are present for ODEs. The third term is present if overlapping occurs. The last term accounts for state-dependencies of the delays.

Implicit Uniform Correction

The solution of the equation system (6.74) with a Newton-type method requires approximation of the corresponding Jacobian matrix, i.e. the derivative of F_{ucp} with respect to g_{l+1}^*. This Jacobian can be recomputed once per integration step before the iterations are started. In order to express the Jacobian in a compact form, the following definitions are introduced

$$\left(\frac{\partial f}{\partial y}\right)_{l+1}^{*} := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(t_{l+1}^*,(y_{l+1}^*)^0,c,(v_{l+1}^*)^0)} \tag{6.95a}$$

$$\left(\frac{\partial f}{\partial v}\right)_{l+1}^{*} := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(t_{l+1}^*,(y_{l+1}^*)^0,c,(v_{l+1}^*)^0)} \tag{6.95b}$$

$$\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{*} := \left.\frac{\partial \tau_1(t,y,c)}{\partial y}\right|_{(t_{l+1}^*,(y_{l+1}^*)^0,c)} \tag{6.95c}$$

With these definitions, the Jacobian is given by

$$\begin{aligned}\left.\frac{\partial F_{ucp}(x)}{\partial x}\right|_{x=(g_{l+1}^*)^0} =& \mathbf{1}_{n_y,n_y} - \left(\frac{\partial f}{\partial y}\right)_{l+1}^{*} h_{l+1}b_*(\theta^*) - \left(\frac{\partial f}{\partial v}\right)_{l+1}^{*} h_{l+1}b_*(\theta_{l,*})(1-\Theta_{l,*}) \\ &+ \left(\frac{\partial f}{\partial v}\right)_{l+1}^{*} h_{l+1}\left[\sum_{i=1}^{\nu} \dot{b}_i(\theta_{l,*})g_{l'+1}^j + \dot{b}_*(\theta_{l,*})g_{l'+1}^*\right]\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{*} b_*(\theta^*).\end{aligned} \tag{6.96}$$

Herein, $\Theta_{l,*} = 0$ in the overlapping case and $\Theta_{l,*} = 1$ otherwise.

Implicit Quadrature Rule

Consider the equation system (6.76) that needs to be solved for applying the implicit quadrature rule. The Jacobian of F_{iqr} with respect to the unknowns $g^{\diamondsuit}_{l+1}$ can be recomputed once per integration step before the iterations are started. Accordingly, define

$$\left(\frac{\partial f}{\partial y}\right)^{\diamondsuit}_{l+1} := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(t_{l+1},(y_{l+1,p+1})^0,c,(v^{\diamondsuit}_{l+1})^0)} \tag{6.97a}$$

$$\left(\frac{\partial f}{\partial v}\right)^{\diamondsuit}_{l+1} := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(t_{l+1},(y_{l+1,p+1})^0,c,(v^{\diamondsuit}_{l+1})^0)} \tag{6.97b}$$

$$\left(\frac{\partial \tau_1}{\partial y}\right)^{\diamondsuit}_{l+1} := \left.\frac{\partial \tau_1(t,y,c)}{\partial y}\right|_{(t_{l+1},(y_{l+1,p+1})^0,c)}. \tag{6.97c}$$

By this, the Jacobian of the equation system (6.76) can be written as:

$$\begin{aligned}\left.\frac{\partial F_{iqr}(x)}{\partial x}\right|_{x=(g^{\diamondsuit}_{l+1})^0} =& \mathbf{1}_{n_y,n_y} - \left(\frac{\partial f}{\partial y}\right)^{\diamondsuit}_{l+1} h_{l+1} B_\mu \\ &+ \left(\frac{\partial f}{\partial v}\right)^{\diamondsuit}_{l+1} h_{l+1} \left[\sum_{i=1}^{\nu} \dot{b}_i(\theta_{l,\diamondsuit}) g^j_{l'+1} + \dot{b}_*(\theta_{l,\diamondsuit}) g^*_{l'+1}\right] \left(\frac{\partial \tau_1}{\partial y}\right)^{\diamondsuit}_{l+1} B_\mu. \end{aligned} \tag{6.98}$$

6.5.5. Decomposition of the Jacobian Matrix

For the computation of the increment Δx^k in Newton's method by equation (6.81), the inverse of the Jacobian is needed. One option for numerically inverting a matrix is based on the computation of a *singular value decomposition* of the Jacobian:

$$\mathbf{J}(x) = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \tag{6.99}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{\Sigma}$ is a diagonal matrix that contains the singular values of $\mathbf{J}$, which are denoted by σ_i, $1 \leqslant i \leqslant n_x$ of $\mathbf{J}(x)$:

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & \sigma_{n_x} \end{pmatrix}. \tag{6.100}$$

If such a singular value decomposition is available, the inverse of $\mathbf{J}(x)$ can easily be computed from $\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T$, if $\sigma_i > 0$ for all $1 \leqslant i \leqslant n_x$.

It is remarked that the Jacobian matrices given in equations (6.94), (6.96), and (6.98) converge, for $h_{l+1} \to 0$, to the identity matrix. Hence, for sufficiently small stepsize h_{l+1}, the matrix $\mathbf{J}(x)$ is always well-conditioned. In practice, when a new Jacobian is computed and decomposed, Colsol-DDE checks that

$$\frac{\sigma_1}{\sigma_{n_x}} \leqslant \kappa_{max}, \tag{6.101}$$

where κ_{max} is a user-given bound on the condition number of the Jacobian matrix that is implied by the spectral norm. If the ratio σ_1/σ_{n_x} exceeds this bound, the stepsize is reduced.

The Jacobian matrix is a quantity that depends on the scaling that the user chooses for the variables of the problem, e.g. whether a position is measured in meters or kilometers. More precisely, the element (i,j) of the Jacobian matrix is affected by $[x_i]/[x_j]$, where $[\,\cdot\,]$ represents the user-chosen unit of the variable.

The singular values and thus the condition of the Jacobian (in the spectral norm) may depend heavily on the user-chosen scaling. In order to make the decomposition of the matrix independent of a probably inappropriate user-scaling, Colsol-DDE uses scaling factors s_{xi}, $1 \leqslant i \leqslant n_x$, for the

unknowns x_i. How these scaling factors are obtained is the subject of Subsection 6.5.8. At this point, it is simply assumed that the scaling factors roughly represent the "typical" magnitude of x_i, i.e. it should hold that $s_{xi} \approx x_i$.

Assuming that appropriate scaling factors are available, the following scaling matrix is defined:

$$\mathbf{S} := \begin{pmatrix} s_{x1} & 0 & \dots & 0 \\ 0 & s_{x2} & & 0 \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & s_{xn} \end{pmatrix}. \tag{6.102}$$

It turns out that the result of the singular value decomposition of $\mathbf{S}^{-1}\mathbf{J}(x)\mathbf{S}$ (instead of $\mathbf{J}(x)$) is (almost) independent of the user scaling. Hence, Colsol-DDE practically uses the decomposition

$$\mathbf{S}^{-1}\mathbf{J}(x)\mathbf{S} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T \tag{6.103}$$

instead of equation (6.99).

6.5.6. Iterative Solution

According to the decomposition of the Jacobian matrix given in equation (6.103), the increment Δx is determined by

$$\Delta x^k = -\mathbf{S}\mathbf{V}\mathbf{\Sigma}^{-1}\mathbf{U}^T\mathbf{S}^{-1}F(x^k). \tag{6.104}$$

The user may specify two integer numbers n^1_{itmax} and n^2_{itmax} such that $1 \leqslant n^1_{itmax} < n^2_{itmax}$. If more than n^1_{itmax} iterations are needed in the current integration step $t_l \to t_{l+1}$, than a recomputation and decomposition of an exact Jacobian matrix is scheduled for the next integration step $t_{l+1} \to t_{l+2}$. If the number of iterations exceeds n^2_{itmax}, then the iterations of Newton's method are stopped and recomputation and decomposition of an exact Jacobian matrix is done in the current integration step. After that, Newton's method is restarted with the initial guess x^0. If convergence can still not be obtained within n^2_{itmax} iteration steps, the stepsize h_{l+1} is reduced.

The above-described mechanism for the iterative solution and Jacobian matrix recomputation is applied for the collocation method, the implicit uniform correction, and the implicit quadrature rule.

6.5.7. Termination Criterion

Colsol-DDE uses a component-dependent termination criterion. This means that the system is considered to be solved successfully if, for every component of the increment, Δx_i, $1 \leqslant i \leqslant n_x$, it holds that

$$\frac{|\Delta x_i|}{s_{x_i}} \leqslant (\delta_{term})_i. \tag{6.105}$$

The components of δ_{term} are computed in such a way that, if a user-given parameter η_{term} is chosen as 10^{-m}, then approximately m digits are valid in the quantity that is used for error estimation. The precise formula for δ_{term} is therefore given in Section 6.6 after the quantities for error estimation have been introduced.

6.5.8. Scaling Factors

The decomposition of the Jacobian matrices for the numerical solution of the equation systems (6.72), (6.74), and (6.76), depends on scaling factors s_{xi} for the unknowns. Moreover, also the termination criterion (6.105) depends on the scaling factors. Since the unknowns are g^j_{l+1}, g^*_{l+1}, and $g^\diamond_{l+1}$, respectively, scaling factors s_{g_i} are needed that represent the "typical" order of magnitude of the i-th component of a right-hand-side function evaluation.

In a different context, see Section 9.1.7, scaling factors s_{y_i} are needed that represent the "typical" order of magnitude of the i-th component of the state vector.

In the following, the heuristic is described that Colsol-DDE uses for computing scaling factors. This is done for the computation of s_{y_i}, but the techniques for choosing the scaling factors s_{g_i} are completely analogous.

Consider the task of finding a scaling factor $(s_y)_i$ for y_i such that $y_i \approx (s_y)_i$, i.e. $(s_y)_i$ should represent the "typical" order of magnitude of y_i. The main issue in doing this is how the scaling factors should be chosen if a variable is zero or close to zero.

A possible strategy for the initialization of the scaling variable at the initial time, where only an initial value y_0 is available, is as follows:

$$(s_y)_i = \begin{cases} |(y_0)_i| & \text{if} \quad (y_0)_i > \epsilon_{thresh} \\ 1 & \text{else} \end{cases} . \tag{6.106}$$

Herein, ϵ_{thresh} is a user-given threshold value. If the initial value of a state vector component is below this value, it is considered to be practically zero and the scaling factor, due to a lack of information, is chosen as 1.

However, it happens frequently that the initial guess is 0 but that the state becomes non-zero in the first integration step. Hence, doing a single integration step would be sufficient to get a better approximation of the "typical" order of magnitude of y_i. For this reason, Colsol-DDE performs one step with the classical explicit 4-stage Runge-Kutta method (which does not need scaling factors) and uses, componentwise, the maximum value over the 4 stages to initialize the scaling factor.

Once the scaling factor is initialized, it is updated after each successful integration step. More precisely, the scaling factor in the step $l+1$, called $(s_{y_{l+1}})_i$, is obtained from the old scaling factor in step l, called $(s_{y_l})_i$, by the following rule:

$$(s_{y_{l+1}})_i = (1-\alpha_{mem})s_{new} + \alpha_{mem}(s_{y_l})_i \tag{6.107a}$$

$$s_{new} = \max(|(y_l)_i|, |(y_{l+1})_i|, \epsilon_{thresh}). \tag{6.107b}$$

Herein, $\alpha_{mem} \in [0,1]$ is a user-defined "memory factor" that gives the relative importance of the old scaling factor for the computation of the new scaling factor. For smaller values of α_{mem}, the scaling factor is quickly adapted if the corresponding solution component varies significantly, whereas for a larger value of α_{mem} the information gathered on the "typical" order of magnitude is changed only slightly in each integration step.

The updating rule (6.107) is also applied in discontinuities (see Section 6.9), e.g. if a jump is applied in a root discontinuity, then the state after a jump is also used for updating the scaling factor.

For the sake of completeness it is mentioned that for the increment computation (6.104) those scaling factors are used that were used for the decomposition of the matrix. Contrariwise, in the termination criterion (6.105) the scaling factors of the current integration step are used.

6.6. Error Control

This section discusses the error control mechanism that is realized in Colsol-DDE.

6.6.1. Advancing and Error-Estimating Method

In Section 5.5 it was discussed that error control strategies are typically based on two discrete and two continuous approximations, for which the discrete and uniform local errors have different orders. This is also the case in Colsol-DDE. More specifically, two discrete approximations $y_{l+1,p}$ and $y_{l+1,p+1}$ or orders p and $p+1$, and two continuous approximations $\eta_{l+1,q}$ and $\eta_{l+1,q+1}$ of orders q and $q+1$, are available. Thereby, (p,q) denote the discrete and uniform local orders of the basic collocation method.

At first, it needs to be specified, which discrete and which continuous approximation is used for advancing the step. In Colsol-DDE, the lower order discrete approximation obtained with the collocation method and the higher order continuous approximation obtained by the implicit uniform correction are used, i.e.

$$y_{l+1} := y_{l+1,p}, \quad \eta_{l+1}(t) := \eta_{l+1,q+1}(t). \tag{6.108}$$

Accordingly, $y_{l+1,p}$ and $\eta_{l+1,q+1}(t)$ are the *advancing methods*, whereas $y_{l+1,p+1}$ and $\eta_{l+1,q}(t)$ are the *error-estimating methods*. Once a step is accepted, the latter results are no longer needed for the further application of the method.

It is mentioned that, consequently, the advancing methods in Colsol-DDE are a realization of the "augmented CRK methods" whose Butcher tableaus and continuous weight functions were given in Section 6.2.

6.6.2. Error Estimation

According to the findings in Section 5.5, the maximum difference between the lower and the higher order continuous representation can be used to estimate the uniform local error:

$$\hat{\tilde{\delta}}_{l+1} = \max_{t_l \leqslant t \leqslant t_{l+1}} \|\eta_{l+1,q}(t) - \eta_{l+1,q+1}(t)\|. \tag{6.109}$$

For all three methods that are implemented in Colsol-DDE, this difference between the two continuous representations is simply given by

$$\hat{\tilde{\delta}}_{l+1} = h_{l+1}\|g^*_{l+1}\| \underbrace{\max_{0 \leqslant \theta \leqslant 1} |b_*(\theta)|}_{=:b_{*,max}}. \tag{6.110}$$

Since $b_*(\theta)$ is a problem-independent polynomial function, the maximum value of $|b_*(\theta)|$ is assumed at the same value θ_{max} on all integration intervals regardless of the specific right-hand-side function f. More precisely, it holds that $\theta_{max} = 1/2$ and $b_{*,max} = 3/4$ for the one-stage Gauss collocation method, $\theta_{max} = 1/3$ and $b_{*,max} = 16/45$ for the two-stage Radau IIA collocation method, and $\theta_{max} = 1/2$ and $b_{*,max} = 1/3$ for the three-stage Lobatto IIIA collocation method.

For estimating the discrete local error, the difference between the result of the collocation method, $y_{l+1,p}$, and the result of the implicit quadrature rule, $y_{l+1,p+1}$, can be used (cf. Zennaro [268]):

$$\hat{\delta}_{l+1} = \|y_{l+1,p+1} - y_{l+1,p}\|. \tag{6.111}$$

6.6.3. Error Control

As discussed before, the lower order discrete approximation and the higher order continuous representation are used as advancing methods. Accordingly, the uniform local error estimation employs local extrapolation, whereas the discrete local error estimation does not.

According to the discussion in Section 5.5, proportionality of the global error to a user-defined local tolerance σ_{tol} is obtained if the following conditions are fulfilled by the error estimates in every integration step: $\hat{\tilde{\delta}}_{l+1} \leqslant \sigma_{tol}(t^{fin}(c) - t^{ini}(c))/h_{l+1}$ and $\hat{\delta}_{l+1} \leqslant \sigma_{tol}h_{l+1}/(t^{fin}(c) - t^{ini}(c))$ (and if the implied strategy for suggesting new stepsizes is used).

It remains to choose a specific norm in $\mathbb{R}^{n_y}$ and to deal with the fact that the individual components of the state vector may, in practice, have very different orders of magnitude. Colsol-DDE therefore uses the following variation of the two conditions (cf. Hairer, Nørsett, and Wanner [126], page 167f):

$$C_{disc} := \max_{1 \leqslant i \leqslant n_y} \left(\frac{|(y_{l+1,p+1})_i - (y_{l+1,p})_i|}{\epsilon_i} \right) \cdot \frac{(t^{fin}(c) - t^{ini}(c))}{h_{l+1}} \leqslant 1 \tag{6.112a}$$

$$C_{unif} := h_{l+1}b_{*,max} \max_{1 \leqslant i \leqslant n_y} \left(\frac{|(g^*_{l+1})_i|}{\epsilon_i} \right) \cdot \frac{h_{l+1}}{(t^{fin}(c) - t^{ini}(c))} \leqslant 1. \tag{6.112b}$$

where

$$\epsilon_i = \max(|(y_l)_i|, |(y_{l+1,p})_i|) \cdot \sigma^{rel}_{tol} + \sigma^{abs}_{tol}. \tag{6.113}$$

Both the *absolute tolerance* σ^{abs}_{tol} and the *relative tolerance* σ^{rel}_{tol} are input parameters that are specified by the user.

In accordance with the chosen error criteria, Colsol-DDE uses the following formula for suggesting

a new stepsize h_{l+2}:

$$h_{l+2} = \rho_{safe} h_{l+1} \min\left(\sqrt[p]{\frac{1}{C_{disc}}},\ \sqrt[q+2]{\frac{1}{C_{unif}}} \right), \tag{6.114}$$

where ρ_{safe} is a user-given safety factor.

6.6.4. Termination Criterion for Newton's Method

With the practically used error conditions at hand, the discussion of the termination criterion for Newton's method can be resumed (recall Subsection 6.5.7). For reasons of efficiency, it is reasonable to stop the Newton iterations when C_{disc} and C_{unif} have, say, two valid digits. For example, for the uniform error estimation, this motivates to determine the termination criterion $(\delta_{term})_i$ for $(g^*_{l+1})_i$ such that

$$(\delta_{term})_i = \left(10^{-2} \frac{(t^{fin}(c) - t^{ini}(c))}{h_{l+1}} \epsilon_i \right) \cdot \frac{1}{h_{l+1} s_{g_i}}. \tag{6.115}$$

Herein, s_{g_i} is the scaling factor for the i-th component of an evaluation of the right-hand-side function f. For a given value of the term in brackets, the termination criterion is smaller (i.e. higher relative accuracy) if the "typical" absolute value of the corresponding component of the right-hand-side function – represented by the scaling factor – is larger.

For the discrete error estimation, i.e. for C_{disc}, the motivation is analogous, because the difference $(y_{l+1,p+1})_i - (y_{l+1,p})_i$ is also a weighted sum of right-hand-side function evaluations multiplied by h_{l+1}. This gives the following termination criterion for g^j_{l+1} and $g^\diamond_{l+1}$, respectively:

$$(\delta_{term})_i = \left(10^{-2} \frac{h_{l+1}}{(t^{fin}(c) - t^{ini}(c))} \epsilon_i \right) \cdot \frac{1}{h_{l+1} s_{g_i}}. \tag{6.116}$$

In Colsol-DDE, the factor 10^{-2} is replaced by some user-given input parameter η_{term}, in order to give the user the opportunity to choose a different termination criterion.

6.7. Basic Stability Properties

For initial value problems, not necessarily in the context of DDEs, it is well-known from practical experience that some methods, mostly explicit methods, become very inefficient when the solution evolves on very different time scales, e.g. a slow oscillation in one state vector component and a rapid transient to an equilibrium in another state vector component. In order to analyze such a behavior, stability concepts are used. In this section, some elementary stability concepts are recalled and it is discussed whether the methods implemented in Colsol-DDE are stable with respect to these concepts.

6.7.1. A-Stability and L-Stability for Discrete Runge-Kutta Methods

Theoretical Background

An elementary stability concept is related to the following simple test equation:

$$\dot{\mathfrak{y}}(t) = \lambda \mathfrak{y}(t) \tag{6.117a}$$

$$\mathfrak{y}(0) = y^{ini}. \tag{6.117b}$$

This is considered for $\lambda \in \mathbb{C}$, the real and imaginary parts of which are denoted by $\mathfrak{Re}(\lambda)$ and $\mathfrak{Im}(\lambda)$. The exact solution of the test problem is known to be $y(t) = y^{ini} \exp(\lambda t)$. Clearly, whenever $\mathfrak{Re}(\lambda) < 0$ then it holds that $|y(t)| < |y^{ini}|$ for $t \geqslant 0$ and $\lim_{t\to\infty} y(t) \to 0$. Moreover, for $\mathfrak{Re}(\lambda) \ll 0$, the test problem can be considered as a role model for rapidly decaying states, e.g. for the situation that some components of the state vector in a differential equation system show a rapid transient behavior.

A discrete Runge-Kutta method, when applied to the problem (6.117), yields in the first step

$$y_1 = y_0 + h \sum_{j=1}^{\nu} \beta_j \lambda y_1^j \tag{6.118a}$$

$$y_1^j = y_0 + h \sum_{k=1}^{\nu} a_{j,k} \lambda y_1^k, \tag{6.118b}$$

with $y_0 = y^{ini}$. This can be expressed in the compact form

$$y_1 = \left[1 + h\lambda\beta^T(\mathbf{1} - h\lambda\mathbf{A})^{-1}e\right] y_0. \tag{6.119}$$

where $\beta^T = (\beta_1, \ldots, \beta_\nu)$ is a vector containing the weights, $\mathbf{A}$ is a matrix containing the coefficients $a_{j,k}$, and $e = (1, 1, \ldots, 1)^T$ is a ν-dimensional vector. Further, $\mathbf{1}$ is the identity matrix of the appropriate dimension. If the absolute value of the term in square brackets is less than 1, then it holds that $|y_1| < |y_0|$, i.e. the numerical solution is decreasing. For constant stepsizes, and for variable stepsizes that are bounded, it holds in addition that the solution y_l vanishes asymptotically for $l \to \infty$. This leads to the following definition.

Definition 6.15 (Stability Function, Stability Domain)

The function

$$R(z) := 1 + z\beta^T(1 - z\mathbf{A})^{-1}e \tag{6.120}$$

is called the stability function *of a discrete Runge-Kutta method and the set*

$$S := \{z \in \mathbb{C} \mid |R(z)| < 1\} \tag{6.121}$$

is called the stability domain *of a discrete Runge-Kutta method.*

The well-known concept of *A-stability* demands that the stability domain of a numerical method should contain the half-plane $\mathbb{C}^- := \{z \in \mathbb{C} \mid \mathfrak{Re}(z) < 0\}$ so that the numerical solution is decreasing whenever the exact solution is decreasing.

Definition 6.16 (A-Stability)

A numerical method is called A-stable *if it holds that*

$$S \supset \mathbb{C}^-. \tag{6.122}$$

This property ensures that, if the exact solution is rapidly going to zero ($\mathfrak{Re}(\lambda) \ll 0$), the numerical method may use large stepsizes $h \gg 1$ and it will still possess the contractivity property $|y_1| < |y_0|$.

However, even if a method is A-stable, its contraction for $h\lambda \to -\infty$ might be very slow if it holds that $\lim_{h\lambda\to-\infty} |R(h\lambda)| = 1$, even though the exact solution approaches the zero rapidly for $\mathfrak{Re}(\lambda) \ll 0$. This leads to the following stronger stability concept.

Definition 6.17 (L-Stability)

A numerical method is called L-stable *if it is A-stable and, in addition, it holds that*

$$\lim_{z\to-\infty} |R(z)| = 0. \tag{6.123}$$

Properties of the Collocation Methods in Colsol-DDE

The stability functions R_G of the one-stage Gauss collocation method, R_R of the two-stage Radau IIA collocation method, and R_L of the three-stage Lobatto IIIA collocation method can be obtained by standard analysis. Using the definition $z = h\lambda$, the following expressions are obtained:

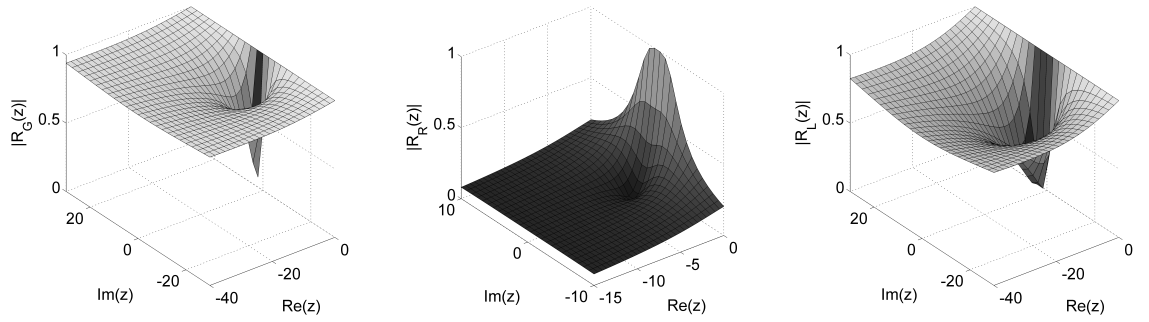
$$R_G(z) = 1 + \frac{-2z}{z - 2} \tag{6.124a}$$

$$R_R(z) = 1 + \frac{z(6 - z)}{(z - (2 + \sqrt{2}i))(z - (2 - \sqrt{2}i))} \tag{6.124b}$$

$$R_L(z) = 1 + \frac{12z}{(z - (3 + \sqrt{3}i))(z - (3 - \sqrt{3}i))}. \tag{6.124c}$$

It is well-known that the resulting stability domains for all three methods cover the half-plane $\mathbb{C}^-$, i.e. all three methods are A-stable (see e.g. Hairer and Wanner [127]).

The plots given in Figure 6.1 show the behavior of the absolute values $|R_G(z)|$, $|R_R(z)|$, and $|R_L(z)|$ for small and moderate negative real parts of z. For $|z \to \infty|$, $\mathfrak{Re}(z) < 0$, it holds that $|R_G(z)| \to 1$, $|R_R(z)| \to 0$, and $|R_L(z)| \to 1$. Accordingly, of the three methods only the two-stage Radau IIA method is L-stable.



(a) (b) (c)

Figure 6.1.: Absolute values $|R_G(z))|$, $|R_R(z))|$, and $|R_L(z))|$ of the stability functions for the Gauss, Radau IIA, and Lobatto IIIA collocation methods used in Colsol-DDE (from left to right).

6.7.2. Stability of Continuous Representations

Theoretical Background

For any given discrete Runge-Kutta method applied to an ODE-IVP for which the stage values are given by $g^j_{l+1} = f(t^j_{l+1}, y^j_{l+1}, c)$, $1 \leqslant j \leqslant \nu$, it is possible to find one or several continuous representations $\eta_{l+1}(t_l + \theta h_{l+1})$, which obey the continuity conditions $\eta_{l+1}(t_l) = y_l$ and $\eta_{l+1}(t_{l+1}) = y_{l+1}$, and which use only the stage values g^j_{l+1} of the discrete method. In the following, the stability properties of the continuous representation are discussed.

Consider a continuous Runge-Kutta method applied to the test equation (6.117). In the first step the continuous representation is given by

$$\eta_1(h\theta) = y_0 + h \sum_{j=1}^{\nu} b_i(\theta) \lambda y^j_1, \tag{6.125}$$

which can shortly be expressed as

$$\eta_1(h\theta) = \left[1 + h\lambda b(\theta)^T (1 - h\lambda \mathbf{A})^{-1} e\right] y_0. \tag{6.126}$$

Herein, $b(\theta)$ is a ν-dimensional fuction whose components are the continuous weight functions $b_i(\theta)$.

In analogy to the treatment of discrete Runge-Kutta methods the *stability function for continuous representations* is defined as follows:

Definition 6.18 (Stability Function for Continuous Representations)

The function

$$R_\eta(z, \theta) = 1 + z b(\theta)^T (1 - z\mathbf{A})^{-1} e \tag{6.127}$$

is called the stability function for continuous representations.

It is important to note that A-stability of a discrete Runge-Kutta method does not imply that for $\mathfrak{Re}(\lambda) < 0$ and for any possible continuous representation it holds that $\max_{0 \leqslant \theta \leqslant 1} |\eta_1(h\theta)| \leqslant |y_0|$.

In other words, from $|R(z)| < 1$ it does not follow that $\max_{0\leqslant\theta\leqslant 1} |R_\eta(z,\theta)| \leqslant 1$. In fact, $|R_\eta(z,\theta)|$ may even be unbounded.

Bellen and Zennaro [25] have introduced the following stability concept for continuous representations.

Definition 6.19 (Stability of Continuous Representations)

Consider a CRK method applied to the test equation (6.117). A continuous representation $\eta_1(h\theta)$ is called stable (with respect to the discrete Runge-Kutta method)*, if there exists a constant $M \geqslant 1$ such that*

$$\max_{0\leqslant\theta\leqslant 1} |\eta_1(h\theta)| \leqslant M \max(|y_0|, |y_1|) \tag{6.128}$$

for every fixed choice $z \in \mathbb{C}_0^- := \mathbb{C}^- \cup \{z \mid \mathfrak{Re}(z) = 0\}$.

For the special case of A-stable methods, the conditions simplifies to $\max_{0\leqslant\theta\leqslant 1} |\eta_1(h\theta)| \leqslant M|y_0|$, i.e. boundedness of the continuous representation for all $z \in \mathbb{C}_0^-$.

For the case of polynomial continuous representations, the stability function in equation (6.127) is a rational function in z. Hence, unboundedness of the stability function for $\mathfrak{Re}(z) \leqslant 0$ may occur for two reasons. Either the degree of the polynomial in the nominator is higher than the degree of the polynomial in the denominator (for at least one $\theta \in [0,1]$), or the polynomial in the denominator has a zero for some $z \in \mathbb{C}_0^-$. For the continuous representation of an A-stable method that uses only the stage values of the discrete method, the latter is not the case. Hence, in this case, only the degrees of the polynomials in the nominator and in the denominator need to be compared.

Properties of the Collocation Polynomials

Consider the collocation polynomials (of uniform local order q) of the three collocation methods that are implemented in Colsol-DDE. The corresponding stability functions R_{G,η_q}, R_{R,η_q}, and R_{L,η_q} for the one-stage Gauss collocation method, the two-stage Radau IIA collocation method, and the three-stage Lobatto IIIA collocation method, are given by follows:

$$R_{G,\eta_q}(z,\theta) = 1 + \frac{-2\theta z}{z-2} \tag{6.129a}$$

$$R_{R,\eta_q}(z,\theta) = 1 + \frac{6\theta z(1 + z(-\frac{2}{3} + \frac{1}{2}\theta))}{(z-(2+\sqrt{2}i))(z-(2-\sqrt{2}i))} \tag{6.129b}$$

$$R_{L,\eta_q}(z,\theta) = 1 + \frac{12z\left[\theta + z(-\frac{1}{2}\theta + \frac{1}{2}\theta^2) + z^2(\frac{1}{6}\theta^3 - \frac{1}{4}\theta^2 + \frac{1}{12}\theta)\right]}{(z-(3+\sqrt{3}i))(z-(3-\sqrt{3}i))}. \tag{6.129c}$$

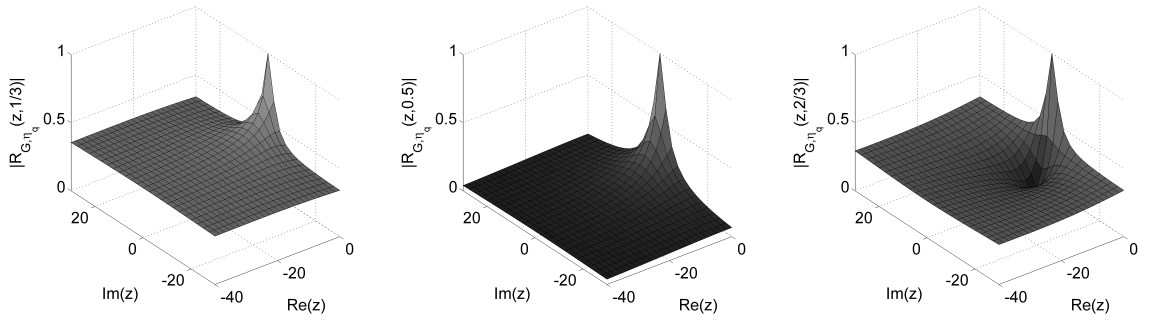
Observe that, as can be expected from the continuity condition, it holds that $R_\eta(z,1) = R(z)$ for all three methods.

From the degrees of the polynomials it is clear that $|R_{G,\eta_q}(z,\theta)|$ and $|R_{R,\eta_q}(z,\theta)|$ are bounded for all $z \in \mathbb{C}_0^-$, $\theta \in [0,1]$. Hence, the collocation polynomials are *stable* with respect to the corresponding discrete Runge-Kutta method. In fact, this is a special case of a general result found by Bellen and Zennaro [25]: Whenever the abscissae of a collocation method do not contain both 0 and 1, then the collocation polynomial is a stable continuous representation in the sense of Definition 6.19.

Some illustrations of the stability functions of the Gauss collocation polynomial and the Radau IIA collocation polynomial for three different values of θ are given in Figures 6.2 and 6.3, respectively. In both cases, the bound of the stability function is 1, which, for the Gauss method, is obvious from the fact that the continuous representation is just a linear function between the values y_0 and y_1.

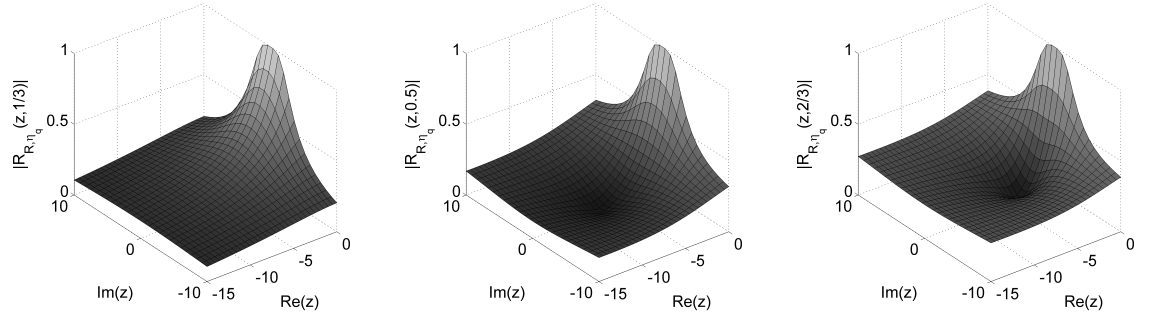
It remains to deal with the collocation polynomial of the three-stage Lobatto IIIA method. Clearly, for $\theta \notin \{0, \frac{1}{2}, 1\}$, the degree of the polynomial in the nominator is higher than the degree of the polynomial in the denominator. Therefore the collocation polynomial is unbounded and not stable in the sense of Definition 6.19. However, since at least the poles have positive real part, the absolute value $|R_{L,\eta}(z,\theta)|$ remains bounded in each bounded set $\mathbb{C}_0^- \cap \{z \mid |z| \leqslant \rho < \infty\}$.

The behavior is investigated in more detail. Since the function $z/(z-(3+\sqrt{3}i))$ is complex analytic in $\mathbb{C}_0^-$ and because $\mathbb{C}_0^-$ is a closed set, the function assumes its maximum on the imaginary



(a) (b) (c)

Figure 6.2.: Absolute values $|R_{G,\eta_q}(z,1/3))|$, $|R_{G,\eta_q}(z,0.5))|$, and $|R_{G,\eta_q}(z,2/3))|$, of the stability function for the collocation polynomial in the one-stage Gauss method.



(a) (b) (c)

Figure 6.3.: Absolute values $|R_{R,\eta_q}(z,1/3))|$, $|R_{R,\eta_q}(z,0.5))|$, and $|R_{R,\eta_q}(z,2/3))|$, of the stability function for the collocation polynomial in the two-stage Radau IIA method.

axis. More precisely, it holds that

$$\max_{z\in\mathbb{C}_0^-}\left|\frac{z}{z-(3+\sqrt{3}i)}\right| = \max_{z,\mathfrak{Re}(z)=0}\left|\frac{z}{z-(3+\sqrt{3}i)}\right| < 1.2. \tag{6.130}$$

The same bound holds if $3+\sqrt{3}i$ is replaced by $3-\sqrt{3}i$.

Furthermore, it holds that

$$\max_{0\leqslant\theta\leqslant 1}\left|\frac{1}{6}\theta^3-\frac{1}{4}\theta^2+\frac{1}{12}\theta\right| < 0.0081. \tag{6.131}$$

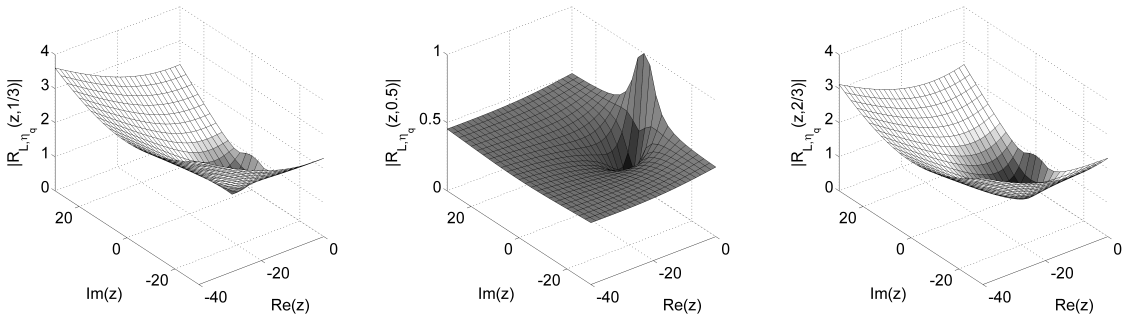
With this, it follows that asymptotically, for $|z|\to\infty$, $\mathfrak{Re}(z)\leqslant 0$, there is a linear increase with a very moderate prefactor 0.14. In view of the fact that the Lobatto IIIA method is not L-stable and therefore provides reasonable approximation only if $|z|$ does not become too large, this is acceptable from a practical point of view.

The mild asymptotic increase can be observed in the Figures 6.4a and 6.4c.

Properties of the Uniformly Corrected Polynomials

For the uniformly corrected polynomials it is recalled that the resulting continuous Runge-Kutta method can be expressed by the augmented schemes given in Subsection 6.2.2. Denoting the coefficient matrix of the augmented scheme by $\hat{\mathbf{A}}$ and setting $\hat{b}(\theta)^T := \begin{pmatrix} b(\theta)^T & b_*(\theta)\end{pmatrix}$, this yields the following stability function for the uniformly corrected polynomial of uniform local order $p = q+1$:

$$R_{\eta_p}(z,\theta) = 1 + z\hat{b}(\theta)^T(1-z\hat{\mathbf{A}})^{-1}e. \tag{6.132}$$



(a) (b) (c)

Figure 6.4.: Absolute values $|R_{L,\eta_q}(z,1/3))|$, $|R_{L,\eta_q}(z,0.5))|$, and $|R_{L,\eta_q}(z,2/3))|$, of the stability function for the collocation polynomial in the three-stage Lobatto IIIA method.

For the three methods considered in Colsol-DDE, this gives

$$R_{G,\eta_p}(z,\theta) = 1 + \frac{3z\left(\left(\frac{1}{2}\theta^2 - \frac{7}{6}\theta\right)z + 4\theta - 3\theta^2\right)}{(z-2)(z-\frac{3}{2})} \tag{6.133a}$$

$$\begin{aligned} R_{R,\eta_p}(z,\theta) = 1 + \frac{-\frac{18}{5}z}{(z-(2+\sqrt{2}i))(z-(2-\sqrt{2}i))(z-\frac{18}{5})} \\ \cdot \left(z^2\left[\theta^3 - \frac{17}{6}\theta^2 + \frac{19}{9}\theta\right] + z\left[-\frac{36}{5}\theta^3 + \frac{87}{5}\theta^2 - \frac{193}{15}\theta\right]\right. \\ \left. + \left[\frac{72}{5}\theta^3 - \frac{144}{5}\theta^2 + \frac{102}{5}\theta\right]\right) \end{aligned} \tag{6.133b}$$

$$\begin{aligned} R_{L,\eta_p}(z,\theta) = 1 + \frac{4z}{(z-(3+\sqrt{3}i))(z-(3-\sqrt{3}i))(z-\frac{16}{3})} \\ \cdot \left(z^3\left[-\frac{2}{3}\theta^4 + \frac{11}{6}\theta^3 - \frac{17}{12}\theta^2 + \frac{1}{4}\theta\right] + z^2\left[\frac{8}{9}\theta^4 - \frac{40}{9}\theta^3 + \frac{115}{18}\theta^2 - \frac{17}{6}\theta\right]\right. \\ \left. + z\left[\frac{64}{3}\theta^4 - \frac{128}{3}\theta^3 + \frac{40}{3}\theta^2 + 11\theta\right] + \left[-\frac{256}{3}\theta^4 + \frac{512}{3}\theta^3 - \frac{256}{3}\theta^2 - 16\theta\right]\right). \end{aligned} \tag{6.133c}$$

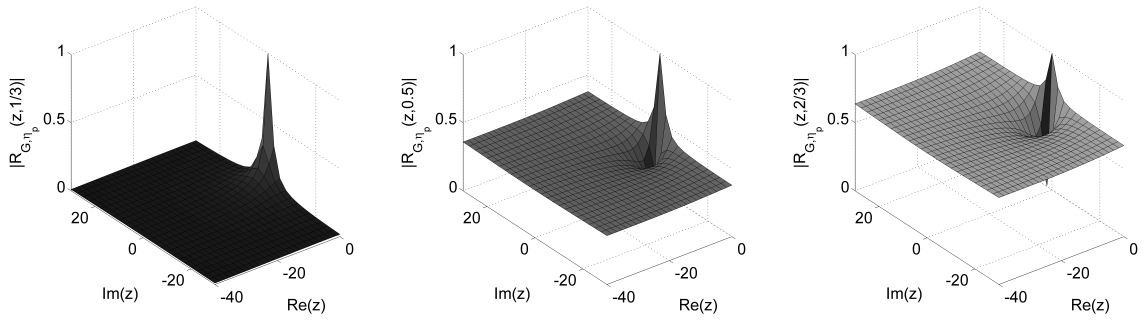
The additional poles of these rational functions compared to those in equation (6.129) are located at 3/2, 18/5, and 16/3, and thus all have positive real part. Therefore, no unboundedness for any of the three stability functions occurs due to the presence of poles in $\mathbb{C}_0^-$.

It can be shown that this property is obtained only if the additional abscissa for the implicit uniform correction is within $(0, \frac{1}{2})$ for the Gauss method and the Lobatto IIIA method, and within $(0, \frac{1}{3})$ for the Radau IIA method. This motivates, a posteriori, the selection of the additional abscissa in Subsection 6.2.2.

For the stability functions $R_{G,\eta_p}(z,\theta)$ and $R_{R,\eta_p}(z,\theta)$ it holds, in addition, that the degree of the polynomial in the nominator equals the degree of the polynomial in the denominator. Hence, also the corrected polynomials are stable with respect to the corresponding discrete method. The stability of these polynomials can also be concluded from a general theorem on the stability of continuous representations that are obtained by implicit uniform corrections, see Bellen and Zennaro [25].

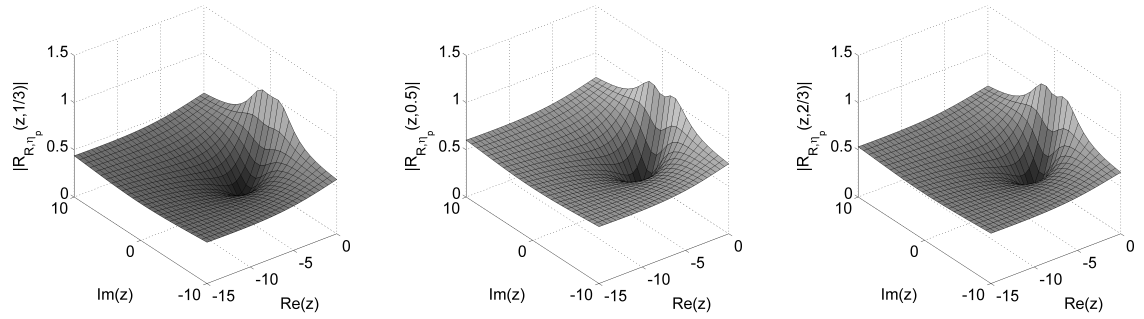
The Figures 6.5 and 6.6 show the behavior of the stability functions for the corrected polynomials in case of the Gauss and Radau IIA method for three different values of θ.

For the corrected polynomial in the case of the Lobatto IIIA method, the degree of the polynomial in the nominator is one higher than in the denominator. Hence, the polynomial is unbounded and not stable in the sense of Definition 6.19. In order to compute, once again, the asymptotic linear



(a) (b) (c)

Figure 6.5.: Absolute values $|R_{G,\eta_p}(z,1/3))|$, $|R_{G,\eta_p}(z,0.5))|$, and $|R_{G,\eta_p}(z,2/3))|$, of the stability function for the corrected polynomial in case of the one-stage Gauss method.



(a) (b) (c)

Figure 6.6.: Absolute values $|R_{R,\eta_p}(z,1/3))|$, $|R_{R,\eta_p}(z,0.5))|$, and $|R_{R,\eta_p}(z,2/3))|$, of the stability function for the corrected polynomial in case of the two-stage Radau IIA method.

increase, observe that

$$\max_{0\leqslant\theta\leqslant 1}\left|-\frac{2}{3}\theta^4+\frac{11}{6}\theta^3-\frac{17}{12}\theta^2+\frac{1}{4}\theta\right|<0.053. \tag{6.134}$$

From this it follows that the prefactor in the asymptotic linear increase is less than 0.36. Therefore the stability function assumes moderate values of less than 10 for $|z|\leqslant 20$, see Figure 6.7.

6.7.3. Error Estimation

Error Estimation for the Continuous Representation

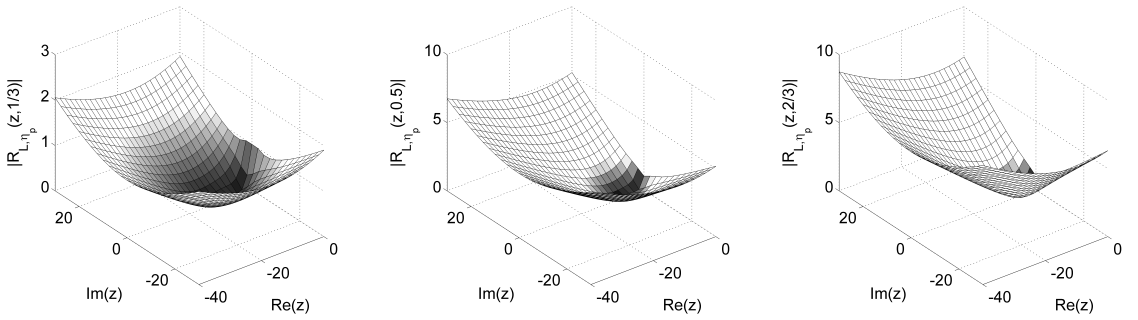
Eventually, the behavior of the employed error estimators is regarded for the asymptotics $|z|\to\infty$, $\mathfrak{Re}(z)\leqslant 0$. For the error estimation of the continuous representation, the following expression is obtained in case of the Gauss collocation method.

$$\begin{aligned}\max_{0\leqslant\theta\leqslant 1}|\eta_q(h\theta)-\eta_p(h\theta)| &= \max_{0\leqslant\theta\leqslant q}|R_{G,\eta_q}(z,\theta)-R_{G,\eta_p}(z,\theta)||y_0|\\ &= \left|R_{G,\eta_q}\left(z,\frac{1}{2}\right)-R_{G,\eta_p}\left(z,\frac{1}{2}\right)\right||y_0|\\ &\to \frac{3}{8}|y_0| \quad \text{for} \quad z\to\infty,\ \mathfrak{Re}(z)\leqslant 0.\end{aligned} \tag{6.135}$$

The sole subscript of η thereby represents its uniform local order.

Further, for the Radau IIA collocation method, the asymptotics

$$\max_{0\leqslant\theta\leqslant 1}|\eta_q(h\theta)-\eta_p(h\theta)|\to\frac{8}{15}|y_0| \tag{6.136}$$



(a) (b) (c)

Figure 6.7.: Absolute values $|R_{L,\eta_p}(z,1/3))|$, $|R_{L,\eta_p}(z,0.5))|$, and $|R_{L,\eta_p}(z,2/3))|$, of the stability function for the corrected polynomial in case of the Lobatto IIIA method used in Colsol-DDE.

is obtained, and for the Lobatto IIIA collocation method, the result is

$$\max_{0\leqslant\theta\leqslant 1} |\eta_q(h\theta) - \eta_p(h\theta)| \to \frac{1}{6}|z||y_0|. \tag{6.137}$$

Hence, for the Gauss and for the Radau IIA collocation method, the error estimate is bounded, whereas for the Lobatto IIIA collocation method the error estimate increases linearly with $|z|$. In all three cases, the asymptotic behavior of the error estimate therefore matches the asymptotic behavior of the numerical error.

Error Estimation for the Discrete Method

The higher order discrete method yields, for the test equation (6.117), the following result:

$$y_{p+1} = \frac{1 + z\sum_{i=1}^{\mu-1} B_i \eta_p(\Gamma_i h)}{1 - zB_\mu} y_0 \tag{6.138}$$

Since the weight B_μ for the last stage is positive for all three employed quadrature formulae, there is no pole for $z \in \mathbb{C}_0^-$.

In case of the Gauss method, the higher order continuous representation η_p is bounded, and hence y_{p+1} is bounded as well. In fact, it can be shown that $|y_{p+1}| \to 0$ for $|z| \to \infty$, $\mathfrak{Re}(z) \leqslant 0$. As a consequence, $|y_{p+1} - y_p|$ remains bounded in the limit $|z| \to \infty$, $\mathfrak{Re}(z) \leqslant 0$, which is the same behavior as the true error of the Gauss collocation method (recall Figure 6.1a).

For the Radau IIA method, it also holds that the higher order continuous representation η_p is bounded, and hence y_{p+1} is bounded as well. It can further be shown that $|y_{p+1} - y_p| \to 1.8$ for $|z| \to \infty$, $\mathfrak{Re}(z) \leqslant 0$. Unfortunately, this does not correspond to the behavior of the true error of the Radau collocation method, which, instead, vanishes asymptotically (recall Figure 6.1b).

For the Lobatto method, the higher order continuous representation η_p is unbounded, and the same holds for the result y_{p+1} of the higher order method. Furthermore, it can be shown that $|y_{p+1} - y_p| \to 0.8|z|$ for $|z| \to \infty$, $\mathfrak{Re}(z) \leqslant 0$. This means that the error estimate is unbounded, although it holds for the true error of the Lobatto IIIA collocation method method that $|y_p - y_0 \exp(\lambda h)| \to 1$, i.e. the true error is bounded.

Apparently, for both the Radau IIA method and the Lobatto IIIA collocation method, the estimated error exhibits a different asymptotic behavior than the true error of the numerical method. A similar problem is known in the literature for the code RADAU5 and its modification for DDEs, called RADAR5, see Hairer and Wanner [127]. In analogy to the remedy proposed therein, the error estimation in Colsol-DDE can optionally be done by replacing the quantity $|(y_{p+1})_i - (y_p)_i|$ in equation (6.112a) with $|\mathbf{P}\,((y_{p+1})_i - (y_p)_i)|$. Thereby, $\mathbf{P}$ is a matrix that is given by

$$\mathbf{P} = (\mathbf{1} - hB_\mu \mathbf{J})^{-1} \tag{6.139}$$

and $\mathbf{J}$ is the most recently computed approximation of the Jacobian of the function F_{iqr}. For the test equation (6.117) it holds that $\mathbf{J} = \lambda$ and thus $\mathbf{P} = 1/(1 - zB_\mu)$. Consequently, for vanishing stiffness or stepsize, $z \to 0$, the error estimate is identical to $|y_{p+1} - y_p|$. On the other hand, for $|z| \to \infty$, $\mathfrak{Re}(z) \leqslant 0$, it ensures that the error estimates for the Radau IIA and for the Lobatto IIIA method have the same behavior as the exact error of the numerical method, i.e. the estimate approaches zero for the former and is bounded for the latter.

6.7.4. Other Stability Concepts for ODEs and DDEs

So far the methods implemented in Colsol-DDE were investigated with respect to the elementary concepts of A-stability and L-stability as well as with respect to one specific concept for the stability of continuous representations. All of them were related to the simple ODE test equation (6.117).

There exists a large number of different stability concepts. For example, AN-stability is related to a test equation where the prefactor λ in equation (6.117) is replaced by a time-varying coefficient $\lambda(t)$, whereas the so-called B-stability is related to a general nonlinear ODE.

With regard to DDEs, the simplest extension of the test equation is to consider $\dot{\mathbf{y}}(t) = \lambda_1 y(t) + \lambda_2 y(t-\tau)$, where τ is a constant delay. For certain pairs $(\lambda_1, \lambda_2) \in \mathbb{C} \times \mathbb{C}$ it can be shown that the solution $y(t)$ vanishes asymptotically for $t \to \infty$, independent of the delay τ. Requesting the same property for the numerical solution leads to so-called P- and GP-stability, where the difference between the two is that the first is related to stepsizes fulfilling $h = \tau/m$, $m \in \mathbb{N}$, whereas the second allows arbitrary constant stepsizes. Alternatively, it is also possible to make an analysis for a fixed delay τ, which leads, for constant stepsizes $h = \tau/m$, to the definition of D-stability. Further stability concepts exist as generalizations of AN- and B-stability.

For an overview on this topic, the reader is referred to Bellen and Zennaro [26]. However, a comprehensive discussion of the methods implemented in Colsol-DDE with respect to the large number of existing stability concepts is beyond the scope of this thesis.

6.8. The Main Algorithm

This section is about the algorithm that is implemented in Colsol-DDE for solving DDE-IVPs.

For its presentation, assume that n_s denotes the number of discontinuities that were found until the mesh point t_l. Further, let $\underline{l}$ be the index such that $t_{\underline{l}} = s_{n_s}$ denotes the left border of the current discontinuity interval. Initially, $n_s = 0$, and $s_0 = t^{ini}(c)$ is the initial time.

Using this notation, Algorithm 6.20 presents the main algorithm for solving DDE-IVPs. The algorithm is applicable for solving the IVP in an interval $[t_{\underline{l}}, t_{\bar{l}}]$, where $t_{\bar{l}}$ is either equal to the next discontinuity point s_{n_s+1} (which is yet to be determined) or it is equal to the final time $t^{fin}(c)$. The practical determination of the next discontinuity point s_{n_s+1} by monitoring the discontinuity interval indicators is discussed in Section 6.9.

Algorithm 6.20 (Main Solution Algorithm in Colsol-DDE)

1. *Start with $l = \underline{l}$, with some given t_l, y_l, $g_l := f(t_l, y_l, c, v_l)$, and with a proposed stepsize h_{l+1}. Let $\Xi^{col} = \Xi^{ucp} = \Xi^{pqr} = 1$ if $l = 0$ (i.e. $t_l = t_0$ is the initial time) or if t_l is a time point of discontinuity of order 0 or 1 in y. Otherwise, let $\Xi^{col} = \Xi^{ucp} = \Xi^{iqr} = 0$. Let h_{min} be a user-given minimum stepsize.*

2. *Determine the initial guess $(g^j_{l+1})^0$ for the stage values of the collocation method as described in Subsection 6.5.3.*

3. *If $\Xi^{col} = 1$, compute and decompose the Jacobian of the equation system (6.72) with respect to g^j_{l+1}.*

4. *Apply a Newton-type method to the nonlinear equation system (6.72).*
   a) *If convergence is achieved in $n \leqslant n^1_{itmax}$ iterations, set $\Xi^{col} = 0$ and proceed to step 5.*
   b) *If convergence is achieved in $n^1_{itmax} < n \leqslant n^2_{itmax}$ iterations, set $\Xi^{col} = 1$ and proceed to step 5.*
   c) *If no convergence is achieved after $n = n^2_{itmax}$ iterations and if $\Xi^{col} = 0$, set $\Xi^{col} = 1$ and go back to step 3.*

d) If no convergence is achieved after $n = n^2_{itmax}$ iterations, if $\Xi^{col} = 1$ and $h_{l+1} > h_{min}$, set $h_{l+1} = \max(h_{l+1}/2, h_{min})$ and go back to step 2.

*e) If no convergence is achieved after $n = n^2_{itmax}$ iterations, if $\Xi^{col} = 1$ and $h_{l+1} = h_{min}$, **stop** and exit with an error message.*

5. *Determine the initial guess $(g^*_{l+1})^0$ for the stage value of the implicit uniform correction as described in Subsection 6.5.3.*

6. *If $\Xi^{ucp} = 1$, compute and decompose the Jacobian of the equation system (6.74) with respect to g^*_{l+1}.*

7. *Apply a Newton-type method to the nonlinear equation system (6.74).*

   a) If convergence is achieved in $n \leqslant n^1_{itmax}$ iterations, set $\Xi^{ucp} = 0$ and proceed to step 8.

   b) If convergence is achieved in $n^1_{itmax} < n \leqslant n^2_{itmax}$ iterations, set $\Xi^{ucp} = 1$ and proceed to step 8.

   c) If no convergence is achieved after $n = n^2_{itmax}$ iterations and if $\Xi^{ucp} = 0$, set $\Xi^{ucp} = 1$ and go back to step 6.

   d) If no convergence is achieved after $n = n^2_{itmax}$ iterations, if $\Xi^{ucp} = 1$ and $h_{l+1} > h_{min}$, set $h_{l+1} = \max(h_{l+1}/2, h_{min})$ and go back to step 2.

   *e) If no convergence is achieved after $n = n^2_{itmax}$ iterations, if $\Xi^{ucp} = 1$ and $h_{l+1} = h_{min}$, **stop** and exit with error message.*

8. *Determine the initial guess $(g^{\diamondsuit}_{l+1})^0$ for the stage value of the implicit quadrature rule as described in Subsection 6.5.3.*

9. *If $\Xi^{iqr} = 1$, compute and decompose the Jacobian of the equation system (6.76) with respect to $g^{\diamondsuit}_{l+1}$.*

10. *Apply a Newton-type method to the nonlinear equation system (6.76).*

    a) If convergence is achieved in $n \leqslant n^1_{itmax}$ iterations, set $\Xi^{iqr} = 0$ and proceed to step 11.

    b) If convergence is achieved in $n^1_{itmax} < n \leqslant n^2_{itmax}$ iterations, set $\Xi^{iqr} = 1$ and proceed to step 11.

    c) If no convergence is achieved after $n = n^2_{itmax}$ iterations and if $\Xi^{iqr} = 0$, set $\Xi^{iqr} = 1$ and go back to step 9.

    d) If no convergence is achieved after $n = n^2_{itmax}$ iterations, if $\Xi^{iqr} = 1$ and $h_{l+1} > h_{min}$, set $h_{l+1} = \max(h_{l+1}/2, h_{min})$ and go back to step 2.

    *e) If no convergence is achieved after $n = n^2_{itmax}$ iterations, if $\Xi^{iqr} = 1$ and $h_{l+1} = h_{min}$, **stop** and exit with error message.*

11. *Compute C_{disc} and C_{unif} as defined in equation (6.112) and a proposition stepsize h_{prop} as given by the right hand side of equation (6.114).*

    a) If $C_{disc} > 1$ or $C_{unif} > 1$, set $h_{l+1} = h_{prop}$ and go back to step 2.

    b) If $C_{disc} \leqslant 1$ and $C_{unif} \leqslant 1$, set $h_{l+2} = h_{prop}$ and proceed with step 12.

12. *Set $l = l + 1$ and go to step 2.*

For the case that t_{l+1} in step 12 is an approximation of a discontinuity point, the reader is referred to the embedding of Algorithm 6.20 into Algorithm 6.21. If $t_{l+1} = t^{fin}(c)$, then stop and signal successful integration to the user.

6.9. Detecting and Locating Discontinuity Points

A key feature of Colsol-DDE is that its implementation closely follows the definition of the practical variant of the modified standard approach. This means that the code monitors the discontinuity interval indicators and, accordingly, computes past states from sufficiently smooth branches of the numerical solution.

The following algorithm works under the (much) simplifying assumptions that there is only one delay and that the propagated discontinuities occur well separated of each other, i.e. in each integration step the deviating argument crosses at most one past discontinuity point. Several modifications for more general problem classes as well as additional regularity checks and fine tuning of the algorithm are discussed later.

For simplicity of notation, the discontinuity interval indicator ξ_1^α for the sole deviating argument is in the following denoted by ξ.

Algorithm 6.21 ((Simplified) Version of Discontinuity Treatment in Colsol-DDE)

1. *Start with $l = 0$, the initial time $t_l = t_0 = t^{ini}(c)$, the initial value $y_0 = y^{ini}(c)$, and initial stepsize h_1. Denote the time points of the initial discontinuities up to order $p+1$ by $\hat{s}_i$, $-n_s^\phi \leqslant i \leqslant -1$, and set $\hat{s}_0 = t_0$. Let the orders of the discontinuity at $\hat{s}_i$ be denoted by o_i, for $-n_s^\phi \leqslant i \leqslant -1$. Further, set $n_s = 0$ and let ξ be the discontinuity interval indicator for the sole delay τ_1, whose initial value is such that $-n_s^\phi \leqslant \xi \leqslant 0$, i.e. the indicator points to a smooth branch $\phi_i(t,c)$, $-n_s^\phi \leqslant i \leqslant 0$, of the initial function. Set $\alpha_{search} = 0$.*

2. *Compute $t_{l+1} = t_l + h_{l+1}$, and y_{l+1} and $\eta_{l+1}(t_l + \theta h_{l+1})$ by the augmented CRK method as in steps 2-11 of Algorithm 6.20 until the criteria for error control, $C_{disc} \leqslant 1$, $C_{unif} \leqslant 1$, are fulfilled. All past states are obtained from the discontinuity interval indicated by ξ, i.e. from $[\hat{s}_{\xi-1}, \hat{s}_\xi]$. Extrapolations are used if necessary.*

3. *Compute for $\hat{s}_k$, $k = \xi - 1$ and, if $\xi \neq n_s + 1$, then also for $k = \xi$, the following quantity:*

$$\mu_k = \begin{cases} +1 & \text{if} \quad t_{l+1} - \tau_1(t_{l+1}, y_{l+1}, c) - \hat{s}_k > (t^{fin}(c) - t^{ini}(c)) \cdot \gamma_{crit} \\ 0 & \text{if} \quad |t_{l+1} - \tau_1(t_{l+1}, y_{l+1}, c) - \hat{s}_k| \leqslant (t^{fin}(c) - t^{ini}(c)) \cdot \gamma_{crit} \\ -1 & \text{if} \quad t_{l+1} - \tau_1(t_{l+1}, y_{l+1}, c) - \hat{s}_k < (t^{fin}(c) - t^{ini}(c)) \cdot \gamma_{crit} \end{cases} \tag{6.140}$$

   Herein, γ_{crit} is a user-given criterion for the detection of zeros.

   a) *If $\mu_{\xi-1} = +1$ and $\mu_\xi = -1$ and $\alpha_{search} = 0$, no propagated discontinuity is detected. Proceed with step 6.*

   b) *If $\mu_{\xi-1} = 0$ or $\mu_\xi = 0$, then a propagated discontinuity is detected. Proceed with step 5.*

   c) *If $\mu_{\xi-1} = -1$ or $\mu_\xi = +1$, then a propagated discontinuity is located somewhere on $[t_l, t_{l+1}]$. Determine the zero $t^\star$ of $t - \tau_1(t, \eta_{l+1}(t), c) - \hat{s}_{k'}$ by using* regula falsi*, where $k' = \xi - 1$ if $\mu_{\xi-1} = -1$ and $k' = \xi$ if $\mu_\xi = +1$. Save $t_{old} = t_{l+1}$, $x_{old} = t_{l+1} - \tau_1(t_{l+1}, y_{l+1}, c) - \hat{s}_{k'}$. Set $h_{l+1} = t^\star - t_l$, $\alpha_{search} = 1$, and go back to step 2.*

   d) *If $\mu_{\xi-1} = +1$ and $\mu_\xi = -1$ and $\alpha_{search} = 1$, no propagated discontinuity is detected but the zero search has previously been initialized. Compute $x = t_{l+1} - \tau_1(t_{l+1}, y_{l+1}, c) - \hat{s}_{k'}$, where k' is the index of the previous discontinuity whose propagation was detected in the last time that c) was called. Determine the zero $t^\star$ of a linear function that interpolates (t_{l+1}, x) and (t_{old}, x_{old}), set $h_{l+1} = t^\star - t_l$, and go back to step 2.*

4. *If the order of the parent discontinuity (given by $o_{\xi-1}$ or o_ξ) is less than or equal to p, then set $\hat{s}_{n_s+1} = t_{l+1}$, $o_{n_s+1} = o_{\xi'} + 1$ (with $\xi' = \xi - 1$ or $\xi' = \xi$) and $n_s = n_s + 1$.*

5. *If $\mu_{\xi-1} = 0$, then set $\xi = \xi - 1$. If $\mu_\xi = 0$, then set $\xi = \xi + 1$.*

6. *Set $\alpha_{search} = 0$, $l = l + 1$, and proceed with step 2, i.e. with the next integration step.*

Even though this algorithm captures the essential points of the discontinuity detection mechanism of Colsol-DDE, the practical realization is significantly more involved for the following reasons.

- Checking Uniqueness of Consistent Choice for ξ: According to the definition of the practical variant of the modified standard approach (Definition 5.22), it needs to be ensured that there is only one consistent choice of the discontinuity interval indicator. If the assumptions of "the special case" as discussed in Section 5.4 are fulfilled (discontinuity points are well-separated and the propagation switching functions have zeros of multiplicity one), then Algorithm 6.21 detects, for sufficiently small stepsizes, all zeros of the propagtion switching functions.

  In order to verify that the behavior of the propagation switching functions is "regular" in the vicinity of a zero, and that the discontinuity interval indicator is the unique consistent choice,

Colsol-DDE performs the following checks. First, a check is done that the time derivative of the propagation switching function in the determined discontinuity point is "significantly" non-zero. Second, it is checked that the propagation switching function does not fulfill the zero criterion in equation (6.140) a "short" time before and after the determined discontinuity point. Thirdly, it is verified that the propagation switching function leaves – into both time directions – its zero set into the direction indicated by ξ. The criteria for "significantly non-zero" and "short time before/after a discontinuity point" thereby depend on user-given parameters.

If the assumptions of "the special case" discussed in Section 5.4 are fulfilled, then Colsol-DDE corresponds to a realization of the practical variant of the modified standard approach, meaning that all sign changes in the propagation switching functions are detected, and that the discontinuity interval indicator is constant and consistent between two mesh points. In addition, the above-described safeguard checks allow to detect most irregularities in the behavior of the propagation switching functions.

- Decision on further propagation: After a discontinuity point has been successfully included in the mesh, it needs to be decided whether the current discontinuity needs to be propagated further. Ideally, this decision should be based upon the order of the discontinuity in the exact solution, which, however, is in practice typically unknown.

  Colsol-DDE therefore determines a lower bound of the order of propagated discontinuities. More precisely, it is checked for the initial discontinuities, whether the initial function is continuous or not and, accordingly, the order is set to 0 or 1. For the propagation of discontinuities, it is exploited in step 4 of the algorithm that a lower bound for the order of the child discontinuity is given by $o_c = o_p + 1$, where o_p is the order of the parent discontinuity. The strategy in step 4 further guarantees that all time points of discontinuity up to order $p + 1$ are propagated, as it is necessary for the application of the implicit quadrature rule.

- Multiple delays: The technical realization of Algorithm 6.21 is severely complicated due to the fact that there maybe multiple delays. In particular, this requires to check, after every integration step, the zero criteria for the propagation switching functions for all deviating arguments. An approximation y_{l+1} of the state at t_{l+1} is then accepted in step 3a, if none of the propagation switching function has changed its sign, and it is accepted in step 3b if one or several propagated discontinuities are detected. In step 3c, the zeros of all propagation switching functions which have changed their sign need to be determined, and the earliest of all these zeros is used to recompute h_{l+1}. In step 3d, linear interpolation is used for all propagation switching functions, for which a zero has been detected the last time 3c was called.

  In view of the "special case" as discussed in Section 5.4 it is, strictly speaking, necessary that zeros of different propagation switching functions do not coincide. Colsol-DDE is less rigorous at this point and allows arbitrarily many propagated discontinuities of order greater or equal to 2 to coincide. Similarly, it is allowed that they occur arbitrarily close after each other. Only coinciding propagated discontinuities of order 1 are excluded by suitable checks, because otherwise the IVP solution is typically not differentiable with respect to parameters (see Chapter 7).

- Multiple switching functions (i.e. extension to IHDDE-IVPs):
  The presence of multiple switching functions in IHDDEs adds another layer of complexity to the development of a practical code. Apparently, this requires to check also the signs of the switching functions and to make the decision in step 3c dependent on these signs. In addition, non-zero impulses generally affect the values of other state-dependent switching functions and propagation switching functions. Accordingly, after an impulse occured, the switching function signs and discontinuity interval indicators need to be updated accordingly.

  With regard to coinciding discontinuities, Colsol-DDE makes several checks that are motivated by the theory for differentiability of IVP solutions with respect to parameters (see Chapter 7). This includes, in particular, that switching functions have a non-zero time derivative at their zeros. For a further discussion of the practical realization in Colsol-DDE, it is referred to Subsection 9.1.11.

- Dealing with Inaccurate Extrapolations: The usage of the modified standard approach guarantees, in contrast to the standard approach, that past states are always obtained by evaluations of sufficiently smooth functions. However, the use of extrapolations may also become problematic when the deviating argument assumes values far outside of the interval $[t_{l'}, t_{l'+1}]$ from which the continuous representation is used. This may happen, e.g., when the stepsize in the past, $h_{l'+1}$, is very small because $t_{l'}$ was coincidentally very close to the next discontinuity point $t_{l'+1}$.

  In the prescribed situation, the result of the extrapolation and thus the discrete and continuous approximations in a trial step may be completely unreliable. This typically leads to repeated calls of step 3d. As a simple but effective remedy, Colsol-DDE allows the user to specify a maximum number of repeated calls of step 3d after which a bijection step is done instead of a linear interpolation.

- Separate treatment of constant delays and simple time-dependent switching functions: The zeros of (simple) time-dependent switching functions as well as the zeros of propagation switching functions that are associated to constant or time-dependent delays can be computed a priori. It is therefore ineffcient to take, as suggested in algorithm 6.21, first the integration step to a time point t_{l+1} (which requires the solution of three nonlinear equation systems) and to check later whether the signs of switching functions or propagation switching functions have changed. Instead, this can be done before the integration step is done, which saves the computational effort for an integration step that is rejected anyway.

  Colsol-DDE therefore provides separate interfaces for simple time-dependent switching functions and constant delays. A special treatment of time-dependent switching functions and time-dependent delays is not yet realized.

Another important issue not mentioned in the above list is that of coinciding discontinuities, i.e. a situation in which several switching functions or propagation switching functions are zero at the same time (either exactly or at least according to the employed zero criteria). A simple example for this is given by a DDE-IVP with two constant delays, in which one delay is an integer multiple of the other.

It is clear that some discontinuities should not coincide or otherwise the solution is not unique. For example, imagine an IHODE-IVP with two switching functions, whose associated impulse function are not identically zero. If the switching functions become zero at the same time point, then the application of the impulses is not generally commutative. Accordingly, there may exist two different ways of continuing the solution beyond the discontinuity point. Colsol-DDE detects such a situation, aborts the solution and informs the user with an error message.

In general, coinciding zeros of switching functions and propagation switching functions may not only lead to non-uniqueness of the solution, but also to a non-continuous or non-differentiable dependence on the parameters c that occur in the model functions of the considered IVP. Modelers, and thus users of practical codes, typically expect that IVP solutions for their models are differentiable with respect to the parameters; if this is not the case, this may hint to errors in the model. Therefore, a good code should be able to detect points of non-differentiability with respect to the parameters.

Finding sufficient conditions for differentiability of IHDDE-IVP solutions with respect to parameters is one of the main topics in the next part of the thesis. The theoretical results are then, in Chapter 9, used to discuss the practical checks that Colsol-DDE does in order to ensure that the computed IVP solutions are differentiable.

Part III.

Sensitivities of IHDDE-IVP Solutions with Respect to Parameters

7. Differentiability Theory

Differentiability results with respect to parameters, beside the obvious theoretical importance, have a natural application in the problem of identification of parameters of the equations.

Hartung and Turi, in the introduction of their paper "On Differentiability of Solutions with respect to Parameters in State–Dependent Delay Equations" [139].

In Chapter 1, impulsive hybrid discrete-continuous delay differential equations (IHDDEs) and the corresponding initial value problems (IHDDE-IVPs) were defined in such a way that all model functions – e.g. the right-hand-side function, the delay functions, and the initial function – depend on parameters c (see Definitions 1.1 and 1.2). The results for the existence and uniqueness of IHDDE-IVP solutions in Chapter 4, however, were presented for arbitrary but fixed parameter values, i.e. the parameter dependence was essentially ignored.

The situation is different in this chapter. Here, it becomes important that the solution of an IHDDE-IVP generally depends on the parameters c, and that it should therefore be denoted by $y(t;c)$, cf. Section 2.4. Well-posedness of a mathematical problem in the sense of Hadamard comprises, besides existence and uniquess of the solution, also a continuous dependence on input quantities, which in the context of this thesis is equivalent to a continuous dependence on the parameters. However, the subject of this chapter is to go directly one step further, i.e. to continuous differentiability of IHDDE-IVP solutions with respect to the parameters.

More precisely, the issue treated in this chapter is as follows: Given some *nominal values* $\tilde{c}$ of the parameters (also called *nominal parameters*), under which conditions does there exist a neighborhood $\mathcal{U}^c$ of $\tilde{c}$ such that for $c \in \mathcal{U}^c$ and (almost) all times $t \in \mathcal{T}^f(c) = (-\infty, t^{fin}(c)]$ the IVP-solution $y(t;c)$ is continuously differentiable with respect to the parameters? In the case that such a neighborhood exists, the derivative is denoted by

$$\mathbf{W}(t;c) = \frac{\partial y(t;c)}{\partial c}, \tag{7.1}$$

where the function $\mathbf{W} : \mathcal{T}^f(c) \times \mathcal{U}^c \to \mathbb{R}^{n_y \times n_c}$ is called the *Wronskian matrix*. The Wronskian matrix itself is, like the state y, also a function of the time and the parameters.

Literature Survey

In the context of delay differential equations (DDEs), Hale and Ladeira [131] made an early contribution to the differentiability of IVP solutions with respect to constant delays. Their approach is to convert the DDE-IVP into an equivalent integral equation and to reformulate the integral equation as a fixed-point problem. Then an extension of the uniform contraction principle is applied in order to obtain the fixed point as a differentiable map of the delays.

For DDEs with state-dependent delays, Hartung [133, 135] has obtained differentiability of IVP solutions with respect to parameters in the right-hand-side function, in the initial function, and in the delay function, in a pointwise sense. Furthermore – inspired by the techniques of Hale and Ladeira [131] – Hartung and Turi [139] and Chen, Hu, and Wu [63] present results for first and second order differentiability in a weaker sense. Hale and Verduyn Lunel [130], page 48f, discuss differentiability of IVP solutions with respect to initial data in a fairly general class of functional differential equations; their proof is also based upon the differentiability of a fixed-point of a contraction mapping. For compact overviews over the differences in the various differentiability theorems regarding, in particular, assumptions on the smoothness of the initial function, the selection of function spaces, and the choice of norms on these function spaces, it is referred to Hartung et al. [137] and Hartung [135].

A different approach for proving differentiability of DDE-IVP solutions with respect to parameters has been pursued by Brewer [48], Robbins [216], and Banks, Robbins, and Sutton [18]. There, the DDE-IVP is first converted into a so-called *abstract Cauchy problem* and then differentiability results for this abstract Cauchy problem are derived.

All of the results cited above make, at some point, smoothness assumptions on the initial function and on the link of the initial function $\phi(t)$ to $y(t)$ for $t \geqslant t^{ini}(c)$. Typically, $\phi(t,c)$ is assumed to be absolutely continuous and the link to $y^{ini}(c)$ is assumed to be continuous (Hartung [133] requires even a continuously differentiable link at the initial time). In the context of this thesis, however, the initial function is only asssumed to be piecewise continuous and therefore the above results cannot be used.

A more suitable starting point for discussing the differentiability of IHDDE-IVP solutions with respect to parameters is obtained by looking at the theory for hybrid discrete-continuous ordinary differential equations (HODEs), impulsive ordinary differential equations (IODEs), and impulsive hybrid discrete-continuous ordinary differential equations (IHODEs). The Wronskian matrix $\mathbf{W}(t;c)$ is thereby typically assumed to be in the space of piecewise continuously differentiable function, i.e. $\mathbf{W}(\cdot;c) \in \mathscr{PD}(\mathcal{T}(c), \mathbb{R}^{n_y \times n_c})$ (where, as usual, $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$). For the IHODE-case, Bock [35] has first formulated a theorem that guarantees differentiability of the discontinuity points and of the IVP solution with respect to parameters. A sequence of later works, Bock [39], page 191ff, Lakshmikantham [169], page 73ff, von Schwerin, Winckler, and Schulz [252], and Galán, Feheery, and Barton [111] have found similar sets of sufficient differentiability conditions. In addition, these works present expressions that quantify the jump in the Wronskian.

With regard to DDEs, derivatives of IVP solutions with respect to parameters that are discontinuous in time have rarely been considered. To the knowledge of the author, Baker and Paul [13] were the first who explicitly allowed this case. Later, ZivariPiran [271] and ZivariPiran and Enright [273] used the formalism known from IHODE-IVPs for deriving the size of the jump in the Wronskian at the discontinuity points. However, the important issue of differentiability of DDE-IVP solutions with respect to parameters is barely touched in their works. In other words, there is a lack of rigorous theoretical foundation for the existence of the derivatives that they practically compute. Lenz, Schlöder, and Bock [173] make a first step in this direction and provide a differentiability result for DDE-IVPs with multiple constant delays. This work serves as a basis for several new developments presented in this chapter, as outlined in the following.

Novel Results Presented in This Chapter

In this chapter, it is demonstrated how the ideas of the proof of the differentiability result in Lenz, Schlöder, and Bock [173] can be transferred to IVPs in more complicated differential equations, namely to DDEs with time-dependent delays and even to IHDDEs with (simple) time-dependent switching functions. Furthermore, a differentiability theorem for DDEs with state-dependent delay functions is given, which is based on the assumption that a solution of the DDE-IVP is available. The general case of IHDDEs with state-dependent delay and switching functions is also addressed.

It should be remarked that the differentiability of IVP solutions as discussed in this chapter means differentiability with respect to the finite-dimensional parameter vector c. This is in contrast to many of the above-mentioned differentiability results, where the initial function ϕ is treated as an infinite-dimensional parameter and differentiability is meant in the sense of Fréchet. For practical purposes, however, the restriction to finite-dimensional parameter vectors c is not an essential one, because numerical approximation of the Wronskian by using computers is anyway coupled to the need to find a finite-dimensional parametrization $\phi(t,c)$ of the initial function.

Organization of This Chapter

The basic idea of the proof of the differentiability theorem in Lenz, Schlöder, and Bock [173] is to apply the method of steps and to use differentiability theorems of ODE theory. Therefore, the first issue is to recall the ODE theory on differentiability of IVP solutions; this is done in Section 7.1. Section 7.2 recapitulates the differentiability theorem recently given Lenz, Schlöder, and Bock [173] and also reproduces the proof. Section 7.3 discusses how the ideas of the proof transfer to IHDDEs with time-dependent delay functions and (simple) time-dependent switching functions. Section 7.4 deals with IHODEs and recalls a differentiability result for this class of equations. Finally, Sections 7.5 and 7.6 show how the result for IHODEs can be transferred to DDEs and IHDDEs.

Notation

Several times in this chapter it is necessary to consider, for some reference interval $\hat{\mathcal{T}}$, an extension into both directions by some value Δa. For this purpose, the notation $\hat{\mathcal{T}}^{\Delta a}$ is used in this chapter as a symbol for the extended open interval

$$\hat{\mathcal{T}}^{\Delta a} := (a_1 - |\Delta a|, a_2 + |\Delta a|). \tag{7.2}$$

Further, for the special case that the lower boundary is $-\infty$, i.e. $\hat{\mathcal{T}} = (-\infty, a_2]$, then

$$\hat{\mathcal{T}}^{\Delta a} := (-\infty, a_2 + |\Delta a|). \tag{7.3}$$

For notational clarity, fraktal letters $\mathfrak{y}$ and $\mathfrak{w}$ are used in definitions of IVPs, and y and $\mathbf{W}$ are used as symbols for the solutions of the IVPs in this chapter.

7.1. Preliminaries: ODEs

A result for the differentiability of the solution $y(t;c)$ of an ODE-IVP as in Definition 1.4 with respect to parameters goes back to Gronwall [121], and is given in slightly different variants in various textbooks on ODEs, see e.g. Amann [5], page 126ff, and Hairer, Nørsett, and Wanner [126], page 92ff. Here, a version given in the book by Hartman [132], page 95f, is adapted to the notation of this thesis.

Theorem 7.1 (Local Differentiability of ODE-IVP Solutions)

Let an ODE-IVP (1.25) be given with nominal parameters $\tilde{c}$, open sets $\mathcal{V}^y \subset \mathbb{R}^{n_z}$, $\mathcal{V}^c \subset \mathbb{R}^{n_c}$ and an open interval $\mathcal{T}_0$ such that $(t^{ini}(\tilde{c}), y^{ini}(\tilde{c}), \tilde{c}) \in \mathcal{T}_0 \times \mathcal{V}^y \times \mathcal{V}^c$. Assume that the functions t^{ini} and y^{ini} are continuously differentiable for $c \in \mathcal{V}^c$ and that the right-hand-side function $f(t,y,c)$ is continuous in t and continuously differentiable with respect to y and c for $(t,y,c) \in \mathcal{T}_0 \times \mathcal{V}^y \times \mathcal{V}^c$.

Then the solution $y(t;c)$ is unique and has a continuous partial derivative $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$ on $(\omega_-(c), \omega_+(c)) \times \mathcal{V}^c$, where $(\omega_-(c), \omega_+(c))$ is the maximal interval of existence *i.e. the maximal interval for which $(t, y(t;c))$ stays in $\mathcal{T}_0 \times \mathcal{V}^y$.*

Proof
See Hartman [132], page 95f. ■

Theorem 7.1 is a differentiability result for the solution $y(t;c)$ that is local both in time and in the parameters because it holds only for a (possibly very small) open set $\mathcal{V}^c \supset \tilde{c}$ and a (possibly very small) maximal interval of existence $(\omega_-(c), \omega_+(c))$. Since all differentiability theorems in this chapter are local in the parameter space, this property is not emphasized and the attribute "local" in the name of the theorem therefore only refers to the fact that it is local in time. Accordingly, a "global" differentiability result in the sense of this chapter is a differentiability result on a given, parameter-dependent interval. More precisely, the interval of interest is $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$ in case of an IVPs without delays and $\mathcal{T}^f(c) = (-\infty, t^{fin}(c)]$ in the case of IVPs with delays.

As a first example of a global differentiability theorem, the following ODE result is formulated.

Theorem 7.2 (Global Differentiability of ODE-IVP Solutions)

Let an ODE-IVP (1.25) be given with nominal parameters $\tilde{c}$ and an open neighborhood $\mathcal{V}^c$ of $\tilde{c}$. Let $\mathcal{T}(\tilde{c}) = [t^{ini}(\tilde{c}), t^{fin}(\tilde{c})]$ be the interval for the nominal parameters, and choose $\Delta t > 0$ such that also the extended interval $\mathcal{T}^{\Delta t}(\tilde{c}) := (t^{ini}(\tilde{c}) - \Delta t, t^{fin}(\tilde{c}) + \Delta t)$ is defined. Further, choose $\Delta y > 0$ and define the open set

$$\mathcal{V}^y := \{y \mid \|y - y^{ini}(\tilde{c})\|_\infty < \Delta y\} \subset \mathbb{R}^{n_y}. \tag{7.4}$$

Assume that the model functions of the ODE-IVP (1.25) fulfill the following conditions:

(S) Smoothness*: The functions $t^{ini}(c)$, $y^{ini}(c)$, and $t^{fin}(c)$ are continuously differentiable for $c \in \mathcal{V}^c$ and the right-hand-side function $f(t,y,c)$ is continuous in t, uniformly Lipschitz continuous with respect to y, and continuously differentiable with respect to y and c for $(t,y,c) \in \mathcal{T}^{\Delta t}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{V}^c$.*

(B) Boundedness*: The right-hand-side function f is bounded by*

$$\|f(t,y,c)\|_\infty \leqslant M_f \tag{7.5a}$$

$$M_f \leqslant \frac{\Delta y}{2(t^{fin}(\tilde{c}) + \Delta t - t^{ini}(\tilde{c}))} \tag{7.5b}$$

for $(t,y,c) \in \mathcal{T}^{\Delta t}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{V}^c$.

Then there exists an open interval $\mathcal{T}^{\delta t}(\tilde{c}) := (t^{ini}(\tilde{c}) - \delta t, t^{fin}(\tilde{c}) + \delta t)$, $\delta t > 0$, and an open neighborhood $\mathcal{U}^c \subset \mathcal{V}^c$ of $\tilde{c}$ such that $\mathcal{T}(c) \subset \mathcal{T}^{\delta t}(\tilde{c})$ for $c \in \mathcal{U}^c$, and for $(t,c) \in \mathcal{T}^{\delta t}(\tilde{c}) \times \mathcal{U}^c$ the following assertions hold:

1. *A unique solution $y(t;c)$ exists.*
2. *The Wronskian $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$ is continuous.*
3. *The Wronskian $\mathbf{W}(t;c)$ is given by the solution of the problem*

$$\dot{\mathfrak{w}}(t;c) = \frac{\partial f(t,y(t;c),c)}{\partial y}\mathfrak{w}(t;c) + \frac{\partial f(t,y(t;c),c)}{\partial c} \tag{7.6a}$$

$$\mathfrak{w}(t^{ini}(c);c) = \frac{dy^{ini}(c)}{dc} - f^{ini}(c) \cdot \frac{dt^{ini}(c)}{dc}. \tag{7.6b}$$

with $f^{ini}(c) := f(t^{ini}(c), y(t^{ini}(c);c), c)$.

4. *The total derivative of the state at the final time is given by*

$$\frac{dy(t^{fin}(c);c)}{dc} = \mathbf{W}(t^{fin}(c);c) + f^{fin}(c) \cdot \frac{dt^{fin}(c)}{dc}, \tag{7.7}$$

with $f^{fin}(c) := f(t^{fin}(c), y(t^{fin}(c);c), c)$.

Before coming to the proof of the theorem, it is first remarked that for any (possibly parameter-dependent) time $t(c)$, the total derivative of the state, $dy(t(c);c)/dc$, can be expressed in terms of the partial derivative $\partial y(t(c);c)/\partial c$ as follows:

$$\frac{dy(t(c);c)}{dc} = \frac{\partial y(t(c);c)}{\partial c} + \frac{dy(t(c);c)}{dt}\frac{dt(c)}{dc}. \tag{7.8}$$

In particular, this relation holds for the initial time $t^{ini}(c)$, see equation (7.6b), and for the final time $t^{fin}(c)$, see equation (7.7). The Wronskian matrix, i.e. the partial derivative, represents the derivative of the state with respect to the parameters for a fixed, unchanged time point, and only the total derivative takes into account that also the evaluation time changes with the parameters.

Proof (of Theorem 7.2)
The theorem is essentially a combination of the local differentiability result in Theorem 7.1 with suitable additional assumptions that ensure the existence and uniqueness of a solution on $\mathcal{T}(c) \subset \mathcal{T}^{\delta t}(\tilde{c})$ for some neighborhood of the nominal parameters $\tilde{c}$. More precisely, the assumptions (S) and (B) of the theorem are restrictive enough so that the Picard-Lindelöf theorem (Theorem 4.1) can be applied to the parameter-dependent interval $\mathcal{T}(c)$ for arbitrary parameters c that are sufficiently close to the nominal parameters $\tilde{c}$. This proves assertion 1, and assertion 2 follows immediately from Theorem 7.1.

The representation of the Wronskian as solution of an initial value problem as given in assertion 3 is a standard result for ODEs, see e.g. Hairer, Nørsett, and Wanner [126], page 95 and Hartman [132], page 95. However, the textbooks usually formulate different initial value problems for derivatives with respect to the initial value, the initial time, and the parameters. In order to prove the variant given here, consider the integral representation of the ODE-IVP solution

$$y(t;c) = y^{ini}(c) + \int_{t^{ini}(c)}^{t} f(t', y(t';c), c)dt' \tag{7.9}$$

and take the derivative with respect to c. At this point it is relevant that the smoothness assumption (S) was formulated for an extended time interval $\mathcal{T}^{\Delta t}(\tilde{c})$, so that for some sufficiently small neighborhood $\mathcal{U}^c$ the initial and final time remain within a domain where (S) holds. This is crucial for the application of the differentiation rule for parameter-dependent integrals (see e.g. Bronstein et al. [49], page 475f), which yields

$$\begin{aligned}\frac{\partial y(t;c)}{\partial c} =& \frac{d}{dc}y^{ini}(c) - f(t^{ini}(c), y(t^{ini}(c);c), c)\frac{d}{dc}t^{ini}(c) \\ &+ \int_{t^{ini}(c)}^{t} \frac{\partial f(t', y(t';c), c)}{\partial y}\frac{\partial y(t';c)}{\partial c} + \frac{\partial f(t', y(t';c), c)}{\partial c}dt'. \end{aligned} \tag{7.10}$$

Differentiation with respect to t gives the differential equation (7.6a) for $\mathbf{W}$, and the first two terms in equation (7.10) are identified as the associated initial value, i.e. equation (7.6b).

Finally, assertion 4 is verified by doing the same analysis for the integral representation of $y(t^{fin}(c);c)$, i.e. replacing the upper integral boundary in equation (7.9) by $t^{fin}(c)$ and taking the total derivative with respect to the parameters, which also accounts for the parameter dependency of the final time. ■

One result of Theorem 7.2 is that the Wronskian is given as solution of an ODE-IVP. In order to distinguish between this ODE-IVP for $\mathbf{W}(t;c)$ and the original ODE-IVP for $y(t;c)$ the following terminology is used:

Definition 7.3 (Nominal IVP, Variational IVP)

The ODE-IVP for the state $y(t;c)$ is called nominal ODE-IVP*, and the initial value problem for the Wronskian (7.6) is called* variational ODE-IVP*. An analogous terminology is used later in the remainder of this thesis for initial value problems in HODEs, IODEs, IHODEs, DDEs, HDDEs, IDDEs, and IHDDEs.*

Note that the variational ODE-IVP is defined in such a way that the partial derivatives of the right-hand-side function are taken at $y(t;c)$, i.e. along the solution of the nominal ODE-IVP.

7.2. DDEs with Constant Delays

As a first generalization of ODE-IVPs, DDE-IVPs with constant delays are considered.

Definition 7.4 (Initial Value Problem in DDEs with Constant Delays)

An Initial Value Problem in DDEs with constant delays *for the state $\mathfrak{y} : \mathcal{T}^f(c) \to \mathcal{D}^y$ is given by*

$$\dot{\mathfrak{y}}(t;c) = f(t, \mathfrak{y}(t;c), c, \{\mathfrak{y}(t-\tau_i(c);c)\}_{i=1}^{n_\tau}) \quad \textit{for} \quad t \in \mathcal{T}(c) \tag{7.11a}$$

$$\mathfrak{y}(t^{ini}(c);c) = y^{ini}(c) \tag{7.11b}$$

$$\mathfrak{y}(t;c) = \phi(t,c) \quad \textit{for} \quad t < t^{ini}(c). \tag{7.11c}$$

All definitions of functions, intervals, and sets carry over from the Definitions 1.11 and 1.12 with the exception that $\tau_i : \mathcal{D}^c \to \mathbb{R}^+$ are here constant and positive delay functions of the parameters.

A global differentiability result for solutions $y(t;c)$ of DDE-IVPs with constant delays can be found in Lenz, Schlöder, and Bock [173], which is recalled in the following.

Theorem 7.5 (Global Differentiability of DDE-IVP Solutions)

Let a DDE-IVP with constant delays be given as in Definition 7.4. Let $\tilde{c}$ be the nominal parameters, and consider a neighborhood $\mathcal{V}^c \subset \mathbb{R}^{n_c}$ of $\tilde{c}$ and the interval $\mathcal{T}^{f,\Delta t}(\tilde{c})$ with an extension $\Delta t > 0$. Further, choose $\Delta y > 0$ and $\Delta\phi > 0$ for the definition of the open sets

$$\mathcal{V}^y := \{y \mid \|y - y^{ini}(\tilde{c})\|_\infty < \Delta y\} \subset \mathbb{R}^{n_y} \tag{7.12a}$$

$$\begin{aligned}\mathcal{V}^\phi := \{y \mid & \inf_{t\in\mathcal{T}^{f,\Delta t}(\tilde{c})} (\phi_i(t,\tilde{c})) - \Delta\phi < y_i < \sup_{t\in\mathcal{T}^{f,\Delta t}(\tilde{c})} (\phi_i(t,\tilde{c})) + \Delta\phi\} \subset \mathbb{R}^{n_y} \\ & \textit{for} \quad 1 \leqslant i \leqslant n_y\}\end{aligned} \tag{7.12b}$$

Let the following assumptions be fulfilled by the model functions of the DDE-IVP (7.11):

(S) Smoothness*: The functions $t^{ini}(c)$, $y^{ini}(c)$, $t^{fin}(c)$, and $\tau_k(c)$ are continuously differentiable for $c \in \mathcal{V}^c$. The right-hand-side function $f(t, y, c, \{v_k\}_{k=1}^{n_\tau})$ is continuous in t, uniformly Lipschitz-continuous with respect to y, and continuously differentiable with respect to y, c, and v_k, $1 \leqslant k \leqslant n_\tau$, for $(t, y, c, \{v_k\}_{k=1}^{n_\tau}) \in \mathcal{T}^{f,\Delta t}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{V}^c \times (\mathcal{V}^\phi \cup \mathcal{V}^y)^{n_\tau}$. The initial function $\phi(t, c)$ is continuously differentiable for $(t, c) \in \mathcal{T}^{f,\Delta t}(\tilde{c}) \times \mathcal{V}^c$.*

(B) Boundedness *f: The right-hand-side function f is bounded by*

$$\|f(t, y, c, \{v_k\}_{k=1}^{n_\tau})\|_\infty \leqslant M_f \tag{7.13a}$$

$$M_f \leqslant \frac{\Delta y}{2(t^{fin}(\tilde{c}) + \Delta t - t^{ini}(\tilde{c}))} \tag{7.13b}$$

for $(t, y, c, \{v_k\}_{k=1}^{n_\tau}) \in \mathcal{T}^{f,\Delta t}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{V}^c \times (\mathcal{V}^\phi \cup \mathcal{V}^y)^{n_\tau}$.

(D) Distinctness*: the values of the delays are pairwise distinct, i.e. $\tau_k(c) \neq \tau_j(c)$ for $j \neq k$ and for all $c \in \mathcal{V}^c$, and it holds that $\tau_k(c) \neq 0$ and $\tau_k(c) \neq t^{fin}(c) - t^{ini}(c)$ for $1 \leqslant k \leqslant n_\tau$, $c \in \mathcal{V}^c$.*

Without loss of generality the delays shall be ascendingly ordered, $\tau_k(c) < \tau_{k+1}(c)$ for $1 \leqslant k \leqslant n_\tau - 1$, and $K \in \{0, \ldots, n_\tau\}$ shall be the largest integer[1] such that $t^{ini}(c) + \tau_K(c) < t^{fin}(c)$ ($K = 0$ if the relation is violated for all delays). Further, define $s_k(c)$, $0 \leqslant k \leqslant K+1$ as follows: $s_0(c) = t^{ini}(c)$, $s_{K+1}(c) = t^{fin}(c)$, and, if $K \geqslant 1$, $s_k(c)$ for $1 \leqslant k \leqslant K$ as the parameter-dependent time points of the child discontinuities of the discontinuity at $t^{ini}(c)$ within the considered time interval, i.e. $s_k(c) = t^{ini}(c) + \tau_k(c)$.

Then there exists $\delta t > 0$ and an open neighborhood $\mathcal{U}^c$ of $\tilde{c}$, such that $\mathcal{T}^f(c) \subset \mathcal{T}^{f,\delta t}(\tilde{c})$ for $c \in \mathcal{U}^c$, and for $(t, c) \in \mathcal{T}^{f,\delta t}(\tilde{c}) \times \mathcal{U}^c$ the following assertions hold:

1. *A unique solution $y(t; c)$ of the nominal DDE-IVP exists.*

2. *The Wronskian $\mathbf{W}(t; c) = \partial y(t; c)/\partial c$ is continuous for $t \neq s_k(c)$, $0 \leqslant k \leqslant K$. At the times $t = s_k(c)$ for $0 \leqslant k \leqslant K$ the Wronskian is right-continuous.*

3. *At the times $s_k(c)$, $1 \leqslant k \leqslant K$, the Wronskian $\mathbf{W}(t; c)$ is generally discontinuous, and the difference between the left-sided limit and the right-sided limit is given by*

$$\mathbf{W}^+(s_k(c); c) - \mathbf{W}^-(s_k(c); c) = \left(f_k^-(c) - f_k^+(c)\right) \frac{ds_k(c)}{dc}. \tag{7.14}$$

Herein, $ds_k(c)/dc = dt^{ini}(c)/dc + d\tau_k(c)/dc$, and $f_k^-(c)$ and $f_k^+(c)$ are defined as the left-sided and right-sided limit of the evaluation of the right-hand-side function f, i.e.

$$f_k^-(c) := f(s_k(c), y^-(s_k(c); c), c, \{y^-(s_k(c) - \tau_i(c); c)\}_{i=1}^{n_\tau}) \tag{7.15a}$$

$$f_k^+(c) := f(s_k(c), y^+(s_k(c); c), c, \{y^+(s_k(c) - \tau_i(c); c)\}_{i=1}^{n_\tau}). \tag{7.15b}$$

These right-hand-side function evaluations differ in exactly one argument, namely in the value used for the state at the time point $s_k(c) - \tau_k(c) = t^{ini}(c)$. Once, the left-sided limit is taken (which gives $\phi(t^{ini}(c), c)$) and once the right-sided limit is taken (which gives $y(t^{ini}(c); c)$). Further, it holds that $s_k(c) - \tau_i(c) \neq t^{ini}(c)$ for $i \neq k$, and hence $y^-(s_k(c) - \tau_i(c); c) = y^+(s_k(c) - \tau_i(c); c)$.

4. *On the right-open interval $[s_k(c), s_{k+1}(c))$, $0 \leqslant k \leqslant K$, the Wronskian $\mathbf{W}(t; c)$ is given by the solution of the variational DDE-IVP*

$$\begin{aligned}\dot{\mathfrak{W}}(t; c) = &\frac{\partial f(t, y(t; c), c, \{y(t - \tau_i(c); c)\}_{i=1}^{n_\tau})}{\partial y} \mathfrak{W}(t; c) + \frac{\partial f(t, y(t; c), c, \{y(t - \tau_i(c); c)\}_{i=1}^{n_\tau})}{\partial c} \\ &+ \sum_{m=1}^{n_\tau} \frac{\partial f(t, y(t; c), c, \{y(t - \tau_i(c); c)\}_{i=1}^{n_\tau})}{\partial v_m} \\ &\qquad \cdot \left[\mathfrak{W}(t - \tau_m(c); c) - \dot{y}(t - \tau_m(c); c) \frac{d\tau_m(c)}{dc}\right]\Bigg)\end{aligned} \tag{7.16a}$$

[1] Note that by the Distinctness assumption (D) and the Smoothness assumption (S) on the delays $\tau_k(c)$ the total number K of discontinuities within $[t^{ini}(c), t^{fin}(c)]$ (and their order) is the same for all $c \in \mathcal{V}^c$.

$$\mathfrak{w}(s_k(c);c) = \frac{dy_k(c)}{dc} - f_k^+(c)\frac{ds_k(c)}{dc} \tag{7.16b}$$

$$\mathfrak{w}(t;c) = \begin{cases} \frac{\partial \phi(t,c)}{\partial c} & \text{for} \quad t < s_0(c), \\ \text{solution of problem (7.16) in} & [s_i(c), s_{i+1}(c)) \\ & \text{for} \quad s_i(c) \leqslant t < s_{i+1}(c), i \in \{0,\ldots,k\} \end{cases} . \tag{7.16c}$$

Herein, $\partial f/\partial v_m$ denotes the partial derivative of the right-hand-side function with respect to the m-th past state, and $f_k^+(c)$ is given by (7.15b) for $1 \leqslant k \leqslant K$, and for $k = 0$ it is given by

$$f_0^+(c) := f^{ini\,+}(c) := f(t^{ini}(c), y^{ini}(c), c, \{y(t^{ini}(c) - \tau_i(c); c)\}_{i=1}^{n_\tau}). \tag{7.17}$$

Further, $y_k(c)$ denotes the initial state for the considered interval, which is given by $y_0(c) = y^{ini}(c)$ for $k = 0$, and by $y_k(c) = y^-(s_k(c);c)$ for $1 \leqslant k \leqslant K$, i.e. by the solution of the nominal DDE-IVP at the end of the preceding interval $[s_{k-1}(c), s_k(c))$. The total derivative of this state, which occurs in equation (7.16b), is given by

$$\frac{dy_k(c)}{dc} = \mathbf{W}^-(s_k(c);c) + f_k^-(c)\frac{ds_k(c)}{dc}, \tag{7.18}$$

so that, by using equation (7.14), equation (7.16b) is equivalent to

$$\mathfrak{w}(s_k(c);c) = \mathbf{W}^+(s_k(c);c). \tag{7.19}$$

5. *The total derivative of the state at the final time is given by*

$$dy(t^{fin}(c);c)/dc = \mathbf{W}(t^{fin}(c);c) + f_{K+1}^-(c) \cdot \frac{dt^{fin}(c)}{dc}, \tag{7.20}$$

with

$$f_{K+1}^-(c) := f^{fin\,-}(c) = f(t^{fin}(c), y^-(t^{fin}(c);c), c, \{y^-(t^{fin}(c) - \tau_i(c);c)\}_{i=1}^{n_\tau}). \tag{7.21}$$

Before giving the proof, it is remarked that according to assumption (S) of the theorem the initial function ϕ is continuously differentiable rather than just piecewise continuous differentiable as it was assumed in Chapters 4 and 5. This restriction is made here in order to reduce the notational complexity, because with this assumption the solution $y(t;c)$ may have only one discontinuity of order 0, namely at the initial time $t^{ini}(c)$. Accordingly, discontinuities of order 1 in y within $\mathcal{T}(c)$ may occur only at the time points $s_k(c) = t^{ini}(c) + \tau_k(c)$, $1 \leqslant k \leqslant K$. Once the proof has been completed, it becomes obvious how to generalize the differentiability result for the more general case of piecewise continuously differentiable initial functions; this generalization is discussed in Section 7.3.

It should further be emphasized that the idea of the proof is very simple. At first, the interval $\mathcal{T}^{f,\delta t}(\tilde{c})$ is subdivided as follows: $\mathcal{T}^{f,\delta t}(\tilde{c}) := \bigcup_{k=0}^{K+2} \mathcal{T}_k(c)$, where $\mathcal{T}_0(c) := (-\infty, t^{ini}(c))$, $\mathcal{T}_k(c) := [s_{k-1}(c), s_k(c))$ for $1 \leqslant k \leqslant K+1$, and $\mathcal{T}_{K+2}(c) := [t^{fin}(c), t^{fin}(\tilde{c}) + \delta t)$. The essential tools for the proof is then to reduce the DDE-IVP to a sequence of ODE-IVPs on these subintervals and to apply Theorem 7.2 on the differentiability of ODE-IVP solutions. The main technical difficulty of the proof is to define the intervals and sets in such a way that the constructed ODE-IVPs do indeed fulfill the assumptions of Theorem 7.2.

Proof ((of Theorem 7.5))
The assertions of the theorem are formulated for $(t,c) \in \mathcal{T}^{f,\delta t}(\tilde{c}) \times \mathcal{U}^c$, for some $\delta t > 0$ and for a neighborhood $\mathcal{U}^c$ of $\tilde{c}$. Consider $\delta t_1 = \Delta t/2$ as a proposition for δt and the set $\mathcal{U}_1^c \subset \mathcal{V}^c$ as a proposition for $\mathcal{U}^c$, where $\mathcal{U}_1^c$ is chosen sufficiently small so that for all elements $c \in \mathcal{U}_1^c$ the following conditions hold:

(I) $|t^{ini}(c) - t^{ini}(\tilde{c})| < \delta t_1/2$ and $|t^{fin}(c) - t^{fin}(\tilde{c})| < \delta t_1/2$

(II) $|\tau_k(c) - \tau_k(\tilde{c})| < \delta t_1/2$ for $1 \leqslant k \leqslant n_\tau$

(III) $\|y^{ini}(c) - y^{ini}(\tilde{c})\|_\infty < \Delta y/4$

(IV) $\phi(t,c) \in \mathcal{V}^\phi$ for $t \in \mathcal{T}^{f,\Delta t}(\tilde{c})$.

A neighborhood $\mathcal{U}_1^c$ of $\tilde{c}$, where these conditions are fulfilled can always be found because of the Smoothness assumption (S) for $t^{ini}(c)$, $t^{fin}(c)$, $\tau_k(c)$, $y^{ini}(c)$, and $\phi(t,c)$. The final choices for δt, $0 < \delta t \leqslant \delta t_1$ and $\mathcal{U}^c \subset \mathcal{U}_1^c$ are given later in the proof.

Proof for $t \in \mathcal{T}_0(c)$
On the first interval, $\mathcal{T}_0(c)$, the proof is trivial, because the assertions 1 and 2 follow from the assumptions on the initial function ϕ, regardless of the specific choices of δt, $0 < \delta t \leqslant \delta t_1$, and $\mathcal{U}^c \subset \mathcal{U}_1^c$. The other assertions make no statement for $t < t^{ini}(c)$.

Proof for $t \in \mathcal{T}_1(c)$ (for $K \geqslant 1$)
On the right-open interval $\mathcal{T}_1(c) = [s_0(c), s_1(c))$, all deviating arguments $t - \tau_k(c)$, $1 \leqslant k \leqslant n_\tau$, assume values to the left of $t^{ini}(c)$, thus all past states are given by evaluations of the initial function ϕ. On this interval, the DDE-IVP is therefore equivalent to an ODE-IVP with initial time $s_0(c)$, final time $s_1(c)$, initial value $y_0(c) = y^{ini}(c)$, and right-hand-side function

$$f_{ODE}(t, \mathfrak{y}(t;c), c) := f(t, \mathfrak{y}(t;c), c, \{\phi(t - \tau_k(c), c)\}_{k=1}^{n_\tau}). \tag{7.22}$$

In order to apply Theorem 7.2 to the equivalent ODE-IVP, the assumptions (S) and (B) need to be verified. It is immediately clear that the initial time, the final time, and the initial value fulfill the assumption (S) of Theorem 7.2 for all $c \in \mathcal{U}_1^c$, so that it remains only to check the assumptions on the right-hand-side function f_{ODE}. Obviously, for this ODE right-hand-side function f_{ODE} the assumptions (S) and (B) of Theorem 7.2 hold if all arguments of the DDE right-hand-side function f remain in the domain where (S), (B) are assumed in Theorem 7.5.

This is verified for $(t,y,c) \in \mathcal{T}_1^{\delta t_1}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{U}_1^c$. Clearly, because of conditions (I) and (II), it holds that $\mathcal{T}_1(c) \subset \mathcal{T}_1^{\delta t_1}(\tilde{c})$ for all $c \in \mathcal{U}_1^c$. Furthermore, condition (II) ensures that the deviating arguments are such that the relation

$$t - \tau_k(c) \in \left(t^{ini}(\tilde{c}) - \delta t_1 - \tau_{n_\tau}(\tilde{c}) - \frac{\delta t_1}{2}, \quad s_1(\tilde{c}) + \delta t_1 - \tau_1(\tilde{c}) + \frac{\delta t_1}{2} \right) \subset \mathcal{T}^{f,\Delta t}(\tilde{c}) \tag{7.23}$$

is fulfilled. Hence, the past time points given by the deviating arguments are in a domain where the initial function fulfills (S).[2] Because of condition (IV), the past states obtained from evaluations of the initial function are in $\mathcal{V}^\phi$. Eventually, because of the bound M_f given in assumption (B) and because of condition (III), the current state remains in $\mathcal{V}^y$. In summary, all arguments of the DDE right-hand-side function f are in domains where (S) and (B) were assumed in Theorem 7.5, so that the ODE right-hand-side function f_{ODE} defined in equation (7.22) does indeed fulfill the asssumptions (S) and (B) of Theorem 7.2.

Hence, Theorem 7.2 can be applied, which gives the existence of some $\bar{\delta}t_1 > 0$ and a neighborhood $\bar{\mathcal{U}}_1^c \subset \mathcal{U}_1^c$ of $\tilde{c}$, such that for $(t,c) \in \mathcal{T}_1^{\bar{\delta}t_1}(\tilde{c}) \times \bar{\mathcal{U}}_1^c$ there exists a unique solution $y(t;c)$ of the ODE-IVP, and the Wronskian $\mathbf{W}(t;c)$ is continuous. For the representation of the Wronskian as solution of a variational ODE-IVP, it is taken into account that the partial derivative of the right-hand-side function with respect to the parameters in the variational ODE, $\partial f_{ODE}(t, y(t;c), c)/\partial c$, also has to incorporate the parameter dependencies of the function f through the delay argument and the initial function, see equation (7.22). This yields that $\mathbf{W}(t;c)$ is the solution of the variational ODE-IVP

$$\begin{aligned}\dot{\mathfrak{w}}(t;c) = &\frac{\partial f(t, y(t;c), c, \{\phi(t-\tau_i(c);c)\}_{i=1}^{n_\tau})}{\partial y}\mathfrak{w}(t;c) + \frac{\partial f(t, y(t;c), c, \{\phi(t-\tau_i(c);c)\}_{i=1}^{n_\tau})}{\partial c} \\ &+ \sum_{m=1}^{n_\tau} \frac{\partial f(t, y(t;c), c, \{\phi(t-\tau_i(c);c)\}_{i=1}^{n_\tau})}{\partial v_m} \\ &\qquad \cdot \left[\frac{\partial \phi(t-\tau_m(c);c)}{\partial c} - \dot{\phi}(t-\tau_m(c);c)\frac{d\tau_m(c)}{dc} \right]\end{aligned} \tag{7.24a}$$

[2] Note that in equation (7.23) the lower bound depends on $\tau_{n_\tau}(\tilde{c})$, which is the largest of the delays because they are assumed to be ascendingly sorted.

$$\mathfrak{W}(t^{ini}(c);c) = \frac{dy_0(c)}{dc} - f(s_0(c), y_0(c), c, \{\phi(s_0(c)-\tau_i(c);c)\}_{i=1}^{n_\tau})\frac{ds_0(c)}{dc}. \tag{7.24b}$$

Since the constructed ODE-IVP and the nominal DDE-IVP are equivalent on the interval $\mathcal{T}_1(c)$, the conclusions regarding existence, uniqueness and differentiability of solutions carry over to the DDE-IVP (assertions 1 and 2 of Theorem 7.5). Assertion 4 is verified by replacing, in the variational ODE-IVP (7.24), $\phi(t-\tau_m(c);c) \to y(t-\tau_m(c);c)$, $\partial\phi(t-\tau_m(c);c)/\partial c \to \mathfrak{W}(t-\tau_m(c);c)$, and adding the equation $\mathfrak{W}(t;c) = \partial\phi(t,c)/\partial c$ for $t < t^{ini}(c)$ to the system. The second and the third of these replacements actually turn the variational ODE-IVP into a variational DDE-IVP.

Finally, assertion 3 makes a statement on the difference between left-sided and right-sided value of the Wronskian $\mathbf{W}(t;c)$ at the time points $s_k(c)$, $1 \leqslant k \leqslant K$. In order to prepare for the proof of this assertion at $s_1(c)$ in the next paragraph, it is observed that the left-sided limit $\mathbf{W}^-(s_1(c);c)$ is simply given by the left-sided limit of the solution $\mathbf{W}(t;c)$ of the variational DDE-IVP, and that the total derivative of $y_1(c) := y^-(s_1(c);c)$ is given by

$$\begin{aligned}\frac{dy_1(c)}{dc} &= \frac{dy^-(s_1(c);c)}{dc} \\ &= \mathbf{W}^-(s_1(c);c) + f_1^-(c)\frac{ds_1(c)}{dc}.\end{aligned} \tag{7.25}$$

Proof for $t \in \mathcal{T}_2(c)$ (for $K \geqslant 2$)
On the right-open interval $\mathcal{T}_2(c) = [s_1(c), s_2(c))$ the deviating arguments $t - \tau_k(c)$ for $2 \leqslant k \leqslant n_\tau$ assume values that are to the left of $t^{ini}(c)$. The corresponding past states are therefore given by evaluations of the initial function ϕ. Further, the deviating argument $t - \tau_1(c)$ assumes values to the right of $t^{ini}(c)$, and it may or may not happen that $t - \tau_1(c) \geqslant s_1(c)$ for some $t \in \mathcal{T}_2(c)$. For simplicity, it is assumed here that the following auxiliary assumption holds:

(A) Auxiliary Assumption: $t - \tau_1(c) < s_1(c)$ for $(t,c) \in \mathcal{T}^{f,\Delta t}(\tilde{c}) \times \mathcal{V}^c$.

Assumption (A) implies that $\tau_k(c) < 2\tau_1(c)$ for $2 \leqslant k \leqslant K$, so that the propagation of the discontinuity at $t^{ini}(c)$ with the delays $\tau_k(c)$, $2 \leqslant k \leqslant K$, occurs before the propagation of the discontinuity at $s_1(c)$ with delay $\tau_1(c)$. Arguments for Assumption (A) to be dispensable are given later.

The next step is to replace, on the interval $\mathcal{T}_2(c)$, the DDE-IVP by an equivalent ODE-IVP. The initial time of this ODE-IVP is $s_1(c)$, and the final time is $s_2(c)$. As initial value, the left-sided limit of the ODE-IVP solution on the preceding interval is taken, $y_1(c) = y^-(s_1(c);c)$. This is in accordance with the continuity property of DDE-IVP solutions for $t \in \mathcal{D}_1^t(\mathcal{T}(c)) = \mathcal{T}(c)$, see Definition 2.5. Finally, the right-hand-side function of the equivalent ODE-IVP is defined by

$$f_{ODE}(t, \mathfrak{y}(t;c), c) := f(t, \mathfrak{y}(t;c), c, y(t-\tau_1(c);c), \{\phi(t-\tau_k(c),c)\}_{k=2}^{n_\tau}). \tag{7.26}$$

The past state corresponding to the first deviating argument, $y(t-\tau_1(c);c)$, shall be given by the solution of the ODE-IVP on the preceding interval. The past states corresponding to the other deviating arguments are given by evaluations of the initial function ϕ.

In order to apply Theorem 7.2 to this ODE-IVP, it is again required to verify the assumptions (S) and (B) on the model functions. The initial time and the final time are given as sums of the initial time and delay functions, and hence as sums of differentiable functions, so they fulfill (S). The initial value, $y_1(c)$, is given as the left-sided limit of the ODE-IVP solution on the preceding interval at $s_1(c)$, i.e. $y^-(s_1(c);c)$. The differentiability of this value follows from recalling the equivalence of the DDE-IVP to an ODE-IVP on the interval $[s_0(c), s_1(c))$ and the fact that the total derivative of the final value of the ODE-IVP solution exists for $c \in \bar{\mathcal{U}}_1^c$ by assertion 4 of Theorem 7.2.

It remains to be checked that the constructed ODE right-hand-side function f_{ODE} in equation (7.26) fulfills the assumptions (S) and (B) of Theorem 7.2. For this purpose, consider $(t,c) \in \mathcal{T}_2^{\delta t_2}(\tilde{c}) \times \mathcal{U}_2^c$, with $\delta t_2 = \bar{\delta} t_1/2$ and a neighborhood $\mathcal{U}_2^c$ of $\tilde{c}$ that is sufficiently small such that the following modified versions of conditions (I) and (II) hold:

(I') $|t^{ini}(c) - t^{ini}(\tilde{c})| < \delta t_2/2$

(II') $|\tau_k(c) - \tau_k(\tilde{c})| < \delta t_2/2$, for $1 \leqslant k \leqslant n_\tau$.

These two conditions ensure, on the one hand, that for $c \in \mathcal{U}_2^c$ the parameter-dependent interval fulfills the relation $\mathcal{T}_2(c) \subset \mathcal{T}_2^{\delta t_2}(\tilde{c})$, i.e. it remains within the extension around the nominal interval $\mathcal{T}_2(\tilde{c})$. On the other hand, the conditions also ensure that for $(t,c) \in \mathcal{T}_2^{\delta t_2}(\tilde{c}) \times \mathcal{U}_2^c$ it holds that

$$t - \tau_1(c) \in \left(s_1(\tilde{c}) - \delta t_2 - \tau_1(\tilde{c}) - \frac{\delta t_2}{2}, \quad s_2(\tilde{c}) + \delta t_2 - \tau_1(\tilde{c}) + \frac{\delta t_2}{2} \right) \subset \mathcal{T}_1^{2\delta t_2}(\tilde{c}) \subset \mathcal{T}_1^{\bar{\delta} t_1}(\tilde{c}). \quad (7.27)$$

Hence, the first deviating argument varies only within an interval where the ODE-IVP solution on the preceding interval is differentiable. Furthermore, all other deviating arguments remain in $\mathcal{T}^{f,\Delta t}(\tilde{c})$, where the initial function ϕ is differentiable. Since also the current state stays, due to the bound M_f on the DDE right-hand-side function f, within the set $\mathcal{V}^y$, the constructed ODE right-hand-side function f_{ODE} does indeed fulfill the assumptions (S) and (B) of Theorem 7.2.[3]

It is now possible to apply Theorem 7.2, which gives existence, uniqueness, and continuous partial differentiability of the ODE-IVP solution $y(t;c)$ for $(t,c) \in \mathcal{T}_2^{\bar{\delta} t_2}(\tilde{c}) \times \bar{\mathcal{U}}_2^c$, with $\bar{\delta} t_2 > 0$ and with $\bar{\mathcal{U}}_2^c \subset \mathcal{U}_2^c$ being some neighborhood of $\tilde{c}$. The continuous partial derivative $\mathbf{W}(t;c)$ is thereby given as solution of the variational ODE-IVP

$$\begin{aligned}
\dot{\mathfrak{W}}(t;c) =& \frac{\partial f(t, y(t;c), c, y(t-\tau_1(c);c), \{\phi(t-\tau_i(c),c)\}_{i=2}^{n_\tau})}{\partial y} \mathfrak{W}(t;c) \\
&+ \frac{\partial f(t, y(t;c), c, y(t-\tau_1(c);c), \{\phi(t-\tau_i(c),c)\}_{i=2}^{n_\tau})}{\partial c} \\
&+ \frac{\partial f(t, y(t;c), c, y(t-\tau_1(c);c), \{\phi(t-\tau_i(c),c)\}_{i=2}^{n_\tau})}{\partial v_1} \\
&\qquad \cdot \left[\frac{\partial y(t-\tau_1(c);c)}{\partial c} - \dot{y}(t-\tau_1(c);c) \frac{d\tau_1(c)}{dc} \right] \\
&+ \sum_{m=2}^{n_\tau} \frac{\partial f(t, y(t;c), c, y(t-\tau_1(c);c), \{\phi(t-\tau_i(c),c)\}_{i=2}^{n_\tau})}{\partial v_m} \\
&\qquad \cdot \left[\frac{\partial \phi(t-\tau_m(c),c)}{\partial c} - \dot{\phi}(t-\tau_m(c),c) \frac{d\tau_m(c)}{dc} \right] \quad (7.28a)
\end{aligned}$$

$$\mathfrak{W}(s_1(c);c) = \frac{dy_1(c)}{dc} - f(s_1(c), y(s_1(c);c), c, y^{ini}(c), \{\phi(s_1(c)-\tau_i(c),c)\}_{i=2}^{n_\tau}) \frac{ds_1(c)}{dc}. \quad (7.28b)$$

Since the constructed ODE-IVP and the original DDE-IVP are equivalent on $\mathcal{T}_2(c)$, the DDE-IVP solution exists and is unique for $t \in \mathcal{T}_2(c)$ (assertion 1), and it has a continuous partial derivative with respect to the parameters (assertion 2). By replacing, in equation (7.28), $\partial y(t-\tau_1(c);c)/\partial c \to \mathfrak{W}(t-\tau_1(c);c)$, $\partial \phi(t-\tau_m(c),c)/\partial c \to \mathfrak{W}(t-\tau_m(c);c)$, $\phi(t-\tau_i(c),c) \to y(t-\tau_i(c);c)$, and adding the equation

$$\mathfrak{W}(t;c) = \begin{cases} \frac{\partial \phi(t,c)}{\partial c} \quad \text{for} \quad t < s_0(c) \\ \text{solution of the variational DDE-IVP} \quad \text{on} \quad t \in [s_0(c), s_1(c)) \quad \text{for} \quad t \in [s_0(c), s_1(c)) \end{cases} \quad (7.29)$$

to the system, assertion 4 on the expression of $\mathbf{W}(t;c)$ as solution of a variational DDE-IVP is verified. Finally, the right-sided limit of the Wronskian at $s_1(c)$, $\mathbf{W}^+(s_1(c);c)$, is given by the initial value in equation (7.28b). By using equation (7.25), this becomes

$$\mathbf{W}^+(s_1(c);c) = \mathbf{W}^-(s_1(c);c) + (f_1^-(c) - f_1^+(c)) \frac{ds_1(c)}{dc}, \quad (7.30)$$

with

$$f_1^-(c) = f(s_1(c), y(s_1(c);c), c, \phi(t^{ini}(c);c), \{\phi(s_1(c)-\tau_i(c);c)\}_{i=2}^{n_\tau}) \quad (7.31a)$$

$$f_1^+(c) = f(s_1(c), y(s_1(c);c), c, y^{ini}(c), \{\phi(s_1(c)-\tau_i(c);c)\}_{i=2}^{n_\tau}), \quad (7.31b)$$

which are equal in all arguments except for the first past state argument. This verifies assertion 3.

[3] The purpose of the bound in assumption (B) of the ODE Theorem 7.2 is simply to ensure that the current state remains within the domain where the assumptions are formulated, and the fact that this is the case can easily be concluded from the corresponding assumption on the DDE right-hand-side function f.

Proof for $t \in \mathcal{T}_k(c)$, $2 < k < K+2$
For the intervals $\mathcal{T}_3(c), \ldots, \mathcal{T}_{K+1}(c)$, the arguments of the proof on the preceding intervals can be repeated. For each of these intervals, an equivalent ODE-IVP can be formulated, and suitable extended intervals $\mathcal{T}_k^{\delta t_k}(\tilde{c})$ and neighborhoods $\mathcal{U}_k^c$ of $\tilde{c}$ can be found so that $\mathcal{T}_k(c) \subset \mathcal{T}_k^{\delta t_k}(\tilde{c})$, and the assumptions (S) and (B) hold for the model functions of the constructed ODE-IVPs. Hence, existence, uniqueness and differentiability of the solutions $y(t;c)$ of the ODE-IVPs is obtained for $(t,c) \in \mathcal{T}_k^{\bar{\delta t}_k}(\tilde{c}) \times \bar{\mathcal{U}}_k^c$, and it is concluded that the Wronskian $\mathbf{W}(t;c)$ can be expressed as solution of a variational ODE-IVP.

The equivalence of ODE-IVP and DDE-IVP on the interval $\mathcal{T}_k(c)$ gives the assertions 1, 2, and 4 for the DDE-IVP solution; in particular, when using the equivalence between ODE-IVP and DDE-IVP, an appropriate change of notation turns the variational ODE-IVP into a variational DDE-IVP for the Wronskian. Further, the jump in the Wronskian $\mathbf{W}(t;c)$ at the time points $s_k(c)$ is derived from

$$\begin{aligned}\mathbf{W}^+(s_k(c);c) &= \frac{dy_k(c)}{dc} - f_k^+(c)\frac{ds_k(c)}{dc} \\ &= \mathbf{W}^-(s_k(c);c) + f_k^-(c)\frac{ds_k(c)}{dc} - f_k^+(c)\frac{ds_k(c)}{dc},\end{aligned} \tag{7.32}$$

which verifies assertion 3.

Proof for $t \in \mathcal{T}_{K+2}(c)$
On the preceding interval $\mathcal{T}_{K+1}(c) := [s_K(c), t^{fin}(c)]$, differentiability of the solution $y(t;c)$ of the ODE-IVP is obtained for $(t,c) \in \mathcal{T}_{K+1}^{\bar{\delta t}_{K+1}}(\tilde{c}) \times \bar{\mathcal{U}}_{K+1}^c$. For the last interval

$$\mathcal{T}_{K+2}(c) := [t^{fin}(c), t^{fin}(\tilde{c}) + \delta t], \tag{7.33}$$

δt is defined by

$$\delta t := \begin{cases} \min(\bar{\delta t}_{K+1}, \quad \frac{1}{2}\cdot(t_0(\tilde{c}) + \tau_{K+1}(\tilde{c}) - t_f(\tilde{c}))) & \text{if} \quad K < n_\tau \\ \bar{\delta t}_{K+1} & \text{if} \quad K = n_\tau. \end{cases} \tag{7.34}$$

This means that δt is chosen small enough such that for the nominal parameters $\tilde{c}$ there is no child discontinuity of $t^{ini}(\tilde{c})$ within $[t^{fin}(\tilde{c}), t^{fin}(\tilde{c}) + 2\delta t)$. Further, the set $\mathcal{U}_{K+2}^c \subset \bar{\mathcal{U}}_{K+1}^c$ is chosen such that the following conditions are fulfilled for all $c \in \mathcal{U}_{K+2}^c$:

(I') $|t^{fin}(c) - t^{fin}(\tilde{c})| < \delta t/2$.

(II') $|\tau_k(c) - \tau_k(\tilde{c})| < \delta t/2$ for $1 \leqslant k \leqslant n_\tau$.

Then an equivalent ODE-IVP is considered for $(t,c) \in \mathcal{T}_{K+2}^{\delta t_{K+2}}(\tilde{c}) \times \mathcal{U}_{K+2}^c$, with $\delta t_{K+2} = \delta t/2$. This ensures that $\mathcal{T}_{K+2}(c) \subset \mathcal{T}_{K+2}^{\delta t_{K+2}}(\tilde{c})$, and the deviating arguments are located in those intervals where differentiability of the solutions of the ODE-IVPs was proven on the preceding intervals. By application of Theorem 7.2 and equivalence of the ODE-IVP solution and the DDE-IVP solution on $\mathcal{T}_{K+2}(c)$ the assertions 1, 2, and 4 of Theorem 7.5 follow for $(t,c) \in \mathcal{T}_{K+2}(c) \times \bar{\mathcal{U}}_{K+2}^c$. Finally, set the neighborhood of $\tilde{c}$ for which Theorem 7.5 is formulated to $\mathcal{U}^c = \bar{\mathcal{U}}_{K+2}^c$.

Assertion 5
Assertion 5 concerns the total derivative of the state at the final time $t^{fin}(c)$. For the proof, simply recall the definition of the ODE-IVP on the interval $\mathcal{T}_{K+1}(c)$ and apply assertion 4 of Theorem 7.2.

Dropping the Auxiliary Assumption (A)
The effect of the Auxiliary Assumption (A) is two-fold: first, it ensures that on all intervals $\mathcal{T}_k(c)$ for $1 \leqslant k \leqslant K+2$, all deviating arguments are located in some previous interval $\mathcal{T}_l(c)$, $l < k$, for which existence, uniqueness, and differentiability of the DDE-IVP solution have been proven before. The second effect is that no children of the first order discontinuities at $s_k(c) = t^{ini}(c) + \tau_k(c)$, $1 \leqslant k \leqslant K$, occur within the considered time interval.

Consider a simple case where (A) is violated, namely that the discontinuity at $s_1(c)$ has a child discontinuity at $s_1(c) + \tau_1(c) =: \tilde{s}(c) \in \mathcal{T}_2(c) = (s_1(c), s_2(c))$, because of the propagation with the

delay $\tau_1(c)$. In order to treat this situation, sub-divide the interval $\mathcal{T}_2(c)$ into $\mathcal{T}_{2,A}(c) = [s_1(c), \tilde{s}(c))$ and $\mathcal{T}_{2,B}(c) = [\tilde{s}(c), s_2(c))$, and consider, on these intervals, equivalent ODE-IVPs with right-hand-side functions defined by

$$f_{ODE,A}(t, \mathfrak{y}(t;c), c) = f(t, \mathfrak{y}(t;c), c, y_1(t-\tau_1(c);c), \{\phi(t-\tau_k(c),c)\}_{k=2}^{n_\tau}) \quad \text{for} \quad t \in \mathcal{T}_{2,A}(c), \tag{7.35a}$$

$$f_{ODE,B}(t, \mathfrak{y}(t;c), c) = f(t, \mathfrak{y}(t;c), c, y_{2,A}(t-\tau_1(c);c), \{\phi(t-\tau_k(c),c)\}_{k=2}^{n_\tau}) \text{ for} \quad t \in \mathcal{T}_{2,B}(c) \tag{7.35b}$$

where for notational clarity the ODE-IVP solution on $\mathcal{T}_1^{\bar{\delta}t_1}(\tilde{c})$ is here denoted by y_1 and the ODE-IVP solution on $\mathcal{T}_{2,A}^{\bar{\delta}t_{2,A}}(\tilde{c})$ is denoted by $y_{2,A}$. Further, the ODE-IVP solution on $\mathcal{T}_{2,B}^{\bar{\delta}t_{2,B}}(\tilde{c})$ is denoted by $y_{2,B}$; hereby, $\bar{\delta}t_{2,A} > 0$ and $\bar{\delta}t_{2,B} > 0$ are suitably chosen values such that the ODE-IVP solutions $y_{2,A}$ and $y_{2,B}$ are differentiable (see below) on the corresponding intervals.

By transferring the arguments that were used before on each of the intervals $\mathcal{T}_k(c)$, $1 \leqslant k \leqslant K+2$, to the two subintervals $\mathcal{T}_{2,A}(c)$ and $\mathcal{T}_{2,B}(c)$, it is easy to see that under a suitable reduction of the neighborhood of $\tilde{c}$, the ODE-IVP solutions $y_{2,A}$ and $y_{2,B}$ exist, are unique, and continuously differentiable with respect to the parameters on the above-used intervals $\mathcal{T}_{2,A}^{\bar{\delta}t_{2,A}}(\tilde{c})$ and $\mathcal{T}_{2,B}^{\bar{\delta}t_{2,B}}(\tilde{c})$. Further, the Wronskian is given as solution of a variational ODE-IVP. At $\tilde{s}(c)$, it holds for the Wronskian matrix $\mathbf{W}(t;c)$ that

$$\mathbf{W}^+(\tilde{s}(c);c) = \mathbf{W}^-(\tilde{s}(c);c) + (\tilde{f}^-(c) - \tilde{f}^+(c))\frac{d\tilde{s}(c)}{dc}, \tag{7.36}$$

with

$$\tilde{f}^-(c) = f(\tilde{s}(c), y^-(\tilde{s}(c);c), c, y_1^-(s_1(c);c), \{\phi^-(\tilde{s}(c)-\tau_i(c);c)\}_{i=2}^{n_\tau}) \tag{7.37a}$$

$$\tilde{f}^+(c) = f(\tilde{s}(c), y^+(\tilde{s}(c);c), c, y_{2,A}^+(s_1(c);c), \{\phi^+(\tilde{s}(c)-\tau_i(c);c)\}_{i=2}^{n_\tau}), \tag{7.37b}$$

being the left-sided and right-sided evaluation of f at the time point $\tilde{s}(c)$. Note that the current state, as well as all past states at $\tilde{s}(c)-\tau_i(c)$, $i \geqslant 2$, are continuous. Furthermore, also the first past state is continuous, because y is continuous at $\tilde{s}(c) - \tau_1(c) = s_1(c)$. It follows that $\tilde{f}^-(c) = \tilde{f}^+(c)$ and thus $\mathbf{W}^+(\tilde{s}(c);c) = \mathbf{W}^-(\tilde{s}(c);c)$, i.e. the Wronskian matrix is continuous in $\tilde{s}(c)$.

Due to the equivalence of the DDE-IVP to the ODE-IVPs on the interval $\mathcal{T}_{2,A}$ and $\mathcal{T}_{2,B}$, it is concluded that the DDE-IVP solution exists, is unique, and continuously partially differentiable with respect to the parameters. Further, by an appropriate change of notation, the variational ODE-IVPs become variational DDE-IVPs of the form (7.16) on the two subintervals. However, it is noted that the right-hand-sides of the two variational DDE-IVPs on the intervals $\mathcal{T}_{2,A}$ and $\mathcal{T}_{2,B}$ link discontinuously at $\tilde{s}(c)$, because the expressions $\dot{y}(t-\tau_1(c);c)$ and $\mathbf{W}(t-\tau_1(c);c)$ are discontinuous at that time point.

It remains to discuss what happens if $\tilde{s}(c) = s_2(c)$, i.e. if it holds for the first two delays that $2\tau_1(c) = \tau_2(c)$ so that the propagation of the discontinuity at $s_1(c)$ with delay $\tau_1(c)$ coincides with the propagation of the discontinuity $t^{ini}(c)$ with delay $\tau_2(c)$. Formally, this can be done by investigating the derivatives $\partial y^-(s_1(c);c)/\partial c$ and $\partial y^+(s_1(c);c)/\partial c$ for both limits $\tilde{s}(c) \to s_2^-(c)$ and $\tilde{s}(c) \to s_2^+(c)$ (i.e. let $\tilde{s}(c)$ approach $s_2(c)$ once from the left and once from the right). However, in the end it turns out that the crucial argument is, again, that the Wronskian matrix is continuous in $\tilde{s}(c)$, which is the reason why both limits lead to identical derivatives. It follows that assumption (A) can indeed be dropped, which completes the proof. ■

The fact that the subdivision of $\mathcal{T}(c)$ must include all time points of discontinuity of order 0 or 1 in y and all time points of discontinuity of order 0 in $\mathbf{W}$ is only a technical issue in the proof. However, it is important to keep this is mind for the extensions of Theorem 7.5 in Section 7.3. Due to the particular role that discontinuities of order 0 or 1 in y or of order 0 in $\mathbf{W}$ play in the following, the following definition is introduced.

Definition 7.6 (Critical Discontinuities)

Discontinuities that are of order 0 *or* 1 *in* y *or of order* 0 *in* $\mathbf{W}$ *are called* critical discontinuities.

It is appropriate to discuss the results of Theorem 7.5 on the differentiability of DDE-IVP solutions in detail. An interesting observation is that it makes very few additional assumpions compared to the related existence and uniqueness result (see Theorem 4.6, for $n_\sigma = 0$). Besides a higher smoothness of the model functions, which is a very natural condition for differentiability of

the DDE-IVP solution with respect to parameters, the only difference is the distinctness assumption (D) for the delays.

This assumption is crucial because a change in the order of the child discontinuities of $t^{ini}(c)$ generally leads to different jumps in the Wronskian (equation (7.14)). An example, where such a situation occurs is when two delays $\tau_1(c)$ and $\tau_2(c)$ coincide for specific parameter values $\tilde{c}$, $\tau_1(\tilde{c}) = \tau_2(\tilde{c})$, but their derivatives differ, $d\tau_1(c)/dc|_{c=\tilde{c}} \neq d\tau_2(c)/dc|_{c=\tilde{c}}$. While such a coincidence is entirely uncritical for the existence and uniqueness of a DDE-IVP solution, differentiability of the DDE-IVP solution is lost for all $t \geqslant t^{ini}(\tilde{c}) + \tau_1(\tilde{c})$. This is demonstrated by the following example.

Example 7.7

Consider, on the interval $\mathcal{T} = [0, 3]$, the DDE-IVP

$$\dot{\mathfrak{y}}(t; c_1) = \mathfrak{y}(t - c_1) \cdot \mathfrak{y}(t - 2) \tag{7.38a}$$

$$\mathfrak{y}(0) = 2 \tag{7.38b}$$

$$\mathfrak{y}(t; c_1) = 1 \quad \text{for} \quad t < 0. \tag{7.38c}$$

The DDE-IVP has only one parameter c_1, which is a delay, and it is assumed that $c_1 \in \mathcal{D}^c = [1.6, 2.4]$. A second delay is present which is not parameter-dependent. The state y is discontinuous at the initial time $t^{ini} = 0$, because $y^{ini} = 2 \neq 1 = \phi(t^{ini})$.

The choice of the interval $\mathcal{T}$ and of the set $\mathcal{D}^c$ is such that both deviating arguments reach, for some time $t \in \mathcal{T}$, the initial time $t^{ini} = 0$, which causes first order discontinuities in y at the time points $t = c_1$ and $t = 2$. Since it holds for the final time that $t^{fin} = 3 < 2 \cdot \min(c_1, 2)$, these discontinuities of order 1 in y do not have any child discontinuities within $\mathcal{T}$.

It is a simple task to determine the forward solution of the problem for all parameter values $c_1 \in [1.6, 2.4]$. This yields

$$y(t; c_1) = \begin{cases} t + 2 & \text{for} \quad t \in [0, c_1) \\ \frac{1}{2}t^2 + (2 - c_1)t + \frac{1}{2}c_1^2 - c_1 + 2 & \text{for} \quad t \in [c_1, 2) \\ \frac{1}{3}t^3 + \frac{2-c_1}{2}t^2 + \frac{1}{2}c_1^2 - c_1 + \frac{4}{3} & \text{for} \quad t \in [2, 3), \end{cases} \quad \text{for} \quad c_1 < 2 \tag{7.39}$$

and

$$y(t; c_1) = \begin{cases} t + 2 & \text{for} \quad t \in [0, 2) \\ \frac{1}{3}t^3 - \frac{4}{3} & \text{for} \quad t \in [2, 3), \end{cases} \quad \text{for} \quad c_1 = 2 \tag{7.40}$$

and

$$y(t; c_1) = \begin{cases} t + 2 & \text{for} \quad t \in [0, 2) \\ \frac{1}{2}t^2 - 2 & \text{for} \quad t \in [2, c_1) \\ \frac{1}{3}t^3 + \frac{2-c_1}{2}t^2 + \frac{1}{6}c_1^3 - \frac{1}{2}c_1^2 + 2 & \text{for} \quad t \in [2, 3) \end{cases} \quad \text{for} \quad c_1 > 2. \tag{7.41}$$

If the distinctness assumption (D) holds, i.e. if $c_1 \neq 2$, then it follows from Theorem 7.5 that the Wronskian – i.e. the partial derivative of the solution $y(t; c_1)$ with respect to the parameter c_1 – can be obtained from the solution of the variational DDE-IVP

$$\dot{\mathfrak{w}}(t; c_1) = y(t - 2)\,(\mathfrak{w}(t - c_1; c_1) - \dot{y}(t - c_1; c_1)) + y(t - c_1)\mathfrak{w}(t - 2; c_1) \tag{7.42a}$$

$$\mathfrak{w}(0) = 0 \tag{7.42b}$$

$$\mathfrak{w}(t; c_1) = 0 \quad \text{for} \quad t < 0. \tag{7.42c}$$

Further, a jump in $\mathbf{W}$ at the propagated discontinuity point $t = c_1$ needs to be taken into account, cf. assertion 3.[4] Alternatively, it is also possible to take directly the derivative of the equations (7.39) and (7.41) with respect to c_1 as long as $c_1 \neq 2$.

[4] The jump in $\mathbf{W}$ at the propagated discontinuity point $t = 2$ vanishes, because this propagated discontinuity point does not depend on the parameter.

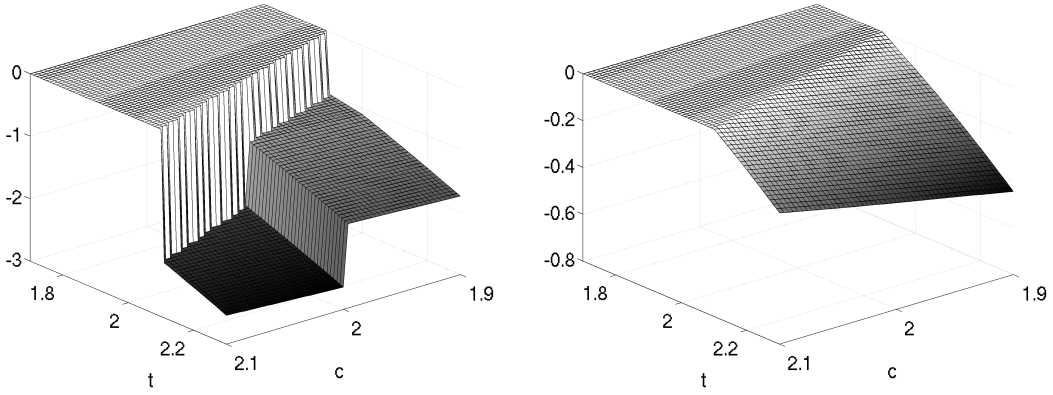
Both approaches lead to

$$\mathbf{W}(t;c_1) = \begin{cases} 0 & \text{for} \quad t \in [0, c_1) \\ -t + c_1 - 1 & \text{for} \quad t \in [c_1, 2) \\ -\frac{1}{2}t^2 + c_1 - 1 & \text{for} \quad t \in [2, 3) \end{cases} \quad \text{for} \quad c_1 < 2. \tag{7.43}$$

and

$$\mathbf{W}(t;c_1) = \begin{cases} 0 & \text{for} \quad t \in [0, c_1) \\ -\frac{1}{2}t^2 + \frac{1}{2}c_1^2 - c_1 & \text{for} \quad t \in [c_1, 3) \end{cases} \quad \text{for} \quad c_1 > 2. \tag{7.44}$$

The Wronskian $\mathbf{W}(t;c_1)$ is displayed in Figure 7.1a for the interesting domain $(t, c_1) \approx (2, 2)$. Clearly visible is, for any fixed parameter value c_1, the discontinuity of order 0 in $\mathbf{W}$ at the time point $t = c_1$, which is a child of the discontinuity at $t^{ini} = 0$ due to the propagation with the delay c_1.



(a) (b)

Figure 7.1.: (a) The Wronskian $\mathbf{W}(t;c_1)$ of the solution of the DDE-IVP (7.38) with respect to the parameter c_1. (b) The Wronskian $\mathbf{W}(t;c_1)$ of the solution of a modified version of DDE-IVP (7.38), where the initial value is set to $y(0) = 1$ such that the state is continuous at the initial time.

For $c_1 < 2$ and $c_1 > 2$, this discontinuity at $t = c_1$ is the only time point of discontinuity in the function $\mathbf{W}(t;c_1)$. However, at $c_1 = 2$, the two expressions for the Wronskian – one for $c_1 < 2$ and the other for $c_1 > 2$ – do not "match" for any time $t \geqslant 2$. The reason is that the temporal order of the child discontinuities $t^{ini} + c_1$ and $t^{ini} + 2$ changes at $c_1 = 2$, which leads to two different jumps in $\mathbf{W}$ in the expression (7.14) at the time point c_1. Hence, for $c_1 = 2$, the DDE-IVP solution is not differentiable with respect to c_1 for $t \geqslant 2$.

At this point, it is also easy to see why the distinctness assumption (D) does not only require pairwise distinctness between the delays, but also that the delays do not become 0 or equal to $t^{fin}(c) - t^{ini}(c)$. In such a situation, additional discontinuities of order 0 in $\mathbf{W}$ would "enter" or "leave" the interval $\mathcal{T}(c)$ either at $t^{ini}(c)$ or at $t^{fin}(c)$. As a consequence, at all times to the right of this additional discontinuity point, the state would not any longer depend continuously differentiable on the parameters.

The situation is different if Example 7.7 is altered in such a way that the state y is continuous at the initial time $t^{ini} = 0$, e.g. by setting $y(0) = 1$. In this case, the solution $y(t;c_1)$ becomes

$$y(t;c_1) = \begin{cases} t + 1 & \text{for} \quad t \in [0, c_1) \\ \frac{1}{2}t^2 + (1 - c_1)t + \frac{1}{2}c_1^2 + 1 & \text{for} \quad t \in [c_1, 2) \\ \frac{1}{3}t^3 - \frac{c_1}{2}t^2 + c_1 t - t + \frac{13}{3} - 2c_1 + \frac{1}{2}c_1^2 & \text{for} \quad t \in [2, 3), \end{cases} \quad \text{for} \quad c_1 < 2 \tag{7.45}$$

and

$$y(t;c_1) = \begin{cases} t+1 & \text{for} \quad t \in [0,2) \\ \frac{1}{3}t^3 - t^2 + t + \frac{7}{3} & \text{for} \quad t \in [2,3), \end{cases} \quad \text{for} \quad c_1 = 2 \tag{7.46}$$

and

$$y(t;c_1) = \begin{cases} t+1 & \text{for} \quad t \in [0,2) \\ \frac{1}{2}t^2 - t + 3 & \text{for} \quad t \in [2,c_1) \\ \frac{1}{3}t^3 - \frac{c_1}{2}t^2 + c_1 t - t + \frac{1}{6}c_1^3 - \frac{1}{2}c_1^2 + 3 & \text{for} \quad t \in [2,3) \end{cases} \quad \text{for} \quad c_1 > 2. \tag{7.47}$$

Following Theorem 7.5, the partial derivative of the IVP solution with respect to parameters can, for $c_1 \neq 2$, be determined by the solution of the variational DDE-IVP (7.42). Further, no discontinuities of order 0 in the Wronskian matrix $\mathbf{W}$ need to be taken into account, because y is continuous at the initial time t^{ini}.

As an alternative to the application of Theorem 7.5, it is also possible to take the derivative of the equations (7.45) and (7.47). Both approaches for the computation of the Wronskian yield the expressions

$$\mathbf{W}(t;c_1) = \begin{cases} 0 & \text{for} \quad t \in [0,c_1) \\ -t + c_1 & \text{for} \quad t \in [c_1,2) \\ -\frac{1}{2}t^2 + t - 2 + c_1 & \text{for} \quad t \in [2,3) \end{cases} \quad \text{for} \quad c_1 < 2. \tag{7.48}$$

and

$$\mathbf{W}(t;c_1) = \begin{cases} 0 & \text{for} \quad t \in [0,c_1) \\ -\frac{1}{2}t^2 + t + \frac{1}{2}c_1^2 - c_1 & \text{for} \quad t \in [c_1,3) \end{cases} \quad \text{for} \quad c_1 > 2. \tag{7.49}$$

In contrast to the original version of Example 7.7, the Wronskian $\mathbf{W}(t;c_1)$ approaches the same function $\mathbf{W}(t;2) = -\frac{1}{2}t^2 + t$ for both $\lim_{\epsilon \to 0^+} \mathbf{W}(t;2+\epsilon)$ and $\lim_{\epsilon \to 0^-} \mathbf{W}(t;2+\epsilon)$. Hence, $y(t;c)$ is continuously differentiable for all $c \in \mathcal{D}^c$, see also Figure 7.1b, even though the distinctness assumption (D) is violated. The reason is that the state is continuous in the initial time, from which it follows that the discontinuity order of $\mathbf{W}$ at the child discontinuities is 1 (instead of 0). Therefore, it does not matter in which temporal order the child discontinuities occur.

This finding can be generalized: the distinctness assumption (D) is dispensable in Theorem 7.5 whenever, for all c in a neighborhood $\mathcal{V}^c$ of the nominal parameters $\tilde{c}$, the state y is continuous at the initial time.

7.3. More General Differentiability Results

7.3.1. Piecewise Continuously-Differentiable Initial Functions

Theorem 7.5 gives a result for the differentiability of solutions of DDE-IVPs with constant delays. The initial function was thereby assumed to be continuously differentiable. Hence, there is at most one discontinuity of order 0 in y, namely at the initial time $t^{ini}(c)$. The question is how this differentiability result can be generalized to piecewise continuously differentiable initial functions, i.e. for $\phi(\cdot,c) \in \mathscr{PD}(\mathcal{T}^f(c), \mathbb{R}^{n_y})$.

In order to find the answer to this question, it is recalled that differentiability of the initial time $t^{ini}(c)$ was assumed in Theorem 7.5 in order to guarantee that the time point of the child discontinuities, given by $t^{ini}(c) + \tau_i(c)$, are differentiable functions of the parameters. Accordingly, in order to obtain differentiability of solutions of DDE-IVPs with piecewise continuously differentiable initial functions, it is necessary to assume that the time points of all critical discontinuities in the initial function are differentiable functions of c. More precisely, if there are n_s^ϕ critical discontinuities in the initial functions, then there must be functions $s_i(c)$, $-n_s^\phi \leqslant i \leqslant -1$, which describe the locations of the discontinuities and which are differentiable with respect to c for a neighborhood $\mathcal{V}^c$ of $\tilde{c}$.

The initial function $\phi(t,c)$ should then be given by

$$\phi(t,c) = \begin{cases} \phi_{-n_s^\phi}(t,c) & \text{for} \quad t < s_{-n_s^\phi}(c) \\ \phi_{-n_s^\phi+1}(t,c) & \text{for} \quad s_{-n_s^\phi}(c) \leqslant t < s_{-n_s^\phi+1}(c) \\ \vdots & \\ \phi_0(t,c) & \text{for} \quad s_{-1}(c) \leqslant t \end{cases}, \tag{7.50}$$

where each function $\phi_i(t,c)$ is continuously differentiable for $(t,c) \in \mathcal{T}^{f,\Delta c}(c) \times \mathcal{V}^c$. They are therefore also called the *smooth branches (of the initial function ϕ)*.

Eventually, the distinctness assumption (D) needs to be modified is such a way that distinctness holds for all children of discontinuities of order 0 in y. Moreover, children of discontinuities of order 0 in y must not enter or leave the interval $\mathcal{T}(c)$. Formally: the discontinuity points $s_j(c) + \tau_i(c)$, $-n_s^\phi \leqslant j \leqslant -1$, $1 \leqslant i \leqslant n_\tau$ and $t^{ini}(c) + \tau_i(c)$, $1 \leqslant i \leqslant n_\tau$, must be pairwise distinct and not equal to $t^{ini}(c)$ or $t^{fin}(c) - t^{ini}(c)$ for all $c \in \mathcal{V}^c$.

With these three modifications – differentiability of the initial discontinuity points, differentiability of the *smooth branches* of the initial function $\phi_i(t,c)$, and a modified distinctness assumption – it is possible to reduce the DDE-IVP to a sequence of ODE-IVPs for which the assumptions of Theorem 7.2 hold. Then, differentiability of the DDE-IVP solution follows.

7.3.2. Non-Vanishing Time-Dependent Delays

As a next step, consider DDE-IVPs with non-vanishing time-dependent delays. In order to obtain an existence, uniqueness, and differentiability result as in Theorem 7.5, a set of conditions needs to be found so that the DDE-IVP can be reduced to a sequence of ODE-IVPs for which Theorem 7.2 can be applied.

In this context, it is immediately clear that requirement (R2) needs to be fulfilled, cf. Subsection 4.3.1 on the existence and uniqueness of DDE-IVPs with time-dependent delays. It is also clear that the delay functions $\tau_i(t,c)$, $1 \leqslant i \leqslant n_\tau$, have to be differentiable for all $(t,c) \in \mathcal{T}^{f,\Delta t}(\tilde{c}) \times \mathcal{V}^c$, with $\Delta t > 0$ and with $\mathcal{V}^c$ being a neighborhood of the nominal parameters $\tilde{c}$. Further, the propagated discontinuities have to satisfy a suitable distinctness assumption. Discontinuities whose parent discontinuity is of order 0 in y should not occur at the same time point, and not at $t^{ini}(c)$ or $t^{fin}(c)$.

In addition, it needs to be guaranteed that the locations of children of critical discontinuities are differentiable functions of the parameters. Differentiability of the critical discontinuities themselves and differentiability of the delay functions are not sufficient in this context, as is illustrated by the following example.

Example 7.8

Consider the following DDE-IVP on the interval $\mathcal{T} = [0,2]$, with nominal parameter value $\tilde{c}_1 = 0$, and $c_1 \in \mathcal{V}^c = (-0.1, 0.1)$:

$$\dot{\mathfrak{y}}(t;c_1) = \mathfrak{y}(t - \tau_1(t,c_1)) \tag{7.51a}$$

$$\mathfrak{y}(0) = 0 \tag{7.51b}$$

$$\mathfrak{y}(t) = 1 \quad \text{for} \quad t \in (-\infty, 0), \tag{7.51c}$$

where the delay function is defined by

$$\tau_1(t,c_1) = \begin{cases} t - (t-0.5)^3 + c_1 & \text{for} \quad t < 0.5 \\ t + c_1 & \text{for} \quad 0.5 \leqslant t < 1.5 \\ t - (t-1.5)^3 + c_1 & \text{for} \quad t \geqslant 1.5. \end{cases} \tag{7.52}$$

For the nominal parameter $\tilde{c}_1 = 0$, the deviating argument $\alpha_1(t,c_1) = t - \tau_1(t,c_1)$ assumes values smaller than 0 for $t < 0.5$, identically 0 for $t \in [0.5, 1.5]$, and it assumes values greater than 0 for $t > 1.5$. Since propagated discontinuities occur, by definition, at those time points where the simplified sign function

$$\zeta_{1,t^{ini}}^{\alpha,+}(t) = \operatorname{sign}^+(\alpha_1(t,c_1) - t^{ini}) \tag{7.53}$$

changes its value either from $+1$ to -1 or vice versa (recall Definition 2.11), there is only one propagated discontinuity for $\tilde{c}_1 = 0$ at $s = 0.5$. However, for any (arbitrarily small) positive value of c_1, the deviating argument assumes values to the left of t^{ini} for all $t \leqslant 1.5$, and the propagated discontinuity now occurs at some time $t > 1.5$. Hence, the time point of the propagated discontinuity is not a continuous, let alone continuously differentiable, function of the parameter c_1, even though the time point of the parant discontinuity (t^{ini}) is differentiable (with derivative 0) and the delay is non-vanishing and a twice-continuously differentiable function that fulfills requirement (R2) for all $c_1 \in \mathcal{V}^c$.

In order to formulate a sufficient condition for the differentiability of the discontinuity points in DDE-IVPs with time-dependent delays, the implicit function theorem is recalled.

Theorem 7.9 (Implicit Function Theorem)
Let $\mathcal{V}^{x_1} \subset \mathbb{R}^{n_{x_1}}$ and $\mathcal{V}^{x_2} \subset \mathbb{R}^{n_{x_2}}$ be open subsets and let $F : \mathcal{V}^{x_1} \times \mathcal{V}^{x_2} \to \mathbb{R}^{n_{n_x}}$, $F : (x_1, x_2) \to F(x_1, x_2)$ be a continuously differentiable function of its arguments. Let further $(a, b) \in \mathcal{V}^{x_1} \times \mathcal{V}^{x_2}$ be a point with $F(a, b) = 0$, and let the $n_{x_2} \times n_{x_2}$ matrix

$$\left.\frac{\partial F}{\partial x_2}(x_1, x_2)\right|_{(x_1,x_2)=(a,b)}$$

be regular. Then there exists an open neighborhood $\mathcal{U}^{x_1} \subset \mathcal{V}^{x_1}$ of a, an open neighborhood $\mathcal{U}^{x_2} \subset \mathcal{V}^{x_2}$ of b, and a continuously differentiable function $G : \mathcal{U}^{x_1} \to \mathcal{U}^{x_2} \subset \mathbb{R}^{n_{x_2}}$ with $G(a) = b$ such that

$$F(x_1, G(x_1)) = 0 \quad \textit{for all} \quad x_1 \in \mathcal{U}^{x_1}. \tag{7.54}$$

If $(x_1, x_2) \in \mathcal{U}^{x_1} \times \mathcal{U}^{x_2}$ is such that $F(x_1, x_2) = 0$, then $x_2 = G(x_1)$.
It holds for the derivative of the function G that

$$\frac{\partial G(x_1)}{\partial x_1} = -\left[\left(\frac{\partial F}{\partial x_2'}(x_1', x_2')\right)^{-1} \frac{\partial F}{\partial x_1'}(x_1', x_2')\right]_{(x_1',x_2')=(x_1,G(x_1))}. \tag{7.55}$$

Proof
See Forster [110], page 89f. ■

In the context of DDE-IVPs with non-vanishing, time-dependent delay functions, the implicit function theorem is applied in the setting $x_1 \to c$, $x_2 \to t$, and

$$F(c, t) := \sigma^{\alpha}_{i,s_j}(t, c) = \alpha_i(t, c) - s_j(c). \tag{7.56}$$

Therein, α_i is the i-th deviating argument, $s_j(c)$ is the time point of a parameter-dependent critical discontinuity in the past, i.e. $s_j(c) < t$, and $\sigma^{\alpha}_{i,s_j(c)}$ is the associated propagation switching function.

Let now, according to Theorem 7.9, $s_k(\tilde{c})$ be the time point of a child discontinuity whose parent discontinuity is located at $s_j(\tilde{c})$, and which is propagated by the delay τ_i. If $\sigma^{\alpha}_{i,s_j(\tilde{c})}$ is differentiable with non-zero time derivative

$$\left.\frac{d\sigma^{\alpha}_{i,s_j(\tilde{c})}(t, c)}{dt}\right|_{(t,c)=(s_k(\tilde{c}),\tilde{c})} \neq 0 \tag{7.57}$$

then there exists a neighborhood of $(s_k(\tilde{c}), \tilde{c})$ where the location of the propagated discontinuity is a differentiable function $s_k(c)$ of the parameters, and

$$\frac{ds_k(c)}{dc} = -\left[\left(\dot{\alpha}_i(t', c')\right)^{-1}\left(\frac{\partial \alpha_i(t', c')}{\partial c'} - \frac{ds_j(c')}{dc'}\right)\right]_{(t',c')=(s_k(c),c)}. \tag{7.58}$$

In a shorter notation, this relation is written as

$$\frac{ds_k(c)}{dc} = -\left(\dot{\alpha}_i(s_k(c), c)\right)^{-1}\left(\frac{\partial \alpha_i(s_k(c), c)}{\partial c} - \frac{ds_j(c)}{dc}\right), \tag{7.59}$$

and this shorter notation is used frequently in the following.

It is appropriate to summarize the assumptions that need to be made for a differentiability result for DDE-IVPs with time-dependent delays:

(a) The time-dependent delay functions are differentiable with respect to all arguments.

(b) The time-dependent delays fulfill requirement (R2).

(c) Children of discontinuities of order 0 in y occur at pairwise distinct time points, and they do not occur at $t^{ini}(c)$ or $t^{fin}(c)$.

(d) All deviating arguments cross the time points of past critical discontinuities with non-zero time derivative. This last condition ensures differentiability of the time points of the child discontinuities. This condition reoccurs in the following sections and subsections as a crucial condition for differentiability of all IVP solutions in which implicitly determined discontinuities play a role.

Note that these are only those assumptions that need to be made in addition to the earlier made assumptions for DDE-IVPs with constant delays and with piecewise continuously differentiable initial functions.

If the sufficient conditions for differentiability are satisfied, then the Wronskian matrix is piecewise given as solution of a variational DDE-IVP. Further, the jump in the Wronskian matrix $\mathbf{W}(t;c)$ at the discontinuity point $s_k(c)$ is given by equation (7.14), with $ds_k(c)/dc$ given by equation (7.59).

7.3.3. HDDEs with (Simple) Time-Dependent Switching Functions

With the conclusions from the last subsection concerning DDEs with time-dependent delays in mind, it is easy to extend the differentiability result of Theorem 7.5 also to HDDEs with simple time-dependent or general time-dependent switching functions. The set of additional conditions is shortly summarized as follows:

(a) The switching functions are differentiable with respect to all arguments.

(b) The number of root discontinuities is finite (i.e. requirement (R1) is fulfilled).

(c) Root discontinuities and children of discontinuities of order 0 in y occur at piecewise distinct time points, and they do not occur at $t^{ini}(c)$ or at $t^{fin}(c)$. In particular, the time points of each of the root discontinuities is a zero of only one switching function.

(d) The switching functions cross their zeros with non-zero time derivative. Note that this last condition is automatically fulfilled for simple time-dependent switching functions, because in this case the time derivative is identically 1.

If the conditions are fulfilled, then the derivative of a time point $s_k(c)$ of a root discontinuity, which is the zero of the time-dependent switching function $\sigma_i(t,c)$, is given by

$$\frac{ds_k(c)}{dc} = -\left(\dot{\sigma}_i(s_k(c),c)\right)^{-1} \frac{\partial \sigma_i(s_k(c),c)}{\partial c}. \tag{7.60}$$

This expression occurs in the equation that gives the size of the jump in the Wronskian:

$$\mathbf{W}^+(s_k(c);c) - \mathbf{W}^-(s_k(c);c) = \left(f_k^-(c) - f_k^+(c)\right) \frac{ds_k(c)}{dc}. \tag{7.61}$$

Herein, the quantity $f_k^-(c) - f_k^+(c)$ represents the difference between the left-sided and the right-sided evaluation of f at $s_k(c)$, which differ in exactly one argument, namely the sign of the switching function σ_i.

7.3.4. Impulses

For the case of non-zero impulses, e.g. in an IHDDE-IVP with time-dependent switching and delay functions, it must at first be ensured that the additional conditions for existence and uniqueness hold (see Subsection 4.3.3). These conditions are that the delay functions must obey requirement (R2) for all discontinuities of order 0 introduced in the root discontinuities, and that the set

$\mathcal{D}^y$ is sufficiently large and the impulses are sufficiently small so that the state remains in the set $\mathcal{D}^y$.

Clearly, for differentiability, it is also necessary that the impulse functions are continuously differentiable with respect to all arguments. It further follows from the previous subsections that switching functions and propagation switching functions of critical discontinuities should have a non-zero time derivative in their zeros. It has also been discussed that root discontinuities and propagations of discontinuities of order 0 in y should occur at pairwise distinct time points, and they should not occur at $t^{ini}(c)$ or at $t^{fin}(c)$. However, it turns out in the following that the last of these conditions, i.e. the distinctness, has to be formulated more restrictively for the treatment of impulses.

In order to derive the jump in the Wronskian $\mathbf{W}(t;c)$ at the time point $s_k(c)$ where a non-zero impulse is applied, it is first remarked that the derivative of the time point of the root discontinuity is still given by expression (7.60). Further, the total derivative of the left-sided limit of the state at the time point of the root discontinuity, i.e. $y^-(s_k(c);c)$, is given by

$$\begin{aligned}\frac{dy^-(s_k(c);c)}{dc} &= \frac{dy^-(s_k(c);c)}{dt}\frac{ds_k(c)}{dc} + \frac{\partial y^-(s_k(c);c)}{\partial c}\\ &= f_k^-(c)\frac{ds_k(c)}{dc} + \mathbf{W}^-(s_k(c);c).\end{aligned} \tag{7.62}$$

Therein, $f_k^-(c)$ denotes, as in Section 7.2, the left-sided limit of the right-hand-side function f at the discontinuity point $s_k(c)$. For the case of IHDDE-IVPs with time-dependent delays, it reads

$$f_k^-(c) := f(s_k(c), y^-(s_k(c);c), c, \{y(s_k(c) - \tau_i(s_k(c),c);c)\}_{i=1}^{n_\tau}, \zeta^k), \tag{7.63}$$

where ζ^k denotes the signs of the switching functions to the left of $s_k(c)$. Note that it has been taken into account in the notation of the past states that they are continuous at the time points $s_k(c) - \tau_i(s_k(c),c)$, $1 \leqslant i \leqslant n_\tau$, because root discontinuities and children of discontinuities of order 0 in y must not coincide (distinctness). Hence, it is not necessary to distinguish between the left-sided limit y^- and the right-sided limit y^+ (or to use the generic expression $y^\bullet$) for the past states in equation (7.63).

The next step is to compute the total derivative of the right-sided limit of the state at $s_k(c)$. This state is determined by the impulse equation. Recall for this purpose that, according to condition c) in the previous subsection on HDDEs with time-dependent switching functions, each time point of a root discontinuity is the zero of only one switching function. Let, in the following, $\sigma_{I(k)}$ be the switching function that is zero at $s_k(c)$, i.e. the function I maps the index of the time point of a root discontinuity to the index of the corresponding switching function.

Recall further that the impulses in IHDDE-IVPs were defined in Chapter 1 by the impulse functions ω, and that one of the arguments of ω was $\zeta(t)$, i.e. the vector of signs of the switching functions. This notation had been used because in general any combination of switching functions may become zero at the same time point. However, under the assumption that zeros of switching functions do not coincide, the use of a different notation is suitable in the following. Therefore, it is from now on understood that $\omega_{I(k)}$ is the impulse function that has to be applied in the zero of the switching function $\sigma_{I(k)}$.

It should be remarked that this notational simplification additionally implies that the impulse does not depend on the (non-zero) signs of the remaining switching functions. However, this limitation is irrelevant for the theoretical analysis, and also sufficient for the treatment of many practical applications.

By characterizing impulses with the index of the sole switching function that is zero, the right-sided limit of the state at $s_k(c)$ is given by

$$y^+(s_k(c);c) = y^-(s_k(c);c) + \omega_{I(k)}(s_k(c), y^-(s_k(c);c), c, \{y(s_k(c) - \tau_i(s_k(c),c);c)\}_{i=1}^{n_\tau}). \tag{7.64}$$

By use of elementary differentiation rules, the total derivative $dy^+(s_k(c);c)/dc$ turns out to be

$$\begin{aligned}\frac{dy^+(s_k(c);c)}{dc} =& \frac{dy^-(s_k(c);c)}{dc} + \left[\frac{\partial \omega_{I(k)}}{\partial t}\frac{ds_k(c)}{dc} + \frac{\partial \omega_{I(k)}}{\partial y}\frac{dy^-(s_k(c);c)}{dc}\right.\\ &\left. + \frac{\partial \omega_{I(k)}}{\partial c} + \sum_{m=1}^{n_\tau}\frac{\partial \omega_{I(k)}}{\partial v_m}\cdot\frac{dy(s_k(c)-\tau_m(s_k(c),c);c)}{dc}\right],\end{aligned} \tag{7.65}$$

where the arguments of the partial derivatives of the impulse functions have been surpressed for notational simplicity; they have to be evaluated at $(s_k(c), y^-(s_k(c);c), c, y(s_k(c)-\tau_i(s_k(c),c);c))$.

According to relation (7.8), the total derivative of the right-sided limit of the state can further be expressed as

$$\frac{dy^+(s_k(c);c)}{dc} = \mathbf{W}^+(s_k(c);c) + f_k^+(c)\frac{ds_k(c)}{dc}. \tag{7.66}$$

Herein,

$$f_k^+(c) := f(s_k(c), y^+(s_k(c);c), c, \{y(s_k(c)-\tau_i(s_k(c),c);c)\}_{i=1}^{n_\tau}, \zeta^{k+1}), \tag{7.67}$$

and ζ^{k+1} denotes the sign of the switching functions to the right of $s_k(c)$.

Finally, the total derivatives of the past states in equation (7.65) are given by

$$\begin{aligned}\frac{dy(s_k(c)-\tau_m(s_k(c),c);c)}{dc} =& \dot{y}(s_k(c)-\tau_m(s_k(c),c);c)\\ &\cdot\left(\frac{ds_k(c)}{dc} - \dot{\tau}_m(s_k(c),c)\frac{ds_k(c)}{dc} - \frac{\partial \tau_m(s_k(c),c)}{\partial c}\right)\\ &+ \mathbf{W}(s_k(c)-\tau_m(s_k(c),c);c).\end{aligned} \tag{7.68}$$

Inserting all this into equation (7.65) yields

$$\begin{aligned}\mathbf{W}^+(s_k(c);c) - \mathbf{W}^-(s_k(c);c) =& \left[\left(\mathbf{1}_{n_y} + \frac{\partial \omega_{I(k)}}{\partial y}\right)f_k^-(c) + \frac{\partial \omega_{I(k)}}{\partial t} - f_k^+(c)\right.\\ &\left. + \sum_{m=1}^{n_\tau}\dot{y}(s_k(c)-\tau_m(s_k(c),c);c)(1-\dot{\tau}_m(s_k(c);c))\right]\frac{ds_k(c)}{dc}\\ &+ \frac{\partial \omega_{I(k)}}{\partial y}\mathbf{W}^-(s_k(c);c) + \frac{\partial \omega_{I(k)}}{\partial c}\\ &+ \sum_{m=1}^{n_\tau}\frac{\partial \omega_{I(k)}}{\partial v_m}\left[-\dot{y}(s_k(c)-\tau_m(s_k(c),c);c)\frac{\partial \tau_m(s_k(c),c)}{\partial c}\right.\\ &\left. + \mathbf{W}(s_k(c)-\tau_m(s_k(c),c);c)\right].\end{aligned} \tag{7.69}$$

This equation describes the jump in the Wronskian $\mathbf{W}$ for the case of non-zero impulses, with $\mathbf{1}_{n_y}$ being the $n_y \times n_y$-dimensional identity matrix.

Note that the right hand side of equation (7.69) depends both on the time derivative of the state $\dot{y}$ and on the Wronskian $\mathbf{W}$ at the past time point $s_k(c) - \tau_m(s_k(c),c)$. Because of this, the distinctness assumption has to be formulated in such a way that the time points of the root discontinuities do not coincide with the time points of children on criticial discontinuities – instead of just avoiding coincidences with the time points of children of discontinuities of order 0 in y, as it has been the case in Subsection 7.3.3 for HDDE-IVPs. However, the more restrictive version of distinctness is needed only if the impulse function actually depends on past states.

7.3.5. Intermediate Summary

So far, differentiability results for IHDDE-IVPs with time-dependent switching and delay functions were considered in this chapter. In comparison to the existence and uniqueness results in Chapter 4, three major additional conditions were worked out.

The first additional condition is that a higher degree of smoothness needs to be assumed for the

model functions. In particular, a smoothness assumption needs to be made for the time points of discontinuity in the initial function, which were not "visible" as model functions in the Chapter 4, because for existence and uniqueness it was sufficient to consider some fixed nominal parameter values $\tilde{c}$ and assume that $\phi(\cdot, \tilde{c}) \in \mathscr{PD}(\mathcal{T}^f, \mathcal{D}^y)$.

The second major modification is the assumption of distinctness. In the case of IHDDE-IVPs with multiple time-dependent delay and switching functions, the distinctness needs to ensure

- that the time points of the root discontinuities are zeros of only one switching function, and that they do not coincide with the time points of children of critical discontinuities,
- that children of discontinuities of order 0 in y do not occur at the same time point,
- and that root discontinuities and children of discontinuities of order 0 in y do not occur at the initial time or at the final time.

As a consequence, IHDDE-IVP solutions that fulfill these conditions for differentiability have a rather simple structure regarding the locations of critical discontinuities, which is a very important observation also for the numerical treatment of these problems.

The third and last assumption for differentiability that goes beyond the assumptions for existence and uniqueness is that switching functions and propagation switching functions of critical discontinuities have non-zero time derivatives in their zeros. This is necessary for the applicability of the implicit function theorem. Thus, this is sufficient for the differentiability of the time points of the root discontinuities and of the propagated discontinuities with respect to parameters.

In analogy to Chapter 4, state dependencies in the switching or delay functions lead to a significant complication of the problem. Therefore, it is assumed in the following sections that a solution exists for some nominal parameters $\tilde{c}$. Then, sufficient conditions are given that ensure differentiability of the IVP solution with respect to c in a neighborhood of $\tilde{c}$. IHODE-IVPs with state-dependent switching functions are discussed first, followed by DDE-IVPs with state-dependent delay functions. Eventually, the general case of IHDDE-IVPs with both state-dependent switching and state-dependent delay functions is considered.

7.4. IHODEs with State-Dependent Switching Functions

As a first problem class that involves state dependencies, IHODE-IVPs as in Definition 1.10 are investigated. For this problem class, the following differentiability theorem holds (cf. Bock [39], Galan, Féehery, and Barton [111]):

Theorem 7.10 (Global Differentiability of IHODE-IVP Solutions)

Consider an IHODE-IVP as in Definition 1.10, with nominal parameters $\tilde{c}$ and a neighborhood $\mathcal{V}^c$ of $\tilde{c}$. Let $\mathcal{V}^y \subset \mathbb{R}^{n_y}$ be an open domain, and let $y : \mathcal{T}(\tilde{c}) \to \mathcal{V}^y$, $y : (t, \tilde{c}) \to y(t; \tilde{c})$ be a solution of the problem for the nominal parameters. Choose $\Delta t > 0$ for the definition of the interval $\mathcal{T}^{\Delta t}(\tilde{c})$.

Further, let $s_1(\tilde{c}) < \cdots < s_{n_s}(\tilde{c})$, $s_k(\tilde{c}) \in (t^{ini}(\tilde{c}), t^{fin}(\tilde{c}))$ for $1 \leqslant k \leqslant n_s$ be the zeros of the switching functions (i.e. the root discontinuities) that occur for the given solution $y(t; \tilde{c})$, and define $s_0(c) := t^{ini}(c)$, $s_{n_s+1}(c) := t^{fin}(c)$. Let $I(k) \in \{1, \ldots, n_\sigma\}$ denote the index of the switching function that is zero at the time point $s_k(\tilde{c})$ for $1 \leqslant k \leqslant n_s$. The signs of the switching functions on the subintervals are denoted by

$$\zeta^k := \zeta(t) \quad \textit{for} \quad t \in (s_{k-1}(\tilde{c}), s_k(\tilde{c})), \quad \textit{for} \quad 1 \leqslant k \leqslant n_s + 1 \tag{7.70}$$

and the left-sided limit and the right-sided limit of the right-hand-side function f at $s_k(\tilde{c})$ for $1 \leqslant k \leqslant n_s$ are denoted by $f_k^-(\tilde{c})$ and $f_k^+(\tilde{c})$, with

$$f_k^-(c) := f(s_k(c), y^-(s_k(c); c), c, \zeta^k) \tag{7.71a}$$

$$f_k^+(c) := f(s_k(c), y^+(s_k(c); c), c, \zeta^{k+1}). \tag{7.71b}$$

Let the following assumptions be fulfilled:

(S) Smoothness*: The initial time $t^{ini}(c)$, the final time $t^{fin}(c)$, and the initial value $y^{ini}(c)$ are continuously differentiable functions for $c \in \mathcal{V}^c$. The right-hand-side function $f(t, y, c, \zeta)$ is continuous in t and continuously differentiable with respect to y and c, and uniformly*

Lipschitz continuous with respect to y for $(t, y, c, \zeta) \in \mathcal{T}^{\Delta t}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{V}^c \times \mathcal{I}_1^\zeta$. The switching functions $\sigma_i(t, y, c)$ for $1 \leqslant i \leqslant n_\sigma$, and the associated impulse functions $\omega_i(t, y, c)$ that have to be evaluated in their zeros, are continuously differentiable with respect to their arguments for $(t, y, c) \in \mathcal{T}^{\Delta t}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{V}^c$.

(B) Boundedness*: The right-hand-side function f is bounded by*

$$\|f(t, y, c, \zeta)\|_\infty \leqslant M_f < \infty \tag{7.72}$$

for $(t, y, c, \zeta) \in \mathcal{T}^{\Delta t}(\tilde{c}) \times \mathcal{V}^y \times \mathcal{V}^c \times \mathcal{I}_1^\zeta$.

(RS) Regularity of the Switching Functions*: The switching functions are non-zero in $t^{ini}(\tilde{c})$ and $t^{fin}(\tilde{c})$, i.e. $\sigma_j(t^{ini}(\tilde{c}), y(t^{ini}(\tilde{c}); \tilde{c}), \tilde{c}) \neq 0$ and $\sigma_j(t^{fin}(\tilde{c}), y(t^{fin}(\tilde{c}); \tilde{c}), \tilde{c}) \neq 0$ for $1 \leqslant j \leqslant n_\sigma$. In addition, depending on the impulse function, the following conditions hold:*

- *if $\omega_{I(k)}(t, y, c) \equiv 0$, then*

$$\left[\frac{\partial \sigma_{I(k)}(t, y, c)}{\partial t} + \frac{\partial \sigma_{I(k)}(t, y, c)}{\partial y} f_k^-(c) \right]_{(t,y,c)=(s_k(\tilde{c}), y^-(s_k(\tilde{c}); \tilde{c}), \tilde{c})} \neq 0. \tag{7.73}$$

  and

$$\left[\frac{\partial \sigma_{I(k)}(t, y, c)}{\partial t} + \frac{\partial \sigma_{I(k)}(t, y, c)}{\partial y} f_k^+(c) \right]_{(t,y,c)=(s_k(\tilde{c}), y^+(s_k(\tilde{c}); \tilde{c}), \tilde{c})} \neq 0. \tag{7.74}$$

- *if $\omega_{I(k)}(t, y, c) \not\equiv 0$, then condition (7.73) is assumed and*

$$\sigma_{I(k)}(s_k(\tilde{c}), y^+(s_k(\tilde{c}); \tilde{c}), \tilde{c}) \neq 0. \tag{7.75}$$

Further, for all other switching functions σ_j, $j \neq I(k)$, it holds that

$$\sigma_j(s_k(\tilde{c}), y^-(s_k(\tilde{c}); \tilde{c}), \tilde{c}) \neq 0 \quad \text{and} \quad \sigma_j(s_k(\tilde{c}), y^+(s_k(\tilde{c}); \tilde{c}), \tilde{c}) \neq 0. \tag{7.76}$$

Then there exists $\delta t > 0$, an open interval $\mathcal{T}^{\delta t}(\tilde{c})$, and a neighborhood $\mathcal{U}^c$ of $\tilde{c}$ such that $\mathcal{T}(c) \subset \mathcal{T}^{\delta t}(\tilde{c})$ for $c \in \mathcal{U}^c$, and for $(t, c) \in \mathcal{T}^{\delta t}(\tilde{c}) \times \mathcal{U}^c$ the following assertions hold.

1. *There exists a unique solution $y(t; c)$.*
2. *The Wronskian $\mathbf{W}(t; c) = \partial y(t; c)/\partial c$ is continuous for $t \neq s_k(c)$, $1 \leqslant k \leqslant n_s$, and right-continuous in $t = s_k(c)$, $1 \leqslant k \leqslant n_s$.*
3. *The discontinuity time points $s_k(c)$ are continuously differentiable functions of the parameters c, and*

$$\frac{ds_k(c)}{dc} = -\frac{\frac{\partial \sigma_{I(k)}}{\partial y} \mathbf{W}^-(s_k(c); c) + \frac{\partial \sigma_{I(k)}}{\partial c}}{\frac{\partial \sigma_{I(k)}}{\partial t} + \frac{\partial \sigma_{I(k)}}{\partial y} f_k^-(c)}, \tag{7.77}$$

   where the partial derivatives of $\sigma_{I(k)}$ have to be evaluated at $(s_k(c), y^-(s_k(c); c), c)$.
4. *At the discontinuity time points $s_k(c)$, $1 \leqslant k \leqslant n_s$, the jump in the Wronskian $\mathbf{W}(t; c)$ (i.e. the difference between left-sided and right-sided) is given by*

$$\begin{aligned} \mathbf{W}^+(s_k(c); c) - \mathbf{W}^-(s_k(c); c) = &+ \frac{\partial \omega_{I(k)}}{\partial y} \mathbf{W}^-(s_k(c); c) + \frac{\partial \omega_{I(k)}}{\partial c} \\ &+ \left(f_k^-(c) + \frac{\partial \omega_{I(k)}}{\partial t} + \frac{\partial \omega_{I(k)}}{\partial y} f_k^-(c) - f_k^+(c) \right) \frac{ds_k(c)}{dc}, \end{aligned} \tag{7.78}$$

   where $ds_k(c)/dc$ is given by equation (7.77) and the partial derivatives of $\omega_{I(k)}$ have to be evaluated at $(s_k(c), y^-(s_k(c); c), c)$.

5. *On the right-open interval* $[s_k(c), s_{k+1}(c))$, $0 \leqslant k \leqslant n_s$, *the Wronskian* $\mathbf{W}(t;c)$ *is given as the solution of the variational IVP*

$$\dot{\mathfrak{W}}(t;c) = \frac{\partial f(t, y(t;c), c, \zeta^{k+1})}{\partial y}\mathfrak{W}(t;c) + \frac{\partial f(t, y(t;c), c, \zeta^{k+1})}{\partial c} \tag{7.79a}$$

$$\mathfrak{W}(s_k(c);c) = \frac{dy_k(c)}{dc} - f_k^+(c)\frac{ds_k(c)}{dc}, \tag{7.79b}$$

where $y_k(c) = y^{ini}(c)$ *for* $k = 0$ *and* $y_k(c) = y^-(s_k(c);c) + \omega_{I(k)}(s_k(c), y^-(s_k(c);c), c)$ *for* $k \neq 0$ *is the initial state for the corresponding interval. For* $k = 0$, *the derivative with respect to the parameters is given by* $dy_0(c)/dc = dy^{ini}(c)/dc$, *whereas for* $k \neq 0$ *the derivative is given by*

$$\begin{aligned}\frac{dy_k(c)}{dc} = &\left(\mathbf{1}_{n_y} + \frac{\partial \omega_{I(k)}}{\partial y}\right)\mathbf{W}^-(s_k(c);c) + \frac{\partial \omega_{I(k)}}{\partial c} \\ &+ \left(f_k^-(c) + \frac{\partial \omega_{I(k)}}{\partial t} + \frac{\partial \omega_{I(k)}}{\partial y} f_k^-(c)\right) \cdot \frac{ds_k(c)}{dc},\end{aligned} \tag{7.80}$$

which, by using equation (7.78), is equivalent to

$$\frac{dy_k(c)}{dc} = f_k^+(c)\frac{ds_i(c)}{dc} + \mathbf{W}^+(s_k(c);c). \tag{7.81}$$

In the special case $k = 0$, *the symbol* $f_0^+(c)$ *in equation (7.79b) is defined by*

$$f_0^+(c) := f(t^{ini}(c), y(t^{ini}(c);c), c, \zeta^1). \tag{7.82}$$

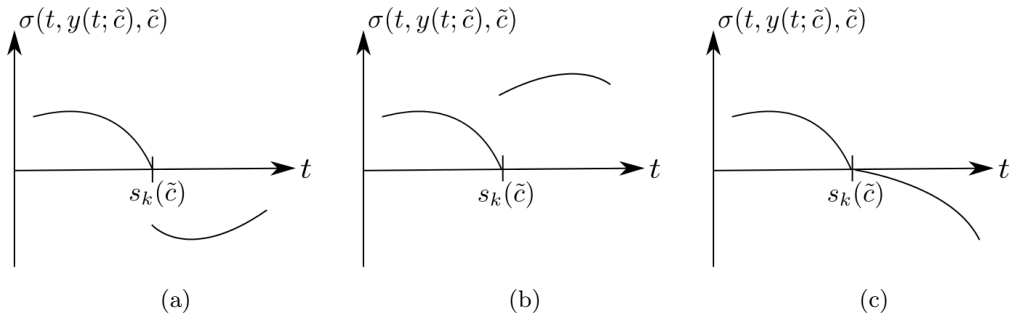
6. *The total derivative of the state at the final time with respect to the parameters is given by*

$$\frac{dy(t^{fin}(c);c)}{dc} = \mathbf{W}^-(t^{fin}(c);c) + f_{n_s+1}^-(c)\frac{dt^{fin}(c)}{dc}. \tag{7.83}$$

with

$$f_{n_s+1}^-(c) := f(t^{fin}(c), y(t^{fin}(c);c), c, \zeta^{n_s+1}). \tag{7.84}$$

Before the proof of the theorem is given, the regularity condition (RS) is discussed and illustrated. Consider first the case $\omega_{I(k)}(t, y, c) \not\equiv 0$, and assume without loss of generality that $\sigma_{I(k)}(t, y(t;\tilde{c}), \tilde{c}) > 0$ for $t < s_k(\tilde{c})$. Then, the conditions (7.73) and (7.75) ensure that the switching function has, in the neighborhood of $s_k(\tilde{c})$, a qualitative behavior that is represented either by Figure 7.2a or by Figure 7.2b: The switching function becomes zero at $s_k(\tilde{c})$ with non-zero, i.e. negative time derivative, and it is either positive or negative after the impulse. Then, if $y^+(t;c)$ depends continuously differentiable on the parameters, this guarantees that the switching function remains non-zero even for small changes in the parameters.



(a) (b) (c)

Figure 7.2.: Behavior of switching functions according to the regularity condition (RS) at a zero. Figures (a), (b) represent the case of a non-zero impulse and Figure (c) represents the case without impulse.

For the case that $\omega_{I(k)}(t,y,c) \equiv 0$, the condition (7.73) ensures that the switching functions becomes zero at $s_k(\tilde{c})$ with non-zero, i.e. negative time derivative. The positive sign is, for $t > s_k(\tilde{c})$, inconsistent, and if $y^-(t;c)$ is continuously differentiable with respect to the parameters, the positive sign remains inconsistent even for small changes in the parameters.[5] Since a solution $y(t;\tilde{c})$ exists, it is guaranteed that there exists a consistent choice for the switching function sign. Since, the old sign +1 is inconsistent, it follows $\zeta_{I(k)}^{k+1} = -1$ (note: $\zeta_{I(k)}^{k+1}$ is the sign of the switching function with index $I(k)$ in the interval (s_k, s_{k+1})). Because of regularity condition (7.74), the switching function leaves the zero right-sided with non-zero, i.e. negative time derivative. If $y^+(t;c) = y^-(t;c)$ depends continuously differentiable on the parameters, this guarantees that the negative sign remains consistent even for small changes in the parameters.

With this geometric interpretation of the regularity assumption (RS) in mind, the next step is the proof of Theorem 7.10.

Proof (of Theorem 7.10)
The proof relies again on the fact that the solution of the IHODE-IVP is locally equivalent to an ODE-IVP. More precisely, since there is no root discontinuity at $t^{ini}(c)$, the IHODE-IVP is locally equivalent to an ODE-IVP with right-hand-side $f_{ODE}(t, \mathfrak{y}(t;c), c) := f(t, \mathfrak{y}(t;c), c, \zeta^1)$. Since $y(t^{ini}(\tilde{c});\tilde{c}) \in \mathcal{V}^y$, with $\mathcal{V}^y$ being an open set, and because the right hand side function f fulfills the conditions (S) and (B), it is possible to find for every c sufficiently close to $\tilde{c}$ a closed set $\mathcal{D}^y \subset \mathcal{V}^y$ and an interval $\left[t^{ini}(c), t^{ini}(c) + \delta t_1\right]$ with $\delta t_1 > 0$, so that Theorem 7.2 can be applied to the constructed ODE-IVP.[6] From this it follows that the ODE-IVP solution is unique, continuously differentiable with respect to the parameters and that the Wronskian is given by the solution of a variational ODE-IVP.

At the time $t^{ini}(\tilde{c}) + \delta t_1$ the same argument can be used for the ODE-IVP on the intervals $\left[t^{ini}(\tilde{c}) + \delta t_1, t^{ini}(\tilde{c}) + \delta t_2\right]$, and subsequently for $\left[t^{ini}(\tilde{c}) + \delta t_2, t^{ini}(\tilde{c}) + \delta t_3\right]$, and so on, until the time point of the first root discontinuity, $s_1(\tilde{c})$, is reached. In fact, since also $y(s_1(\tilde{c});\tilde{c})$ is in the open set $\mathcal{V}^y$, uniqueness and differentiability of the ODE-IVP solution follow also for times beyond $s_1(\tilde{c})$, which is important because the discontinuity point $s_1(c)$ generally varies with the parameters.

On the interval $\left[t^{ini}(c), s_1(c)\right)$, it follows due to the equivalence of ODE-IVP and IHODE-IVP, that also the IHODE-IVP solution is unique and continuously partially differentiable (assertions 1 and 2). Further, the Wronskian can be expressed as solution of the variational ODE-IVP (7.79) (assertion 5).

At the time point $s_1(c)$ of the first root discontinuity, the time derivative of the switching function $\sigma_{I(1)}$ is non-zero according to assumption (RS). This allows to use the implicit function theorem, which gives differentiability of the discontinuity point $s_1(c)$ with respect to the parameters, and in particular

$$\frac{ds_1(c)}{dc} = -\frac{\frac{\partial \sigma_{I(1)}}{\partial y}\mathbf{W}^-(s_1(c);c) + \frac{\partial \sigma_{I(1)}}{\partial c}}{\frac{\partial \sigma_{I(1)}}{\partial t} + \frac{\partial \sigma_{I(1)}}{\partial y} f_1^-(c)}. \tag{7.85}$$

Therein, the partial derivatives of $\sigma_{I(1)}$ are evaluated at $(s_1, y^-(s_1;c), c)$. This verifies assertion 3. It further follows from elementary differentiation rules that

$$\begin{aligned}\frac{dy^-(s_1(c);c)}{dc} &= \frac{dy^-(s_1(c);c)}{dt}\frac{ds_1(c)}{dc} + \frac{\partial y^-(s_1(c);c)}{\partial c} \\ &= f_1^-(c)\frac{ds_1(c)}{dc} + \mathbf{W}^-(s_1(c);c).\end{aligned} \tag{7.86}$$

Further, by recalling that the right-sided limit of the state at the time point of the root discontinuity is given by $y^+(s_1(c);c) = y^-(s_1(c);c) + \omega_{I(1)}(s_1(c), y^-(s_1(c);c), c)$, the following relation is obtained:

$$\frac{dy^+(s_1(c);c)}{dc} = \frac{dy^-(s_1(c);c)}{dc} + \frac{\partial \omega_{I(1)}}{\partial t}\frac{ds_1(c)}{dc} + \frac{\partial \omega_{I(1)}}{\partial y}\frac{dy^-(s_1(c);c)}{dc} + \frac{\partial \omega_{I(1)}}{\partial c}. \tag{7.87}$$

Therein, the partial derivatives of $\omega_{I(k)}$ have to be evaluated at $(s_1(c), y^-(s_1(c);c), c)$.

[5] Showing that $y^-(t;c)$ is indeed continuously differentiable is done in the formal proof of the theorem.

[6] The increment δt_1 only has to be chosen small enough such that the bound M_f in condition (B) of Theorem 7.10 fulfills the more restrictive bound (B) of Theorem 7.2.

Finally, it also holds that

$$\mathbf{W}^+(s_1(c);c) = \frac{dy^+(s_1(c);c)}{dc} - f_1^+(c)\frac{ds_1(c)}{dc}, \tag{7.88}$$

and by inserting the equations (7.86) and (7.87) into equation (7.88), assertion 4 is verified.

For the continuation of the solution to the right of the discontinuity point $s_1(c)$, assumption (RS) is exploited: it guarantees, for both cases $\omega_{I(k)}(t,y,c) \equiv 0$ and $\omega_{I(k)}(t,y,c) \not\equiv 0$, that there is a unique consistent choice of the sign of the relevant switching function $\sigma_{I(k)}$ for all c in a neighborhood $\mathcal{U}^c$ of $\tilde{c}$ (recall Figure 7.2 and the corresponding discussion). Further, also all remaining switching functions have a unique consistent choice of the corresponding signs because of equation (7.76).

It is then possible to continue, for each c sufficiently close to $\tilde{c}$, the solution of the IHODE-IVP by solving a locally equivalent ODE-IVP. The arguments can then be repeated on all subintervals $[s_i(c), s_{i+1}(c))$. By choosing, if necessary, a smaller neighborhood $\mathcal{U}^c$ of $\tilde{c}$ such that the total number and temporal order of discontinuities in $\mathcal{T}(c)$ does not change, the proof is completed. ■

In comparison to Theorem 4.8 (uniqueness of IHODE-IVP solutions), Theorem 7.10 only requires a higher degree of smoothness of the model functions and the additional condition (RS). In this context it should be emphasized that the equation (7.75) for the case $\omega_{I(k)}(t,y,c) \not\equiv 0$ is sufficient, but not necessary. In the case that a switching function remains zero after the impulse for all c in a neighborhood of $\tilde{c}$, it is possible to use condition (7.74) instead and requiring, in addition, that the opposite sign $-\zeta^2_{I(k)}$ is inconsistent for a neighborhood of $\tilde{c}$. This can be ensured, e.g., by an appropriate condition on the time derivative of $\sigma_{I(k)}$ for this opposite sign choice.

7.5. DDEs with State-Dependent Delay Functions

The technique used in the previous section for proving differentiability of IHODE-IVP solutions can immediately be transferred to DDE-IVPs with state-dependent delays. This leads to the following theorem.

Theorem 7.11 (Global Differentiability of DDE-IVP Solutions)

Consider a DDE-IVP as in Definition 1.12, with nominal parameters $\tilde{c}$ and a neighborhood $\mathcal{V}^c$ of $\tilde{c}$. Let $\mathcal{V}^y \subset \mathbb{R}^{n_y}$ be an open domain and let $y : \mathcal{T}^f(c) \to \mathcal{V}^y$, $y : (t,\tilde{c}) \to y(t;\tilde{c})$ be a solution of the DDE-IVP for the nominal parameters. Choose $\Delta t > 0$ such that the interval $\mathcal{T}^{f,\Delta t}(\tilde{c})$ is defined.

Let $r_1(\tilde{c}), \dots, r_{n_r}(\tilde{c})$, $r_i(\tilde{c}) \in \mathcal{T}^f(\tilde{c})$, be the time points of the n_r critical discontinuities in $y(t;\tilde{c})$ for $t \in \mathcal{T}^f(\tilde{c})$, n_r^ϕ of which are located in $(-\infty, t^{ini}(\tilde{c}))$; it is possible that $r_{n_r^\phi+1}(\tilde{c}) = t^{ini}(\tilde{c})$. Further, let $s_1(\tilde{c}), \dots, s_{n_s}(\tilde{c})$ denote the time points of the children of critical discontinuities in $y(t;\tilde{c})$ for $t \in \mathcal{T}(\tilde{c})$, and define $s_0(c) := t^{ini}(c)$, $s_{n_s+1}(c) := t^{fin}(c)$. Let

$$\zeta^{\alpha,k+1}_{l,r_j(\tilde{c})} = \zeta^{\alpha}_{l,r_j(\tilde{c})}(t) \quad for \quad t \in (s_k(\tilde{c}), s_{k+1}(\tilde{c})) \quad for \quad 1 \leqslant k \leqslant n_s+1, \tag{7.89}$$

denote the simplified signs of the propagation switching function (see Definition 2.10), i.e.

$$\zeta^{\alpha}_{l,r_j(\tilde{c})}(t) := sign^+(\alpha_l(t, y(t;\tilde{c}), \tilde{c}) - r_j(\tilde{c})) \tag{7.90}$$

on the subintervals.

In addition, if a time point $r_j(\tilde{c})$ of a critical discontinuity is propagated to a discontinuity point $s_k(\tilde{c})$ with the deviating argument α_l, then let $I_1 : \{1,\dots,n_s\} \to \{1,\dots,n_r\}$ be the function that maps the index k of the time point $s_k(\tilde{c})$ of the child discontinuity to the index $I_1(k) = j$ of the time point $r_{I_1(k)}(\tilde{c})$ of the critical parent discontinuity. Further, let $I_2 : \{1,\dots,n_s\} \to \{1,\dots,n_\tau\}$ be the function that maps k to the index $I_2(k) = l$ of the according deviating argument.[7]

Let the following assumptions be fulfilled:

(S) Smoothness*: The initial time $t^{ini}(c)$, the final time $t^{fin}(c)$, and the initial value $y^{ini}(c)$ are continuously differentiable functions for $c \in \mathcal{V}^c$. The right-hand-side function $f(t,y,c,\{v_i\}_{i=1}^{n_\tau})$*

[7] Note that for every $r_j(\tilde{c}) > t^{ini}(\tilde{c})$ there is some $i \in \{1,\dots,n_s\}$ such that $r_j(\tilde{c}) = s_i(\tilde{c})$, because all critical discontinuities that are located in $(t^{ini}(\tilde{c}), t^{fin}(\tilde{c}))$ are also children of discontinuities of order 0 in y and thus children of critical discontinuities.

is continuous in t, continuously differentiable with respect to y, c, and $\{v_i\}_{i=1}^{n_\tau}$, and uniformly Lipschitz continuous with respect to y for $(t,y,c,\{v_i\}_{i=1}^{n_\tau}) \in \mathcal{T}^{f,\Delta t} \times \mathcal{V}^y \times \mathcal{V}^c \times (\mathcal{V}^y)^{n_\tau}$. The delay functions $\tau_i(t,y,c)$ are continuously differentiable with respect to all arguments and Lipschitz continuous with respect to y for $(t,y,c) \in \mathcal{T}^{f,\Delta t} \times \mathcal{V}^y \times \mathcal{V}^c$.

The time points of the critical discontinuities in the initial function ϕ, i.e. $r_i(c)$ for $1 \leqslant i \leqslant n_r^\phi$, are continuously differentiable functions for $c \in \mathcal{V}^c$. The initial function $\phi(\cdot,c)$ has a representation (7.50), and all functions $\phi_i(\cdot,c)$ are continuously differentiable with respect to both t and c and Lipschitz continuous with respect to t for $(t,c) \in \mathcal{T}^{f,\Delta t}(\tilde{c}) \times \mathcal{V}^c$.

(B) Boundedness*: The right-hand-side function f is bounded by*

$$\|f(t,y,c,\{v_i\}_{i=1}^{n_\tau})\|_\infty < M_f < \infty \tag{7.91}$$

for $(t,y,c,\{v_i\}_{i=1}^{n_\tau}) \in \mathcal{T}^{f,\Delta t} \times \mathcal{V}^y \times \mathcal{V}^c \times (\mathcal{V}^y)^{n_\tau}$.

(NVD) Non-Vanishing Delays*: It holds that $\tau_i(t,y(t;\tilde{c},\tilde{c}) \geqslant \underline{\tau} > 0$ for $1 \leqslant i \leqslant n_\tau$, $t \in \mathcal{T}(\tilde{c})$ and the considered solution $y(t;\tilde{c})$.*

(RS) Regularity of the Propagation Switching Functions*: It holds for the propagation switching functions of critical discontinuities, i.e. $\sigma^\alpha_{i,r_j(c)}(t,y,c) = \alpha_i(t,y,c) - r_j(c)$, that*

$$\sigma^\alpha_{i,r_j(\tilde{c})}(t^{ini}(\tilde{c}), y(t^{ini}(\tilde{c});\tilde{c},\tilde{c}),\tilde{c}) \neq 0 \quad for \quad 1 \leqslant i \leqslant n_\tau, \quad 1 \leqslant j \leqslant n_r^\phi \tag{7.92a}$$

$$\sigma^\alpha_{i,r_j(\tilde{c})}(t^{fin}(\tilde{c}), y(t^{fin}(\tilde{c});\tilde{c},\tilde{c}),\tilde{c}) \neq 0 \quad for \quad 1 \leqslant i \leqslant n_\tau, \quad 1 \leqslant j \leqslant n_r. \tag{7.92b}$$

Further, it holds for $s_k(\tilde{c})$, $1 \leqslant k \leqslant n_s$, with $I_1(k) = j$ and $I_2(k) = l$ that

$$\left[\frac{\partial \sigma^\alpha_{l,r_j(c)}(t,y,c)}{\partial t} + \frac{\partial \sigma^\alpha_{l,r_j(c)}(t,y,c)}{\partial y} f_k^-(c)\right]_{(t,y,c)=(s_k(\tilde{c}),y(s_k(\tilde{c});\tilde{c}),\tilde{c})} \neq 0. \tag{7.93}$$

If $r_j(c)$ is the time point of a discontinuity of order 0 in y, then it holds in addition that

$$\left[\frac{\partial \sigma^\alpha_{l,r_j(c)}(t,y,c)}{\partial t} + \frac{\partial \sigma^\alpha_{l,r_j(c)}(t,y,c)}{\partial y} f_k^+(c)\right]_{(t,y,c)=(s_k(\tilde{c}),y(s_k(\tilde{c});\tilde{c}),\tilde{c})} \neq 0, \tag{7.94}$$

and further that

$$\sigma^\alpha_{l',r_{j'}(c)}(s_k(\tilde{c}), y(s_k(\tilde{c});\tilde{c}),\tilde{c}) \neq 0 \tag{7.95}$$

for $l' \neq l$ and all critical discontinuities $r_{j'}(c)$, $j' \neq j$, that are of order 0 in y.

In equations (7.93), (7.94), $f_k^-(c)$ and $f_k^+(c)$ are defined by

$$\begin{aligned} f_k^-(c) =& f(s_k(c), y(s_k(c);c), c, \\ & \{y(s_k(c) - \tau_{l'}(s_k(c), y(s_k(c);c), c); c)\}_{l'=1,l'\neq l}^{n_\tau}, y^{\bullet,k}(r_j(c);c)) \end{aligned} \tag{7.96a}$$

$$\begin{aligned} f_k^+(c) =& f(s_k(c), y(s_k(c);c), c, \\ & \{y(s_k(c) - \tau_{l'}(s_k(c), y(s_k(c);c), c); c)\}_{l'=1,l'\neq l}^{n_\tau}, y^{\bullet,k+1}(r_j(c);c)) \end{aligned} \tag{7.96b}$$

where

$$y^{\bullet,k'}(r_j(c);c) := y^{\pm}(r_j(c);c) \quad if \quad \zeta^{\alpha,k'}_{l,r_j(c)} = \pm 1 \quad for \quad k' \in \{k, k+1\} \tag{7.97}$$

Note that it has been taken into account in equation (7.96) that the state is continuous in all past time points except for one.

Then there exists $\delta t > 0$, an open interval $\mathcal{T}^{f,\delta t}(\tilde{c})$, and a neighborhood $\mathcal{U}^c$ such that $\mathcal{T}^f(c) \subset \mathcal{T}^{f,\delta t}(\tilde{c})$, and for $(t,c) \in \mathcal{T}^{f,\delta t}(\tilde{c}) \times \mathcal{U}^c$ the following assertions hold.

1. *There exists a unique solution $y(t;c)$.*

2. *The Wronskian* $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$ *is continuous for* $t > t^{ini}(c)$, $t \neq s_k(c)$ *for* $1 \leqslant k \leqslant n_s$, *and right-continuous in* $t = s_k(c)$, $1 \leqslant k \leqslant n_s$.

3. *The discontinuity points* $s_k(c)$, $1 \leqslant k \leqslant n_s$, *are continuously differentiable functions of the parameters* c, *and*

$$\frac{ds_k(c)}{dc} = -\frac{\frac{\partial \alpha_l}{\partial y}\mathbf{W}^-(s_k(c);c) + \frac{\partial \alpha_l}{\partial c} - \frac{dr_j(c)}{dc}}{\frac{\partial \alpha_l}{\partial t} + \frac{\partial \alpha_l}{\partial y} f_k^-(c)}, \tag{7.98}$$

where $j = I_1(k)$, $l = I_2(k)$, *and where the partial derivatives of* α_l *have to be evaluated at* $(s_k(c), y(s_k(c);c), c)$.

4. *At the time points* $s_k(c)$, $1 \leqslant k \leqslant n_s$, *the jump in the Wronskian* $\mathbf{W}(t;c)$ *(i.e. the difference between left-sided limit and right-sided limit) is given by*

$$\mathbf{W}^+(s_k(c);c) - \mathbf{W}^-(s_k(c);c) = (f_k^-(c) - f_k^+(c))\frac{ds_k(c)}{dc}, \tag{7.99}$$

where $ds_k(c)/dc$ *is given by equation (7.98).*

5. *On the right-open interval* $[s_k(c), s_{k+1}(c))$, $0 \leqslant k \leqslant n_s$, *the Wronskian* $\mathbf{W}(t;c)$ *is given as the solution of the variational DDE-IVP*

$$\begin{aligned}\dot{\mathfrak{W}}(t;c) =& \frac{\partial f}{\partial y}\mathfrak{W}(t;c) + \frac{\partial f}{\partial c} \\ &+ \sum_{m=1}^{n_\tau} \frac{\partial f}{\partial v_m}\left(\dot{y}(t-\tau_m(t,y(t;c),c);c)\left[-\frac{\partial \tau_m}{\partial y}\mathfrak{W}(t;c) - \frac{\partial \tau_m}{\partial c}\right]\right. \\ &\left. + \mathfrak{W}(t-\tau_m(t,y(t;c),c);c);c\right)\end{aligned} \tag{7.100a}$$

$$\mathfrak{W}(s_k(c);c) = \frac{dy_k(c)}{dc} - f_k^+(c)\frac{ds_k(c)}{dc} \tag{7.100b}$$

$$\mathfrak{W}(t;c) = \begin{cases} \frac{\partial \phi(t,c)}{\partial c} & \text{for} \quad t < t^{ini}(c) \\ \text{solution of problem} \quad (7.100) \quad \text{in} \quad [s_j(c), s_{j+1}(c)) & \\ \text{for} \quad t \in [s_j(c), s_{j+1}(c)), j \leqslant k & \end{cases}. \tag{7.100c}$$

Herein, $y_k(c) = y^{ini}(c)$ *for* $k = 0$ *and* $y_k(c) = y^-(s_k(c);c)$ *for* $k > 0$ *is the initial state for the corresponding interval. The derivative with respect to the parameters is given by* $dy_0(c)/dc = dy^{ini}(c)/dc$ *and*

$$\frac{dy_k(c)}{dc} = \mathbf{W}^-(s_k(c);c) + f_k^-(c)\frac{ds_k(c)}{dc} \quad \text{for} \quad k > 0. \tag{7.101}$$

The partial derivatives of f *in equation (7.100) have to be evaluated at* $(t, y(t;c), c, \{y(t-\tau_i(t,y(t;c),c))\}_{i=1}^{n_\tau})$, *and the partial derivatives of* τ_m *have to be evaluated at* $(t, y(t;c), c)$.

The time derivative of the state, $\dot{y}$, *and the Wronskian,* $\mathbf{W}$, *have to be evaluated at the past time points given by the deviating arguments, see equation (7.100). For the special case that* $t = s_k(c)$ *and* $m = l = I_2(k)$, *the left-sided limit is taken if* $\zeta_{l,s_j(\tilde{c})}^{\alpha,k+1} = -1$, *and the right-sided limit is taken if* $\zeta_{l,s_j(\tilde{c})}^{\alpha,k+1} = +1$. *The same holds for the corresponding past state in the argument of the partial derivative of* f.

Finally, it is defined that

$$f_0^+(c) := f(t^{ini}(c), y(t^{ini}(c);c), c, \{y(t^{ini}(c) - \tau_i(t^{ini}(c), y(t^{ini}(c);c), c);c)\}_{i=1}^{n_\tau}). \tag{7.102}$$

6. *The total derivative of the state at the final time with respect to the parameters is given by*

$$\frac{dy(t^{fin}(c);c)}{dc} = \mathbf{W}^-(t^{fin}(c);c) + f_{n_s+1}^-(c)\frac{dt^{fin}(c)}{dc}, \tag{7.103}$$

with

$$f^{-}_{n_s+1}(c) := f(t^{fin}(c), y(t^{fin}(c);c), c, \{y(t^{fin}(c) - \tau_i(t^{fin}(c), y(t^{fin}(c);c), c); c)\}_{i=1}^{n_\tau}). \quad (7.104)$$

The proof of this theorem is mostly analogous to the proof of Theorem 7.10.

Proof
The delays are non-vanishing for the solution $y(t;\tilde{c})$, and in particular non-vanishing at the initial time. The DDE-IVP is therefore locally equivalent to an ODE-IVP, in which all past state arguments are replaced by evaluations of differentiable deduced functions. All critical propagation switching functions are non-zero, the initial state $y(t^{ini}(\tilde{c});\tilde{c})$ is in the open domain $\mathcal{V}^y$, and the DDE right-hand-side function f fulfills the conditions (S) and (B). For a sufficiently small neighborhood of $\tilde{c}$, it is therefore possible to find a closed set $\mathcal{D}^y$ and an interval $[t^{ini}(\tilde{c}), t^{ini}(\tilde{c}) + \delta t_1]$, $\delta t_1 > 0$, such that Theorem 7.2 can be applied. Since also $y(t^{ini}(\tilde{c}) + \delta t_1; \tilde{c})$ is in the open set $\mathcal{V}^y$, it is possible to continue this ODE-IVP solution successively until and beyond $s_1(\tilde{c})$, which gives uniqueness of the ODE-IVP solution and continuous differentiability with respect to the parameters. The Wronskian $\mathbf{W}$ can be expressed as the solution of a variational ODE-IVP.

Due to equivalence of DDE-IVP and ODE-IVP on $[s_0(c), s_1(c))$, existence, uniqueness and differentiability of the DDE-IVP solution follow for a neighborhood $\mathcal{U}^c$ of $\tilde{c}$ (assertions 1 and 2). With an appropriate change of notation, the variational ODE-IVP becomes a variational DDE-IVP (assertion 5).

For the discontinuity point $s_1(c)$, differentiability and thus assertion 3 follows from the implicit function theorem, because the left-sided time derivative of the propagation switching function $\sigma^{\alpha}_{l,r_j(\tilde{c})}$, $l = I_2(1)$, $j = I_1(1)$, is non-zero (equation (7.93)). Since the state y is continuous in the time points of the propagated discontinuities, it holds that $y^+(s_1(c);c) = y^-(s_1(c);c)$, and by elementary differentiation rules, equations (7.101) and (7.99) follow (assertion 4).

For the continuation of the DDE-IVP solution to the right of $s_1(c)$, consider at first the case that $s_1(c)$ is the time point of a critical discontinuity itself (i.e. $s_1(c)$ is the time point of a child discontinuity whose parent discontinuity is of order 0 in y). In this case, the equations (7.93) and (7.94) ensure that there is a unique consistent choice for the sign of the relevant propagation switching function $\sigma^{\alpha}_{l,r_j(\tilde{c})}$, $l = I_2(1)$, $j = I_1(1)$, for $\tilde{c}$ and a sufficiently small neighborhood. In the case that $s_k(c)$ is not the time point of a critical discontinuity, the argument is the same, but the condition (7.94) is not needed because it is implied by equation (7.93). All other propagation switching functions of discontinuities of order 0 in y are non-zero, according to equation (7.95).

Hence, there is a unique consistent choice of the signs of all critical propagation switching functions. The DDE-IVP solution can then again be replaced by a locally equivalent ODE-IVP on the interval $[s_1(c), s_2(c))$. By repeating the arguments on all subintervals, the proof is completed.■

7.6. The General Case: IHDDEs

A differentiability result for IHDDE-IVP solutions can be obtained by combining the ideas that led to the Theorems 7.10 and 7.11. In particular, it needs to be ensured that root discontinuities do not coincide neither with other root discontinuities nor with children of critical discontinuities (neither before nor after a possible non-zero impulse). Moreover, children of discontinuities of order 0 in y should not coincide.

As a result, one obtains again that the Wronskian is piecewise given as solution of a variational IVP of the form (7.100), except that the partial derivatives of f now include the switching function signs ζ as an additional argument. Moreover, there are again jumps in the Wronskian $\mathbf{W}$ that need to be taken into account. They may occur at the time points of root discontinuities and at the time points of child discontinuities whose parent discontinuity is of order 0 in y.

Since the formalization of the differentiability result is mainly a technicality, it is omitted here. Instead, only the expressions for the jumps in the Wronskian are discussed in the following.

For children of discontinuities of order 0 in y, the expressions given in Theorem 7.11, namely equations (7.98) and (7.99), remain valid. It only needs to be taken into account that $f_k^-(c)$ and $f_k^+(c)$ are defined in such a way that they get – compared to equations (7.96) - the switching function signs as an additional argument.

The expression for the jump in the Wronskian $\mathbf{W}$ at root discontinuities in IHDDE-IVPs is, however, more involved. In order to derive this expression, it is first observed that the total

derivative of the time point of a root discontinuity with respect to the parameters is given by

$$\begin{aligned}\frac{ds_k(c)}{dc} = &- \frac{1}{\frac{\partial \sigma_{I(k)}}{\partial y} f_k^-(c) + \frac{\partial \sigma_{I(k)}}{\partial t} + \sum_{m=1}^{n_\tau} \frac{\partial \sigma_{I(k)}}{\partial v_m} \dot{y}_{past}^{k,m} \left[1 - \frac{\partial \tau_m}{\partial t} - \frac{\partial \tau_m}{\partial y} f_k^-(c)\right]} \\ &\cdot \left\{ \frac{\partial \sigma_{I(k)}}{\partial y} \mathbf{W}^-(s_k(c); c) + \frac{\partial \sigma_{I(k)}}{\partial c} \right. \\ &\left. + \sum_{m=1}^{n_\tau} \frac{\partial \sigma_{I(k)}}{\partial v_m} \left[\dot{y}_{past}^{k,m} \left(-\frac{\partial \tau_m}{\partial y} \mathbf{W}^-(s_k(c); c) - \frac{\partial \tau_m}{\partial c} \right) + \mathbf{W}_{past}^{k,m} \right] \right\}. \end{aligned} \tag{7.105}$$

Herein, the partial derivatives of the switching function $\sigma_{I(k)}$ are evaluated at the arguments $(s_k(c), y^-(s_k(c); c), c, \{y(s_k(c) - \tau_i(s_k(c), y^-(s_k(c); c), c); c)\}_{i=1}^{n_\tau})$, and the partial derivatives of the delay functions τ_m are evaluated at $(s_k(c), y^-(s_k(c); c), c)$. Further, it holds that

$$f_k^-(c) = f(s_k(c), y^-(s_k(c); c), c, \{y(s_k(c) - \tau_i(s_k(c), y^-(s_k(c); c), c); c)\}_{i=1}^{n_\tau}, \zeta^k), \tag{7.106}$$

where ζ^k are the switching function signs to the left of $s_k(c)$. Eventually, the quantities $\dot{y}_{past}^{k,m}$ and $\mathbf{W}_{past}^{k,m}$ are given by

$$\dot{y}_{past}^{k,m} = \dot{y}(s_k(c) - \tau_m(s_k(c), y^-(s_k(c); c), c); c) \tag{7.107a}$$

$$\mathbf{W}_{past}^{k,m} = \mathbf{W}(s_k(c) - \tau_m(s_k(c), y^-(s_k(c); c), c); c). \tag{7.107b}$$

Note that the past states, the past time derivatives $\dot{y}_{past}^{k,m}$, and the past Wronskians $\mathbf{W}_{past}^{k,m}$ are continuous at the time point of evaluation because it is assumed that root discontinuities do not coincide with children of critical discontinuities.

With $ds_k(c)/dc$ being given by equation (7.105), the jump in the Wronskian matrix at the time point of a root discontinuity can be expressed as

$$\begin{aligned}\mathbf{W}^+(s_k(c); c) - \mathbf{W}^-(s_k(c); c) = &\left[\left(\mathbf{1}_{n_y} + \frac{\partial \omega_{I(k)}}{\partial y} \right) f_k^-(c) + \frac{\partial \omega_{I(k)}}{\partial t} - f_k^+(c) \right. \\ &\left. + \sum_{m=1}^{n_\tau} \frac{\partial \omega_{I(k)}}{\partial v_m} \dot{y}_{past}^{k,m} \left(1 - \frac{\partial \tau_m}{\partial t} - \frac{\partial \tau_m}{\partial y} f_k^-(c)\right) \right] \frac{ds_k(c)}{dc} \\ &+ \frac{\partial \omega_{I(k)}}{\partial y} \mathbf{W}^-(s_k(c); c) + \frac{\partial \omega_{I(k)}}{\partial c} \\ &+ \sum_{m=1}^{n_\tau} \frac{\partial \omega_{I(k)}}{\partial v_m} \left[\dot{y}_{past}^{k,m} \left(-\frac{\partial \tau_m}{\partial y} \mathbf{W}^-(s_k(c), c) - \frac{\partial \tau_m}{\partial c} \right) \right. \\ &\left. + \mathbf{W}_{past}^{k,m} \right]. \end{aligned} \tag{7.108}$$

with

$$f_k^+(c) = f(s_k(c), y^-(s_k(c); c), c, \{y(s_k(c) - \tau_i(s_k(c), y^-(s_k(c); c), c); c)\}_{i=1}^{n_\tau}, \zeta^{k+1}), \tag{7.109}$$

where ζ^{k+1} are the switching function signs to the right of $s_k(c)$.

8. Numerical Sensitivity Computation

> *Summing up, internal numerical differentiation leads to a drastic reduction of computing time (60-80 % especially for low tolerances) compared to external numerical differentiation because of substantial overhead savings (...) and much lower accuracy requirements for the basic integration scheme.*
>
> Bock, in the paper "Recent advances in parameter identification techniques for ODE" [38], summarizing the advantages of Internal Numerical Differentiation for the computation of derivatives of initial value problem solutions with respect to parameters.

The definitions of an impulsive hybrid discrete-continuous delay differential equation (IHDDE) and of the corresponding initial value problem (IHDDE-IVP) in Chapter 1 were formulated in such a way that all model functions depend on parameters $c \in \mathbb{R}^{n_c}$. Accordingly, the solution of an IHDDE-IVP depends not only on the time t but also on the parameters; it is therefore denoted by $y(t;c)$. In the previous chapter, sufficient conditions were given under which IVP solutions depend differentiably on the parameters. More precisely, if these sufficient conditions are fulfilled, then there exists an open domain $\mathcal{U}^c \subset \mathbb{R}^{n_c}$ such that the Wronskian matrix

$$\mathbf{W}(t;c) = \frac{\partial y(t;c)}{\partial c} \tag{8.1}$$

is, for any $c \in \mathcal{U}^c$, a piecewise continuously differentiable function of time with potential jumps at the time points of root discontinuities and at the time points of propagated discontinuities. Throughout the chapter, the derivatives of the IVP solution with respect to the parameters are often shortly called *sensitivities*. This term is motivated by the fact that the derivatives measure how sensitive the IVP solution is with respect to changes in the parameters.

In practice, solutions of IVPs are of interest for specific values $\tilde{c}$ of the parameters called the *nominal parameters*. Numerical methods for the approximation of the function $y(t;\tilde{c})$ were presented in the Chapters 5 and 6. The results of Chapter 7 provide means to assess whether IVP solutions are differentiable with respect to c, i.e. whether $\mathbf{W}(t;\tilde{c})$ exists. The next logical step is, of course, to present and analyze numerical methods for the practical computation of the Wronskian $\mathbf{W}(t;\tilde{c})$. This is the topic of this chapter.

Literature Survey

A simple and straightforward way for sensitivity computation that can be used on any differentiable function is the computation of a difference quotient. For example, if an IVP solver is available, then it can be called twice: once for computing the solution $y(t;\tilde{c})$ for the nominal parameter values $\tilde{c}$, and once for computing a solution for slightly varied parameter values $\tilde{c} + \Delta c$. By computing a difference quotient from these two integration results, a numerical approximation of the derivative in the direction Δc of the parameter variation can be obtained.

This approach is also referred to as *External Numerical Differentiation*, because the differentiation takes place outside of the integrator. An alternative name occasionally found in the literature is "brute-force method (for differentiation)". Derivative computation by this approach can be easily realized. However, the result of practical variable-stepsize IVP solvers is typically a discontinuous function of the parameters, with jumps that are of the order of magnitude of the chosen relative tolerance. This is a consequence of logical decisions in the IVP solvers, e.g. in the error control strategy. Therefore, External Numerical Differentiation requires the computation of highly accurate IVP solutions even if only a low or medium accuracy is required for the sensitivities. This makes the External Numerical Differentiation approach very inefficient.

Several authors have reported this drawback of External Numerical Differentiation, see e.g. the early accounts of Bard [19], Gear and Vu [113], and Bock [36, 38]. These works, as well as a large number of later publications, have dealt with the development of numerical methods that are superior to External Numerical Differentiation regarding reliability and efficiency. Most of the proposed methods fit into the following categorization.

1. Methods for computing *forward sensitivities* aim, like the External Numerical Differentiation method, at the computation of the derivative of all components of the state vector with respect to one or several perturbations in parameter space. This corresponds to the computation of the columns of the Wronskian matrix $\mathbf{W}(t;\tilde{c})$ (or linear combinations of the columns). Two subclasses of methods for forward sensitivity computation are given are distinguished.

   - *Numerical solution of a variational initial value problem by discretization (first differentiate, then discretize)*:

     In view of the results of Chapter 7, it is obvious that one way to compute sensitivities of IVP solutions with respect to parameters is the solution of the corresponding variational initial value problem. In case that root discontinuities or propagated discontinuities occur in the numerical solution that give rise to jumps in the Wronskian, discrete analogues of the expressions for the jump in the Wronskian (see Chapter 7) can be derived and evaluated.

     Examples for early works that approach sensitivity analysis for ordinary differential equations (ODEs) in this way are due to Dickinson and Gelinas [77] and Dunker [83], who called it the "direct method" for sensitivity analysis. Caracotsios and Stewart [59] use the same approach on differential-algebraic equations. Galán, Feehery, and Barton [111] follow this idea for impulsive hybrid discrete-continuous ordinary differential equations (IHODEs). In order to compute the sensitivities in this case, they evaluate the jump expressions for the Wronskian in the time points of the root discontinuities. These jump expressions involve, in particular, the derivative of the time point of the root discontinuity with respect to the parameters. For delay differential equations, the computation of sensitivities by solving a variational IVP and taking into account the jump expressions has been proposed by Bock and Schlöder [44]. Later, this approach has also been used by ZivariPiran [271] and ZivariPiran and Enright [273], also for delay differential equations of so-called "neutral type".

   - *Differentiation of the adaptively generated discretization scheme (first discretize, then differentiate)*:

     Here, the basic idea is to take the derivative of the adaptively generated discretization scheme, under the condition that all logical and discrete decisions – as they occur e.g. in common variable-order, variable-stepsize strategies – are kept fixed. This condition ensures that the call of the numerical integration method can be regarded as a sequence of differentiable mappings. Jumps in the Wronskian matrix can be computed by applying the formalism of Chapter 7 (in particular, the implicit function theorem) to the numerical solution of the nominal IVP.

     The basic idea of this approach goes back to Bock [36, 38], where it is used on extrapolation methods applied to ODEs. The method has therein been called *Internal Numerical Differentiation* as opposed to the earlier mentioned External Numerical Differentiation. Very early, Bock [39] has presented the extension to IHODEs. Von Schwerin, Winckler, and Schulz [252] use Internal Numerical Differentiation on a Runge-Kutta method applied to IHODE-IVPs arising in multi-body systems. Bauer [20], Albersmeyer [1, 2], and also Støren and Hertzberg [242] have used the approach on backward differentiation formulae applied to differential-algebraic equations.

   It is worthwile to remark that for linear integration methods (i.e. methods that are linear in the evaluations of the right-hand-side function f) applied to ODEs or differential-algebraic equations, both approaches lead formally to the same equation systems, see e.g. Bock [39], Körkel [164], and Sandu and Miehe [220]. Hence, if the same stepsizes are used, the differentiation and discretization operators commute. The efficiency and accuracy of the numerically computed sensitivities does therefore not so much depend on the point of view under which the equations are derived, but rather on how they are concretely solved in practice. A variety of propositions has been made in this respect.

In the literature of the "first differentiate, then discretize" approach, it is mainly distinguished between the *coupled direct method* (also called "simultaneous method") and the *decoupled direct method* (also called "staggered method"). The coupled method interpretes the nominal IVP and the variational IVP as one enlarged system of equations, see e.g. Dickinson and Gelinas [77] and Maly and Petzold [183]. Contrariwise, the decoupled method computes the step in the sensitivities independently after the step in the solution of the nominal IVP (but using the same integration stepsize). The decoupled method is, for implicit methods, typically much more efficient. For references on the decoupled direct method, see e.g. Dunker [83], Caracotsios and Stewart [59], and Feehery, Tolsma, and Barton [102]. An overview over the different practical realizations of the direct method is found in Li and Petzold [176].

The "first discretize, then differentiate" approach can be carried out rigorously such that differentiation is applied to every floating point operation in the numerical integration scheme, except for those that affect the adaptive components. This approach is usually called *iterative Internal Numerical Differentiation* and can be interpreted as a special realization of *Automatic Differentiation* (see Kedem [157], Griewank [119] and Griewank and Walther [120] for an introduction of this concept). There exist, however, variants that deviate from this rigorous interpretation. These variants are usually called *direct Internal Numerical Differentiation* and *Internal Numerical Differentiation with varied trajectories*. For overviews on different realizations of Internal Numerical Differentiation, see Bauer [20] and Albersmeyer [1, 2]. It is noted that Internal Numerical Differentiation methods presented in the literature have typically been derived under sophisticated structure exploitation, which has led to the development of highly efficient practical solvers.

2. Methods for computing *adjoint sensitivities*, also called "backward sensitivities", aim at the computation of the derivative of one or several scalar functions of the state vector with respect to all parameters. This corresponds to the computation of the rows of the Wronskian matrix $\mathbf{W}(t;\tilde{c})$ (or linear combinations of the rows). Similar to the computation of forward sensitivities, two subclasses of approaches are distinguished.

   - *Numerical solution of an adjoint initial value problem by discretization*: This approach relies, traditionally, on a Hilbert space scalar product of the variational differential equation with newly introduced *continuous adjoint variables*. From this, an *adjoint initial value problem* can be derived, whose numerical solution yields an approximation of the sought sensitivities. Examples for works in this direction in the context of ODEs and differential-algebraic equations are Cao, Li, Petzold [58], Sandu, Daescu, and Carmichael [219], and Alexe and Sandu [4]. For the discussion of the approach in the context of DDEs, it is referred to Koda [162] and Rihan [214].
   - *Discrete adjoint of a forward method for sensitivity computation*: This approach relies on multiplying the equations that are used for the numerical computation of forward sensitivities by *discrete adjoint variables*. From this, a *discrete adjoint scheme* can be derived for the numerical computation of the sensitivities. This approach has been first suggested by Bock [39], where it was used on Runge-Kutta methods applied to ODEs (see also Wirsching [258] and Kirches et al. [161]). Bock, Schlöder, and Schulz [45], Albersmeyer and Bock [3], and Albersmeyer [2] present the application of this approach to backward differentiation formulae for differential-algebraic equations.

   In contrast to forward sensitivity computation, discretization and differentiation do typically not commute, not even for linear integration methods applied to ODE-IVPs. For example, Bock [39] shows that Runge-Kutta methods have to fulfill rather restrictive conditions on their abscissae, weights, and coefficients in order to be self-adjoint in the sense that a discretization of the adjoint IVP is equivalent to the discrete adjoint scheme.

   As a consequence of the fact that discretization and differentiation do not generally commute for adjoint sensitivity computation, the intermediate quantities in the discrete adjoint scheme can not easily be related to the continuous solution of the adjoint IVP. In the context of backward differentiation formulae applied to ODE-IVPs, Beigel [23] has recently investigated the relationship between continuous adjoints and discrete adjoints by means of a functional analytic framework. Importantly, these results have also led to new efficient methods for goal-oriented global error estimation and control.

Occasionally, it has also been suggested to directly apply Automatic Differentiation to an existing integration scheme in forward mode, see Carmichael, Sandu, and Potra [60] and Ellwein et al. [88], or in adjoint mode, see Sandu, Daescu, and Carmichael [219]. However, since such a straightforward application of forward and adjoint Automatic Differentiation also takes the derivative of the adaptively chosen stepsize into account, this is not equivalent to forward and adjoint Internal Numerical Differentiation, respectively. Naive application of Automatic Differentiation may thus lead to relative errors of more than 100% even for simple test problems, see Eberhard and Bischof [84].

For completeness, it is further appropriate to mention the following related works on sensitivity analysis: Hwang et al. [153], Dougherty, Hwang, and Rabitz [81], Kramer and Calo [165] have introduced a so-called "Greens function method", which aims at the reduction of the computational costs for equation systems with many parameters. Gear and Vu [113] have proposed the construction of a smooth variable-stepsize strategy. Rabitz, Kramer, and Dacol [211], Turányi [250], and Kiehl [158] have published surveys on various methods for sensitivity analysis. Finally, there is a large number of papers that are application-oriented, where sensitivities of IVP solutions need to be computed for a specific differential equation model of a real-world process. In particular, in the context of DDEs, it is referred to Baker and Rihan [14], Horbelt, Timmer, and Voss [152], Reinecke [212], and Wu, Wang, and Shang [260].

Despite the considerable literature that is available in the area of sensitivity analysis for differential equation systems, it seems that only very few works have proposed general-purpose numerical methods for sensitivity analysis of delay differential equations (DDEs). Further, the numerical analysis of these methods is still immature. Hybrid discrete-continuous delay differential equations (HDDEs) and impulsive hybrid discrete-continuous delay differential equations (IHDDEs) have apparently not been studied at all in this respect. This chapter aims at improving on this unsatisfactory state of research by presenting several new results, as described in the following.

Novel Results Presented in This Chapter

This chapter presents and compares two numerical methods for the computation of forward sensitivities of DDE-IVP solutions: On the one hand, differentiation of the continuous Runge-Kutta scheme (CRK scheme) that is used for solving the nominal DDE-IVP, and, on the other hand, discretization of the variational DDE-IVP with the same CRK method. This yields the result that discretization and differentiation do not generally commute for CRK methods applied to DDEs – in contrast to a known result for Runge-Kutta methods applied to ODE-IVPs.

Subsequently, the convergence of the numerically computed sensitivities by the developed methods to the exact derivative of the exact IVP solution with respect to parameters is discussed. Furthermore, also the behavior of the local error in a single integration step is analyzed for both methods. Based on the findings regarding the behavior of the local and of the global error, a generalization of the concept of Internal Numerical Differentiation is proposed, and important aspects of error control strategies for numerically computed sensitivities are discussed. The generalization of the results for the treatment of IHDDE-IVPs is also part of this chaper.

This chapter furthermore introduces the first discrete adjoint scheme for the computation of sensitivities in the context of DDEs and IHDDEs. In particular, it is also discussed how the proposed discrete adjoint approach can be used to compute the derivatives of the state at inner time points of the considered interval.

Organization of This Chapter

Section 8.1 gives a short summary of previous chapters in order to recall some basic concepts and notations. Section 8.2 is about the computation of forward sensitivities. After analysing External Numerical Differentiation and discussing the reasons for its inefficiency, the "first discretize, then differentiate" and the "first differentiate, then discretize" approaches are used in the context of CRK methods applied to DDE-IVPs. Subsequently, the local and global errors of the two methods are studied. Section 8.3 derives the discrete adjoint scheme for the computation of sensitivities in the context of CRK methods applied to DDE-IVPs and IHDDE-IVPs. In particular, the contributions originating from jumps in the Wronskian are discussed in detail.

8.1. Short Summary of Previous Chapters

It is appropriate to start this chapter with a short summary of the concepts and notations that were introduced in the preceding chapters. More precisely, the notation for CRK methods and their realization in the framework of the modified standard approach is recalled, and the main results of the differentiability theory for DDE-IVP solutions are summarized.

8.1.1. CRK Methods and the Modified Standard Approach

Consider a DDE-IVP with a single but possibly state-dependent delay τ_1:

$$\dot{\mathfrak{y}}(t) = f(t, \mathfrak{y}(t), c, \mathfrak{y}(t - \tau_1(t, \mathfrak{y}(t), c))) \tag{8.2a}$$

$$\mathfrak{y}(t^{ini}(c)) = y^{ini}(c) \tag{8.2b}$$

$$\mathfrak{y}(t) = \phi(t, c) \quad \text{for} \quad t < t^{ini}(c). \tag{8.2c}$$

This DDE-IVP is considered on the interval $[t^{ini}(c), t^{fin}(c)]$. Like in Chapter 7, the DDE-IVP (8.2) is often referred to as *nominal DDE-IVP* in order to allow a better distinction from the *variational DDE-IVP*, which was introduced in Section 7.2 and which will also play a role in this section. The solution of problem (8.2), subsequently also called nominal DDE-IVP solution, depends on the parameters c and is therefore denoted by $y(t; c)$ (cf. Section 2.4).

For the solution of the nominal DDE-IVP, a CRK method with discrete local order p and uniform local order q is applied on a mesh $t_0 < t_1 < \cdots < t_{n_m}$, where $n_m + 1$ is the total number of mesh points. The method is initialized with $t_0 = t^{ini}(c)$, $y_0 = y^{ini}(c)$ and proceeds as follows: Given a discrete approximation y_l of $y(t_l; c)$ and a continuous approximation $\eta(t)$ of $y(t; c)$ for $t \leqslant t_l$, the discrete approximation y_{l+1} of $y(t_{l+1}; c)$ and the continuous representation $\eta_{l+1}(t)$ of $y(t; c)$ for $t \in [t_l, t_{l+1}]$ are obtained by:

$$y_{l+1} = y_l + h_{l+1} \sum_{j=1}^{\nu} \beta_j g_{l+1}^j \tag{8.3a}$$

$$\eta_{l+1}(t_l + \theta h_{l+1}) = y_l + h_{l+1} \sum_{j=1}^{\nu} b_j(\theta) g_{l+1}^j \tag{8.3b}$$

$$g_{l+1}^j = f(t_{l+1}^j, y_{l+1}^j, c, v_{l+1}^j) \tag{8.3c}$$

$$y_{l+1}^j = y_l + h_{l+1} \sum_{i=1}^{\nu} a_{j,i} g_{l+1}^i. \tag{8.3d}$$

In order to establish the theoretical results of this chapter, recall first the idealized variant of the modified standard approach as a technique for the computation of past states (cf. Subsection 5.2.2, and in particular Definition 5.16).

The idealized variant of the modified standard approach relies on a number of – quite restrictive – assumptions. The first assumption is that the exact solution $y(t; c)$ is unique and has finitely many discontinuities up to the discrete local order p of the numerical method. The time points of these discontinuities are denoted by s_j, with $-n_s^\phi \leqslant j \leqslant n_s$. Thereby, the discontinuity points s_j with $-n_s^\phi \leqslant j \leqslant -1$ are the time points of the initial discontinuities. Secondly, s_0 is by convention the initial time (even if $y(t)$ is smooth up to order p at that time). And thirdly, s_j with $1 \leqslant j \leqslant n_s$ are the time points of the propagated discontinuities in the interval $(t^{ini}(c), t^{fin}(c)]$.

For these discontinuity points s_j, it is possible to define the associated propagation switching functions

$$\sigma_{1,s_j}^{\alpha}(t, y(t; c), c) = \alpha_1(t, y(t; c), c) - s_j. \tag{8.4}$$

with the deviating argument $\alpha_1(t, y, c) := t - \tau_1(t, y, c)$. It is further possible to define the simplified signs of these propagation switching functions as

$$\zeta_{1,s_j}^{\alpha,+}(t) = \operatorname{sign}^+(\alpha_1(t, y^-(t; c), c) - s_j). \tag{8.5}$$

The simplified sign function sign^+ attributes, to the argument 0, the value 1, see equation (2.14). Note that these signs are defined by an evaluation of the propagation switching function along the

exact solution $y(t;c)$.

Given the signs $\zeta_{1,s_j}^{\alpha,+}(t)$, it is further possible to define the discontinuity interval indicator of the sole deviating argument α_1 as

$$\xi_1^\alpha(t) = n_s + 1 + \frac{1}{2}\sum_{j=-n_s^\phi}^{n_s}(\zeta_{1,s_j}^{\alpha,+}(t) - 1). \tag{8.6}$$

The idealized variant of the modified standard approach further relies on the assumption that $\xi_1^\alpha(t)$ has only finitely many discontinuities. This implies, in particular, that there is a finite number of discontinuities up to order $p+1$ in y.

Eventually, the idealized variant assumes that the discontinuity points s_j and the discontinuity interval indicator $\xi_1^\alpha(t)$ - evaluated for the exact solution $y(t)$ - are known to the numerical method, and that the mesh is chosen in such a way that it contains all of the finitely many discontinuity points in $\xi_1^\alpha(t)$.

In the remainder of this section and in Section 8.2, only a single integration step $t_l \to t_{l+1}$ of the CRK method is considered. Therefore, the notation can be simplified by using the symbol ξ as an abbreviation for $\xi_1^\alpha(t')$ for $t' \in (t_l, t_{l+1})$.

The key aspect of the modified standard approach – in both the idealized variant (Definition 5.16) and the practical variant (Definition 5.22) – is to use extrapolations beyond past discontinuity points, whenever the current integration step is such that the deviating arguments assumes values outside of the discontinuity interval indicated by ξ. In order to formalize this idea, so-called deduced functions $z_{\eta,\eta}^\xi$ are evaluated in order to compute past states:

$$v_{l+1}^j = z_{\eta,\eta}^\xi(t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c)). \tag{8.7}$$

Whenever the past time points given by the deviating argument are located within the "correct" discontinuity interval, the evaluation of the deduced function $z_{\eta,\eta}^\xi$ is equivalent to an evaluation of a smooth branch of the initial function (if $-n_s^\phi \leqslant \xi \leqslant 0$) or to an evaluation of the continuous representation in a previous or in the current integration step (if $1 \leqslant \xi \leqslant n_s + 1$). If the past time points are located outside of the "correct" discontinuity interval, the deduced function is given by extrapolating the smooth branch of the initial function or by extrapolating the continuous representation, see equations (5.63) and (5.60). Formally, one has

$$v_{l+1}^j = \begin{cases} \phi_\xi(t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c), c) & \text{if} \quad -n_s^\phi \leqslant \xi \leqslant 0 \\ y_{l'} + h_{l'+1}\sum_{i=1}^{\nu} b_i(\theta_{l,j}) g_{l'+1}^i & \text{if} \quad 1 \leqslant \xi \leqslant n_s + 1, \end{cases} \tag{8.8}$$

where ϕ_i are the smooth branches of the initial function, see equation (5.57). Further, in the case that $1 \leqslant \xi \leqslant n_s + 1$, $l'+1$ is the index of the integration step from which the continuous representation is used for the computation of the past state, and the symbol $\theta_{l,j}$ denotes the relative position of the past time points in the interval $[t_{l'}, t_{l'+1}]$:

$$\theta_{l,j} = \frac{t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) - t_{l'}}{h_{l'+1}}. \tag{8.9}$$

The index l' generally depends on l and j, but for compactness of the notation this dependency is not written. It is emphasized that the mesh is assumed to be chosen such that the points of discontinuity in the exact solution $y(t;c)$ are included in the mesh. Therefore, it may happen for the numerical solution that the deviating argument assumes values $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c)$ such that $\theta_{l,j} \notin [0,1]$.

8.1.2. Derivatives of IVP Solutions with Respect to Parameters

If the derivative of the nominal DDE-IVP solution $y(t;c)$ with respect to the parameters c exists, then it is denoted by $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$, and $\mathbf{W}(t;c)$ is called the Wronskian matrix. In practice, the goal is typically to compute, on the one hand, the solution $y(t;\tilde{c})$ for specific parameters $\tilde{c}$,

and, on the other hand, the Wronskian matrix along this specific solution, i.e.

$$\mathbf{W}(t;\tilde{c}) = \left.\frac{\partial y(t;c)}{\partial c}\right|_{c=\tilde{c}}. \tag{8.10}$$

However, for simplicity of notation, it is not formally distinguished in this chapter between the parameters c as a variable and the specific evaluation point $\tilde{c}$.

The theory on the differentiability of DDE-IVP solutions was presented in the Sections 7.2, 7.3, and 7.5 (recall, in particular, Theorems 7.5 and 7.11). In the following, it is assumed that the sufficient conditions for differentiability as presented in these sections are fulfilled. In this case, the Wronskian matrix is at the initial time given by

$$\begin{aligned}\mathbf{W}(t^{ini}(c);c) &= \frac{d}{dc}y^{ini}(c) - f^{ini}(c)\frac{d}{dc}t^{ini}(c) \\ f^{ini}(c) &:= f(t^{ini}(c), y^{ini}(c), c, \phi(t^{ini}(c) - \tau_1(t^{ini}(c), y^{ini}(c), c), c)).\end{aligned} \tag{8.11}$$

Further, for $t > t^{ini}(c)$ the Wronskian $\mathbf{W}(t;c)$ is given as solution of the following variational DDE-IVP (cf. equation (7.100)):

$$\begin{aligned}\dot{\mathfrak{w}}(t;c) =& \frac{\partial f}{\partial y}\mathfrak{w}(t;c) + \frac{\partial f}{\partial c} + \frac{\partial f}{\partial v}\Bigg(\dot{y}(t-\tau_1(t,y(t;c),c);c)\left[-\frac{\partial \tau_1}{\partial y}\mathfrak{w}(t;c) - \frac{\partial \tau_1}{\partial c}\right] \\ &+\mathfrak{w}(t-\tau_1(t,y(t;c),c);c)\Bigg)\end{aligned} \tag{8.12a}$$

$$\mathfrak{w}(t^{ini}(c);c) = \frac{dy^{ini}(c)}{dc} - f^{ini}(c)\frac{dt^{ini}(c)}{dc} \tag{8.12b}$$

$$\mathfrak{w}(t;c) = \frac{\partial \phi(t,c)}{\partial c} \quad \text{for} \quad t < t^{ini}(c). \tag{8.12c}$$

Herein, all partial derivatives of f are evaluated at $(t, y(t;c), c, y(t-\tau_1(t,y(t;c),c);c))$, and the partial derivatives of τ_1 are evaluated at $(t, y(t;c), c)$.

In addition, there may be jumps in the derivative $\mathbf{W}$. The jumps may occur at the time point of a propagated discontinuity if the parent discontinuity is of order 0 in y. The jump that needs to be applied is given by equation (7.99), which is recalled here:

$$\mathbf{W}^+(s_i;c) - \mathbf{W}^-(s_i;c) = (f_i^-(c) - f_i^+(c))\frac{ds_i}{dc}, \tag{8.13}$$

with

$$f_i^{\pm}(c) := f(s_i, y(s_i;c), c, y^{\bullet}(s_i - \tau_1(s_i, y(s_i;c), c);c)). \tag{8.14}$$

Thereby, $y^{\bullet}$ represents the left-sided or the right-sided limit at the time point of the parent discontinuity depending on the behavior of the deviating argument to the left and to the right of s_i. Furthermore, the term ds_i/dc in equation (8.13) is given by

$$\frac{ds_i}{dc} = \frac{\frac{\partial \tau_1}{\partial y}\mathbf{W}^-(s_i;c) + \frac{\partial \tau_1}{\partial c} + \frac{ds_j(c)}{dc}}{1 - \frac{\partial \tau_1}{\partial t} - \frac{\partial \tau_1}{\partial y}f_i^-(c)}. \tag{8.15}$$

Herein, s_j is the time point of the initial discontinuity of order 0 in y that is the parent of the discontinuity at s_i. The partial derivatives of τ_1 are evaluated at $(s_i, y(s_i;c), c)$.

There are different ways how to regard the IVP (8.12) for the Wronskian matrix. Either one considers it together with the nominal DDE-IVP (8.2), in which case the combined system occurs as a so-called "DDE-IVP of neutral type", i.e. a DDE-IVP in which the right-hand-side function depends on the time derivative of the state at past time points. Or one considers the variational IVP (8.12) disconnected from the nominal IVP, in which case the nominal DDE-IVP solution $y(t)$ and its time derivative $\dot{y}(t)$ have to be considered as external input functions that are simply assumed to be available.

8.2. Forward Sensitivity Computation

This section is concerned with the numerical computation of forward derivatives of DDE-IVP solutions with respect to parameters in the model functions. As a motivation for the work presented in this section, it is first recalled why the use of difference quotients is a very inefficient strategy for this purpose.

8.2.1. Difference Quotients, External Numerical Differentiation

Error Analysis for Derivative Approximation by Difference Quotients

Whenever an arbitrary function x, depending both on time t and on parameters c, is differentiable with respect to c (for fixed t), then an elementary approach for derivative computation is the use of a difference quotient. More precisely, one obtains from a Taylor expansion the relation

$$\frac{\partial x(t,c)}{\partial c}\Delta c = \frac{x(t, c+\epsilon_{fd}\Delta c) - x(t,c)}{\epsilon_{fd}} + \mathscr{O}(\epsilon_{fd}). \tag{8.16}$$

The first term on the right hand side is a difference quotient approximating the directional derivative $\partial x(t,c)/\partial c \cdot \Delta c$. The second term on the right hand side represents nonlinear dependencies of x on c; it is proportional to ϵ_{fd}.

In theory, the derivative approximation by a difference quotient becomes better and better if the *variational parameter* ϵ_{fd} is chosen smaller and smaller. Unfortunately, if the function x is given in the form of a computer program, then this is not true due to the use of floating point arithmetic. For computer programs, one has to take into account that the result of an evaluation of x are representable numbers $\hat{x}(t,c)$ and $\hat{x}(t, c+\epsilon_{fd}\Delta c)$, which are only approximations of $x(t,c)$ and $x(t, c+\epsilon_{fd}\Delta c)$, respectively. Hence, it holds that

$$\hat{x}(t,c) = \mathtt{fl}(x(t,c)) = x(t,c) + \epsilon_1 \tag{8.17a}$$

$$\hat{x}(t, c+\epsilon_{fd}\Delta c) = \mathtt{fl}(x(t, c+\epsilon_{fd}\Delta c)) = x(t, c+\epsilon_{fd}\Delta c) + \epsilon_2(\epsilon_{fd}), \tag{8.17b}$$

where $\mathtt{fl}$ is the operator that rounds its argument to a representable floating point number and ϵ_1 and $\epsilon_2(\epsilon_{fd})$ are the numerical errors that are made in the computer evaluation of the function x. The latter, ϵ_2, is considered as a function of the variational parameter ϵ_{fd}.

Note that, in fact, there are further sources of errors in equation (8.17) because also the input arguments t, c, and $c+\epsilon_{fd}\Delta c$ have to be rounded to representable numbers $\mathtt{fl}(t)$, $\mathtt{fl}(c)$, and $\mathtt{fl}(c+\epsilon_{fd}\Delta c)$. This issue is neglected here. However, the reader should keep in mind that this principally prevents to regard the asymptotic $\epsilon_{fd} \to 0$ whenever floating point arithmetic is used. Instead, the variation always has to be large enough such that a representable number different from $\mathtt{fl}(c)$ is used as input argument of $\hat{x}$ in equation (8.17b).

The error analysis of a practically computed finite difference approximation yields

$$\frac{\hat{x}(t, c+\epsilon_{fd}\Delta c) - \hat{x}(t,c)}{\epsilon_{fd}} = \frac{\partial x(t,c)}{\partial c}\Delta c + \mathscr{O}(\epsilon_{fd}) + \frac{\epsilon_2(\epsilon_{fd}) - \epsilon_1}{\epsilon_{fd}}. \tag{8.18}$$

The error of the derivative approximation thus has two contributions. On the one hand, there are nonlinear effects in the function x, represented by the second term in the right hand side, which become larger for increasing ϵ_{fd}. On the other hand, there are errors in the evaluation of x. For any arbitrary computer program, e.g. a program evaluating a polynomial function, it must always be expected that $\epsilon_2(\epsilon_{fd}) - \epsilon_1$ is of the size $\epsilon_{mach} \cdot |x(t,c)|$, with ϵ_{mach} being the machine precision. Since this error is divided by the variational parameter, it becomes larger with decreasing ϵ_{fd}. Hence, there typically is a value for ϵ_{fd} that balances the two errors optimally such that the error of the resulting derivative approximation is minimal. Unfortunately, this optimal value (and also the range of "good" values) of ϵ_{fd} is generally unknown.

Difference Quotients for Derivatives of IVP Solutions (External Numerical Differentiation)

Difference quotients may also be used if the function $x(t,c)$ represents the numerically computed solution of an initial value problem. If realized in such a way that the integrator is called twice for

the computation of $\hat{x}(t,c)$ and $\hat{x}(t,c+\epsilon_{fd}\Delta c)$, then the approach is also termed *External Numerical Differentiation*, because the differentiation takes place outside the integrator.

Importantly, it holds in the context of IVP solutions that ϵ_1 and $\epsilon_2(\epsilon_{fd})$ represent the numerical integration errors, which are typically much larger than the machine precision ϵ_{mach}. Moreover, the result obtained from an IVP solver is typically a discontinuous function of the parameters c, because the IVP solver involves discrete decisions, see e.g. the error control strategy of Colsol-DDE in step 11 of Algoritm 6.20, which is representative for IVP solvers. As a consequence, the difference $\epsilon_2(\epsilon_{fd})-\epsilon_1$ must generally be expected to be of the order of the chosen relative tolerance σ_{tol}^{rel} in case that a variable-stepsize solver is used.

As a rule of thumb, even for the optimal choice of ϵ_{fd} the directional derivative is approximated with only half as many valid digits as the nominal solution. Hence, in general, the computation of the sensitivities with four valid digits requires to compute IVP solutions with eight valid digits. Since such highly accurate IVP solutions are available only at very high computational costs, this motivates the search for more efficient methods for sensitivity computation.

Two approaches for a more efficient sensitivity computation are considered in the following. On the one hand, the differentiation of the discrete scheme that is used for the numerical solution of the nominal DDE-IVP ("first discretize, then differentiate"). On the other hand, the discretization of the variational DDE-IVP by the numerical method ("first differentiate, then discretize"). The convergence properties of both approaches are subsequently analyzed, which will lead to a generalization of the concept of Internal Numerical Differentiation.

The two approaches, "first discretize, then differentiate" and "first differentiate, then discretize", are here presented in the context of CRK methods. However, some important peculiarities of sensitivity computation for DDE-IVP solutions discussed in this section are also encountered for other one-step and multi-step methods.

8.2.2. First Discretize, Then Differentiate (Differentiation of CRK Scheme)

In the first approach, the CRK scheme equation (8.3), (8.8) for solving the nominal DDE-IVP (8.2), is differentiated with respect to the parameters. Clearly, all intermediate results of the CRK scheme are generally dependent on the parameters, even though this dependency was so far not explicitly given. For compactness of notation, define the following quantities:

$$\mathbf{W}_{l+1} := \frac{\partial y_{l+1}}{\partial c}, \quad \mathbf{W}_{l+1}^{j} := \frac{\partial y_{l+1}^{j}}{\partial c}, \quad \mathbf{G}_{l+1}^{j} := \frac{\partial g_{l+1}^{j}}{\partial c}. \tag{8.19}$$

Start with $t_0 := t^{ini}(c)$, $y_0 := y^{ini}(c)$, and with the following initialization for the Wronskian matrix:

$$\mathbf{W}_0 := \frac{dy_0}{dc} - f(t_0, y_0, c, \phi(t_0 - \tau_1(t_0, y_0, c)))\frac{dt_0}{dc}. \tag{8.20}$$

Then, in order to obtain the approximation of the Wronskian matrix for $t > t_0$, take the derivative of the CRK scheme (8.3), (8.8). Motivated by the findings for sensitivity computation for ODE-IVPs (see introduction of this chapter, and references given therein), the sequence of stepsizes is thereby kept fixed. This yields the following result:

$$\mathbf{W}_{l+1} = \mathbf{W}_l + h_{l+1}\sum_{j=1}^{\nu}\beta_j\mathbf{G}_{l+1}^{j} \tag{8.21a}$$

$$\mathbf{E}_{l+1}(t_l+\theta h_{l+1}) = \mathbf{W}_l + h_{l+1}\sum_{j=1}^{\nu} b_j(\theta)\mathbf{G}_{l+1}^{j} \tag{8.21b}$$

$$\mathbf{G}_{l+1}^{j} = \left(\frac{\partial f}{\partial y}\right)_{l+1}^{j}\mathbf{W}_{l+1}^{j} + \left(\frac{\partial f}{\partial c}\right)_{l+1}^{j} + \left(\frac{\partial f}{\partial v}\right)_{l+1}^{j}\frac{dv_{l+1}^{j}}{dc} \tag{8.21c}$$

$$\mathbf{W}_{l+1}^{j} = \mathbf{W}_l + h_{l+1}\sum_{i=1}^{\nu} a_{j,i}\mathbf{G}_{l+1}^{i}. \tag{8.21d}$$

Herein and throughout the chapter, the quantity $\mathbf{E}_{l+1}(t_l+\theta h_{l+1})$ denotes the continuous representation for the Wronskian matrix on the interval $[t_l, t_{l+1}]$. The symbols $(\partial f/\partial y)_{l+1}^{j}$, $(\partial f/\partial c)_{l+1}^{j}$,

and $(\partial f/\partial v)_{l+1}^{j}$ represent evaluations of the partial derivatives of f:

$$\left(\frac{\partial f}{\partial y}\right)_{l+1}^{j} := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(t_{l+1}^{j},y_{l+1}^{j},c,v_{l+1}^{j})} \tag{8.22a}$$

$$\left(\frac{\partial f}{\partial c}\right)_{l+1}^{j} := \left.\frac{\partial f(t,y,c,v)}{\partial c}\right|_{(t_{l+1}^{j},y_{l+1}^{j},c,v_{l+1}^{j})} \tag{8.22b}$$

$$\left(\frac{\partial f}{\partial v}\right)_{l+1}^{j} := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(t_{l+1}^{j},y_{l+1}^{j},c,v_{l+1}^{j})} . \tag{8.22c}$$

It remains to give an expression for dv_{l+1}^{j}/dc. A straightforward differentiation of equation (8.8) gives

$$\frac{dv_{l+1}^{j}}{dc} = \begin{cases} \left(\frac{\partial \phi_\xi}{\partial c}\right)_{l+1}^{j} - \left(\frac{d\phi_\xi}{dt}\right)_{l+1}^{j}\left[\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{j}\mathbf{W}_{l+1}^{j} + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{j}\right] & \text{if } -n_s^{\phi} \leqslant \xi \leqslant 0 \\ \mathbf{W}_{l'} + h_{l'+1}\sum_{i=1}^{\nu} b_i(\theta_{l,j})\mathbf{G}_{l'+1}^{i} + h_{l'+1}\sum_{i=1}^{\nu}\dot{b}_i(\theta_{l,j})g_{l'+1}^{i}\frac{d\theta_{l,j}}{dc} & \text{if } 1 \leqslant \xi \leqslant n_s+1. \end{cases} \tag{8.23}$$

Herein, the abbreviations

$$\frac{d\theta_{l,j}}{dc} = -\frac{\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{j}\mathbf{W}_{l+1}^{j} + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{j}}{h_{l'+1}} \tag{8.24}$$

and

$$\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{j} := \left.\frac{\partial \tau_1(t,y,c)}{\partial y}\right|_{(t_{l+1}^{j},y_{l+1}^{j},c)} \tag{8.25a}$$

$$\left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{j} := \left.\frac{\partial \tau_1(t,y,c)}{\partial c}\right|_{(t_{l+1}^{j},y_{l+1}^{j},c)} \tag{8.25b}$$

$$\left(\frac{d\phi_\xi}{dt}\right)_{l+1}^{j} := \left.\frac{d\phi_\xi(t,c)}{dt}\right|_{(t_{l+1}^{j}-\tau_1(t_{l+1}^{j},y_{l+1}^{j},c),c)} \tag{8.25c}$$

$$\left(\frac{\partial \phi_\xi}{\partial c}\right)_{l+1}^{j} := \left.\frac{\partial \phi_\xi(t,c)}{\partial c}\right|_{(t_{l+1}^{j}-\tau_1(t_{l+1}^{j},y_{l+1}^{j},c),c)} \tag{8.25d}$$

have been used.

Inserting equation (8.24) into (8.23) gives, for $1 \leqslant \xi \leqslant n_s+1$, the following expression:

$$\frac{dv_{l+1}^{j}}{dc} = \mathbf{W}_{l'} + h_{l'+1}\sum_{i=1}^{\nu} b_i(\theta_{l,j})\mathbf{G}_{l'+1}^{i} - \sum_{i=1}^{\nu}\dot{b}_i(\theta_{l,j})g_{l'+1}^{i}\left[\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{j}\mathbf{W}_{l+1}^{j} + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{j}\right]. \tag{8.26}$$

Clearly, this is equivalent to

$$\frac{dv_{l+1}^{j}}{dc} = \mathbf{E}_{l'+1}(t_{l'}+\theta_{l,j}h_{l'+1}) - \dot{\eta}_{l'+1}(t_{l'}+\theta_{l,j}h_{l'+1})\left[\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{j}\mathbf{W}_{l+1}^{j} + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{j}\right]. \tag{8.27}$$

It can then be seen that the equations (8.21), (8.23) are a CRK method applied to the variational DDE-IVP (8.12) in which the time derivative of the nominal solution, $\dot{y}(t_{l+1}^{j}-\tau_1(t_{l+1}^{j},y_{l+1}^{j},c);c)$, is approximated by the time derivative of the continuous representation, i.e. by the expression $\dot{\eta}_{l'+1}(t_{l+1}^{j}-\tau_1(t_{l+1}^{j},y_{l+1}^{j},c))$.

Without giving the formulas explicitly, it is remarked that the numerical computation of sensitivities for problems with multiple delays leads to additional terms in the right hand side of equation (8.21c); each of the occuring terms $d(v_m)_{l+1}^{j}/dc$ for each delay τ_m is thereby given by an expression of the form (8.23).

8.2.3. First Differentiate, Then Discretize (CRK Discretization of Variational IVP)

Consider now, as a second approach for sensitivity computation, a CRK method applied to the variational DDE-IVP (8.12). Clearly, this approach for computing sensitivities offers more freedom than the "first discretize, then differentiate" approach, because a different sequence of stepsizes may be used. Moreover, a CRK discretization for the numerical computation of the Wronskian $\mathbf{W}(t;c)$ implies no specific approximation of the nominal DDE-IVP solution $y(t;c)$ or its time derivative $\dot{y}(t;c)$; in fact, it is generally possible to take the exact solution and its time derivative, if it is available.

In the typical situation that the exact nominal DDE-IVP solution is not available, it is practical to use the same sequence of stepsizes and the same stage values y_{l+1}^j, v_{l+1}^j as for the nominal DDE-IVP solution in order to save computational costs. However, the question remains how to approximate $\dot{y}$ at past time points because this quantity is not needed for solving the nominal DDE-IVP.

The use of the time derivative of the continuous representation, $\dot{\eta}$, is only one possible option. Alternatively, it can be exploited that a solution of a DDE-IVP fulfills the differential equation (8.2a). Accordingly, evaluations of the right-hand-side function f at past time points can be used as approximations:

$$\dot{y}(t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c); c) \approx f(t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c), v_{l+1}^j, c, u_{l+1}^j). \tag{8.28}$$

Thereby, u_{l+1}^j is an approximation of the state at $\{t_{past}\}_{l+1}^j - \tau_1(\{t_{past}\}_{l+1}^j, v_{l+1}^j, c)$ with $\{t_{past}\}_{l+1}^j = t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c)$, i.e.:

$$u_{l+1}^j \approx y(\{t_{past}\}_{l+1}^j - \tau_1(\{t_{past}\}_{l+1}^j, v_{l+1}^j, c); c). \tag{8.29}$$

This means that u_{l+1}^j is an approximation of a state at a time point even further in the past. In practice, u_{l+1}^j is given by an evaluation of the initial function or by an evaluation of the continuous representation at that time, depending on the discontinuity interval indicator ξ_{past} that was used for the past integration step $t_{l'} \to t_{l'+1}$.

By using evaluations of the right-hand-side function f at past time points, the obtained CRK scheme is given by the equations (8.21) and

$$\frac{dv_{l+1}^j}{dc} = \begin{cases} \left(\frac{\partial \phi_\xi}{\partial c}\right)_{l+1}^j - \left(\frac{d\phi_\xi}{dt}\right)_{l+1}^j \left[\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^j\right] & \text{if} \quad -n_s^\phi \leqslant \xi \leqslant 0 \\ \mathbf{E}_{l'+1}(t_{l'} + \theta_{l,j} h_{l'+1}) - f(t_{l'} + \theta_{l,j} h_{l'+1}, v_{l+1}^j, c, u_{l+1}^j) & \\ \qquad \cdot \left[\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^j\right] & \text{if} \quad 1 \leqslant \xi \leqslant n_s + 1. \end{cases} \tag{8.30}$$

The sole difference of this equation to equation (8.27) is that the evaluation of $\dot{\eta}$ has been replaced by an evaluation of the right-hand-side function f at the past time point.

For DDE-IVPs with multiple delays, total derivatives $d(v_m)_{l+1}^j/dc$ need to be computed for each delay τ_m by a formula of this form depending on the value of the discontinuity interval indicator for this delay. The computed terms have to be taken into account in the equation (8.21c) of the CRK scheme.

It is remarked that there exists at least one more option for the approximation of the time derivative $\dot{y}(t)$ for $t \in [t_{l'}, t_{l'+1}]$: the use of an interpolation procedure on a set of evaluations of the right-hand-side function f. Some evaluations of f are always available, namely the stage values g_{l+1}^j. This approach is known from the numerical solution of delay differential equations of neutral type, see Bellen and Zennaro [26] and Bellen and Guglielmi [24] for details.

It is remarked that the freedom gained in the "first differentiate, then discretize" approach as compared to "first discretize, then differentiate" is not specific to the use of CRK methods. In fact, it holds also for arbitrary numerical methods that the latter approach is coupled to the differentiation of the continuous representation used by the method, whereas the former allows to use different approximations of $\dot{y}$.

8.2.4. Discontinuities of Order 0 in W

In both approaches discussed in the preceding subsections – "first discretize, then differentiate" vs. "first differentiate, then discretize" – it needs to be taken into account that the Wronskian matrix $\mathbf{W}(t;c)$ may be discontinuous at the time points s_i. More precisely, this is the case whenever the parent of the discontinuity in s_j is of order 0 in y.

Let the parent discontinuity at s_j, $j < i$, be of order 0 in y. Then the jump in the exact derivative $\mathbf{W}(t;c)$ at $t = s_i$ is given by equation (8.13) (see also the equations (8.14) and (8.15)).

In the context of the idealized variant of the modified standard approach, let l be the index of the mesh points that corresponds to the discontinuity point s_i. Further, let $\mathbf{W}_l^-$ be the numerical approximation obtained by applying a CRK scheme to the variational DDE-IVP (8.12) until the mesh point t_l. Then the right-sided limit of $\mathbf{W}$ at $t_l = s_i$, denoted by $\mathbf{W}_l^+$, can numerically be approximated by

$$\mathbf{W}_l^+ = \mathbf{W}_l^- + (f_l^-(c) - f_l^+(c)) \left(\frac{\frac{\partial \tau_1(t_l,y_l,c)}{\partial y}\mathbf{W}_l^- + \frac{\partial \tau_1(t_l,y_l,c)}{\partial c} + \frac{ds_j}{dc}}{1 - \frac{\partial \tau_1(t_l,y_l,c)}{\partial t} - \frac{\partial \tau_1(t_l,y_l,c)}{\partial y} f_l^-(c)} \right). \tag{8.31}$$

Therein, $f_l^+(c)$ and $f_l^-(c)$ are the right-sided limit and the left-sided limit of the right-hand-side function evaluation at t_l, respectively, which are given by

$$f_l^{\pm}(c) = f(t_l, y_l, c, z_{\eta,\eta}^{\xi^{\pm}}(t_l - \tau_1(t_l, y_l, c))). \tag{8.32}$$

The symbol $\xi^{\pm}$ denotes the discontinuity interval indicator $\xi_1^{\alpha}(t)$ for $t \in (t_l, t_{l+1})$ and for $t \in (t_{l-1}, t_l)$, respectively. It is noted that the jump in the Wronskian given by equation (8.31) is obtained from equation (8.13) by replacing $y(s_i) \to y_l$.

After the discontinuity point, the integration of the variational IVP is continued by using $\mathbf{W}_l^+$ as a starting value for the step $t_l \to t_{l+1}$.

8.2.5. Convergence of Derivative Approximations to the Exact Derivative

It was observed in the previous subsections that discretization and differentiation do, in general, not commute for the computation of derivatives of DDE-IVP solutions with respect to parameters, even if the same integration method and the same sequence of stepsizes is used. This is a fundamental difference to sensitivity computation in the context of IVPs in ODEs or in differential-algebraic equations. There, differentiation and discretization do commute, at least for Runge-Kutta methods, extrapolation methods, and for linear multi-step methods (see e.g. Bock [38], Bock [39], page 199, Körkel [164], page 112).

The fact that there are several approaches for the approximation of $\dot{y}$ in a discretization of the variational DDE-IVP leads immediately to the question which approximation should be used in practice.

A first hint to the answer of this question is obtained as follows. From results on the polynomial approximation of functions it is known that, if the continuous representation η is consistent of uniform local order q, then its j-th time derivative $d^j\eta(t)/dt^j$ is a continuous approximation of $d^jy(t)/dt^j$ that is consistent with uniform local order $q - j$, $1 \leqslant j \leqslant q - 1$. Hence, a reduction of the order of consistency is obtained by using time derivatives of the continuous representation.

Consider further the special case that an explicit or implicit Euler method is used with linear interpolation. Then v_{l+1}^j is obtained by

$$\begin{aligned} v_{l+1}^j &= \eta_{l'+1}(t_l + \theta_{l,j} h_{l'+1}) \\ &= y_{l'} + h_{l'+1}\theta_{l,j} g_{l'+1}^1. \end{aligned} \tag{8.33}$$

In this case, the time derivative $\dot{\eta}_{l'+1}(t)$ is a piecewise constant approximation of the time derivative $\dot{y}$ of the exact solution. This approximation is discontinuous at the mesh points t_l, hence the obtained method cannot be interpreted as continuous Runge-Kutta method.

Another important argument against the use of the time derivative of the continuous representation is obtained by analysing the convergence of the numerically computed sensitivities to the exact derivative of the exact IVP solution. Since the exact derivative is given piecewise by the solution of the variational DDE-IVP, with possible discontinuities of order 0 at the time points s_i,

Theorem 5.18 with the extension to IHDDE-IVPs (Section 5.3) can be used. More precisely, the following corollary is obtained, which is formulated for the general case of multiple delays.

Corollary 8.1 (Convergence of CRK Scheme for Sensitivity Computation)

Consider a CRK method of discrete local order p and uniform local order q applied to the variational DDE-IVP (8.12). Assume that the same sequence of stepsizes is used and that y_{l+1}^j and v_{l+1}^j are computed by the same CRK method applied to the nominal DDE-IVP. Let the time derivative of the past state in the variational DDE be approximated with uniform local order $\tilde{q}$.

Assume that the right-hand-side function of the variational DDE-IVP fulfills the conditions (S) and (B) of Theorem 5.18. Assume further that the conditions of Theorem 7.11 are fulfilled (which guarantee existence of the partial derivative $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$).

In addition, assume that there is a finite number of discontinuities up to order p in either y or $\mathbf{W}$, whose locations are denoted by $s_{-n_s^\phi}, \ldots, s_{n_s}$. Assume that the set of mesh points is chosen in such a way that the discontinuity interval indicator $\xi^\alpha(t) = (\xi_1^\alpha(t), \ldots, \xi_{n_\tau}^\alpha(t))^T$ for the above-given set of discontinuities, evaluated along the exact solution, is (componentwise) constant between two mesh points; in particular, this implies that all discontinuities up to order $p+1$ in either y or $\mathbf{W}$ are included in the mesh. Further, in all discontinuities s_j whose parent is of order 0 in y, the jump in the Wronskian is approximated by equation (8.31).

Then the obtained CRK scheme for sensitivity computation, realized in the framework of the idealized variant of the modified standard approach, converges with discrete global order and uniform global order $\tilde{r} = \min(p, q+1, \tilde{q}+1)$, i.e.

$$\max_{1 \leqslant l \leqslant n_m} \|\mathbf{W}(t_l) - \mathbf{W}_l\| = \mathscr{O}(h_{max}^{\tilde{r}}) \tag{8.34a}$$

$$\max_{t_0 \leqslant t \leqslant t_{n_m}} \|\mathbf{W}(t) - \mathbf{E}(t)\| = \mathscr{O}(h_{max}^{\tilde{r}}), \tag{8.34b}$$

where $h_{max} = \max_{1 \leqslant l \leqslant n_m} h_l$.

Proof

Follows directly from Theorem 5.18 applied to the variational DDE-IVP. ■

There are two main differences between the convergence result for the sensitivity computation compared to the convergence result for the nominal solution (Theorem 5.18).

The first difference is that potentially more discontinuity points have to be included in the mesh in order to guarantee sufficient smoothness of the right-hand-side functions of both the nominal and of the variational DDE-IVP.

The second difference is that the discrete and uniform global order for the computed sensitivities, $\tilde{r}$, is lower than the discrete and uniform global order $r = \min(p, q+1)$ for the nominal solution if $\tilde{q} < \min(p-1, q)$. In particular, if the time derivative of the continuous representation is used for approximating $\dot{y}$ (which has uniform local order $\tilde{q} = q-1$) then it follows that $\tilde{r} = \min(p, q+1, q) = q$. This is less than r if $q < p$.

Contrariwise, the evaluation of the right-hand-side function at a past time point (see equation (8.28)) provides an approximation of $\dot{y}$ that is of uniform local order q. Hence, it follows that $\tilde{r} = \min(p, q+1) = r$. This means that the CRK scheme for sensitivity computation converges with the same discrete and uniform global order as the CRK scheme for the nominal solution.

It is worthwile to remark that the above conclusions regarding the convergence order of the computed sensitivities extend to many other one-step and multi-step integration methods. The reason is that the time derivative of a polynomial continuous approximation always has a lower uniform order of consistency than the continuous approximation itself; hence, a reduction of the convergence order may occur.

8.2.6. Internal Numerical Differentiation

The traditional definition of Internal Numerical Differentiation is to differentiate the discretization scheme that was used for computing the nominal solution, thereby keeping the adaptive components of the integration method fixed. In the context of IVPs in ODE and in differential-algebraic equations, and for many numerical integration methods, this turned out to be a suitable definition in order to obtain the desired property that the method for sensitivity computation converges with the same (discrete) order as the method that was used for solving the nominal initial value problem (see e.g. Bock [39], page 199).

For ODEs and differential-algebraic equations with switches or impulses, a straightforward differentiation of the discretization scheme does not take into account that the location of some of the mesh points is enforced by the condition to include the time points of the root discontinuities into the mesh. This led to an adaptation of the principle of Internal Numerical Differentiation, which was introduced in Bock [39], see also Albersmeyer [2].

The observations of the previous subsection, in particular Corollary 8.1, suggest that sensitivity computation in the context of DDEs requires a further extension of the concept of Internal Numerical Differentiation. More precisely, the following definition of Internal Numerical Differentiation for DDEs is proposed:

Definition 8.2 (Internal Numerical Differentiation for DDEs)

An Internal Numerical Differentiation method for computing the sensitivities of DDE-IVP solutions *is obtained by differentiating the discretization scheme that was used for the nominal solution, thereby keeping the adaptive components of the integration scheme fixed. Furthermore, jumps in the Wronskian are taken into account and all approximations of quantities in the past have the same local order of consistency as the approximations of quantities in the past that are used in the nominal scheme.*

Remark 8.3 (Example for an Internal Numerical Differentiation Method)

According to this definition, the scheme (8.21), (8.30) is an Internal Numerical Differentiation method, but the scheme (8.21), (8.23) is not.

It holds for CRK methods applied to DDE-IVPs that sensitivities that are computed by following the above definition of Internal Numerical Differentiation converge to the exact sensitivities with the same order as the nominal solution obtained by the same CRK method. It is remarked, however, that Definition 8.2 is formulated sufficiently general such that the same favorable convergence result should be obtained for other continuous integration methods, for discrete methods applied to DDE-IVPs on so-called constrained meshes, and for the computation of higher order derivatives $d^j y(t;c)/dc^j$ for $j \geqslant 2$ (which involves j-th order time derivatives of y).

8.2.7. Local Error

In the following, the *discrete and uniform local errors* of CRK methods for sensitivity computation are studied, i.e. the errors that are newly introduced in the integration step $t_l \to t_{l+1}$. The results of this subsection are used in Subsection 8.2.8 below for the construction of a numerical method that provides error-controlled sensitivity approximations.

For the numerical solution of the nominal DDE-IVP, the local error was formally defined as the difference between the numerical solution and the "exact solution of the local problem", see Definition 5.20. Therefore, in order to investigate the discrete and uniform local errors of CRK methods for sensitivity computation, it is necessary to define what the equivalent of the local problem (5.58) is in the case of sensitivity computation.

As a prepatory step for this purpose, a new notation is introduced for the right-hand-side function of the variational DDE (8.12a). Thereby, the general case of multiple delays is considered:

$$\begin{aligned}\mathbf{F}(t,\mathbf{W},c,\{\mathbf{V}_i\}_{i=1}^{n_\tau}) :=& \frac{\partial f}{\partial y}\mathbf{W} + \frac{\partial f}{\partial c} \\ &+ \sum_{i=1}^{n_\tau} \frac{\partial f}{\partial v_i}\left(\dot{y}(t-\tau_i(t,y(t;c),c);c)\left[-\frac{\partial \tau_i}{\partial y}\mathbf{W} - \frac{\partial \tau_i}{\partial c}\right] + \mathbf{V}_i\right). \end{aligned} \tag{8.35}$$

As usual, all partial derivatives of f are evaluated at $(t, y(t;c), c, \{y(t-\tau_i(t,y(t;c),c);c)\}_{i=1}^{n_\tau})$ and all partial derivatives of τ_i are evaluated at $(t, y(t;c), c)$. The above introduced notation is motivated by the wish to express the right-hand-side function of the variational DDE-IVP in a way that is similar to the right-hand-side function f of the nominal DDE-IVP. In particular, the second argument of $\mathbf{F}$ represents the Wronskian at the current time and the fourth argument represents the Wronskians at the past time points.

Assume that the variational DDE-IVP has been solved until the mesh point t_l, such that discrete approximations $\mathbf{W}_{l'}$ of $\mathbf{W}(t_{l'})$ and continuous approximations $\mathbf{E}_{l'}(t_{l'-1}+\theta h_{l'})$ of $\mathbf{W}(t_{l'-1}+\theta h_{l'})$

are available for $l' \leqslant l$. Then, in analogy to the local IVP (5.58), the idealized variant of the modified standard approach consists in solving the following local variational IVP:

$$\dot{\mathfrak{U}}_{l+1}(t) = \bar{\mathbf{F}}^{d}_{\mathbf{E},\mathbf{U}_{l+1}}(t, \mathfrak{U}_{l+1}(t), c, \xi^{\alpha}[l]) \tag{8.36a}$$

$$\mathfrak{U}_{l+1}(t_l) = \mathbf{W}_l. \tag{8.36b}$$

with

$$\bar{\mathbf{F}}^{d}_{\mathbf{E},\mathbf{U}_{l+1}}(t, \mathbf{W}, c, \xi^{\alpha}) := \mathbf{F}(t, \mathbf{W}, c, \{\mathbf{Z}^{\xi_i^{\alpha}}_{\mathbf{E},\mathbf{U}_{l+1}}(t - \tau_i(t, y, c))\}_{i=1}^{n_\tau}). \tag{8.37}$$

The past Wronskians in this local IVP are - as usual in the modified standard approach - obtained by evaluating deduced functions. Which of the deduced functions is evaluated depends on the discontinuity interval indicator $\xi^{\alpha}[l]$ in the current integration step, i.e. $\xi^{\alpha}[l] = \xi^{\alpha}(t')$ for an arbitrary $t' \in [t_l, t_{l+1}]$. It thereby holds that j' is the index such that $[t_l, t_{l+1}] \subset [s_{j'}, s_{j'+1}]$.

The deduced functions $\mathbf{Z}^{j}_{\mathbf{E},\mathbf{U}_{l+1}}(t)$ for $-n_s^{\phi} \leqslant j \leqslant j'+1$ are thereby given in complete analogy to equation (5.60). This means that $\mathbf{Z}^{j}_{\mathbf{E},\mathbf{U}_{l+1}}(t)$ for $-n_s^{\phi} \leqslant j \leqslant 0$ denotes a smooth branch of $\partial\phi_j/\partial c(t;c)$ and its smooth extrapolation. Further, $\mathbf{Z}^{j}_{\mathbf{E},\mathbf{U}_{l+1}}(t)$ for $1 \leqslant j \leqslant j'$ is equal to the continuous representation $\mathbf{E}(t)$ if the argument t is in the "correct" discontinuity interval, i.e. if $t \in [s_{j-1}, s_j]$. Otherwise, the continuous representation in the first (last) integration step within that discontinuity interval is extrapolated to the left (right). Finally, $\mathbf{Z}^{j'+1}_{\mathbf{E},\mathbf{U}_{l+1}}(t)$ is equal to $\mathbf{E}(t)$ for $t \in [s_{j'}, t_l]$, extrapolation of the continuous representation is used to the left of $s_{j'}$, and the exact solution $\mathbf{U}_{l+1}(t)$ of problem (8.36) is used if $t > t_l$.

In practical situations, the exact solution $\mathbf{U}_{l+1}(t)$ is unavailable. Therefore, a continuous one-step method has to make use of the continuous representation $\mathbf{E}_{l+1}(t)$ that is implied by the method itself. In particular, this has been the case for the CRK methods for sensitivity computation presented in the Subsections 8.2.2 and 8.2.3. In terms of the discrete and continuous increment functions $\mathbf{\Phi}$ and $\mathbf{\Psi}$ of the CRK method, $\mathbf{W}_{l+1}$ and $\mathbf{E}_{l+1}(t_l + \theta h_{l+1})$ can formally be written as

$$\mathbf{W}_{l+1} = \mathbf{W}_l + h_{l+1}\mathbf{\Phi}(t_l, \mathbf{W}_l, h_{l+1}; \bar{\mathbf{F}}^{d}_{\mathbf{E},\mathbf{E}^1_{l+1}}(\cdot,\cdot,\cdot,\xi^{\alpha})) \tag{8.38a}$$

$$\mathbf{E}_{l+1}(t_l + \theta h_{l+1}) = \mathbf{W}_l + h_{l+1}\mathbf{\Psi}(t_l, \mathbf{W}_l, h_{l+1}, \theta; \bar{\mathbf{F}}^{d}_{\mathbf{E},\mathbf{E}_{l+1}}(\cdot,\cdot,\cdot,\xi^{\alpha})). \tag{8.38b}$$

Herein, $\bar{\mathbf{F}}^{d}_{\mathbf{E},\mathbf{E}_{l+1}}$ is a right-hand-side function that has formally the shape of an ODE:

$$\bar{\mathbf{F}}^{d}_{\mathbf{E},\mathbf{E}_{l+1}}(t, \mathbf{W}, c, \xi) := \mathbf{F}(t, \mathbf{W}, c, \{\mathbf{Z}^{\xi_i^{\alpha}}_{\mathbf{E},\mathbf{E}_{l+1}}(t - \tau_i(t, y, c))\}_{i=1}^{n_\tau}), \tag{8.39}$$

The deduced functions $\mathbf{Z}^{j}_{\mathbf{E},\mathbf{E}_{l+1}}$ are thereby given in analogy to equation (5.63). This means that the evaluation of the deduced functions $\mathbf{Z}^{j}_{\mathbf{E},\mathbf{E}_{l+1}}$ yields the same result as an evaluation of $\mathbf{Z}^{j}_{\mathbf{E},\mathbf{U}_{l+1}}$ except if $j = j'+1$ and if the evaluation time is to the right of t_l. In the exceptional case, the continuous representation implied by the method itself is used instead of the exact solution $\mathbf{U}_{l+1}$ of the local variational IVP (8.36a).

With these preparations, a corollary of Theorem 5.21 for sensitivity approximations can be formulated.

Corollary 8.4 (Local Errors of CRK Methods for Sensitivity Computation with the Modified Standard Approach)

Consider the two local IVPs (5.58) and (8.36), and denote their exact solutions by $u_{l+1}(t)$ and $\mathbf{U}_{l+1}(t)$, respectively. Consider further a CRK method with discrete local order p and uniform local order q, which is applied to numerically solve these local problem with the modified standard approach. The results are denoted by y_{l+1} and $\eta_{l+1}(t_l + \theta h_{l+1})$ for the local nominal IVP (see equation (5.108)) and as $\mathbf{W}_{l+1}$ and $\mathbf{E}_{l+1}(t_l + \theta h_{l+1})$ for the local variational IVP (see equation (8.38)).

Assume that the conditions of Corollary 8.1 hold and further that the state $y(t;c)$ and the Wronskian $\mathbf{W}(t;c)$ at times to the left of t_l are approximated by an expression $\mathbf{E}(t)$ that has uniform global order r_1:

$$\max_{t \leqslant t_l} \|\eta(t) - y(t)\| = \mathscr{O}(h^{r_1}) \tag{8.40a}$$

$$\max_{t \leqslant t_l} \|\mathbf{E}(t) - \mathbf{W}(t)\| = \mathscr{O}(h^{r_1}), \tag{8.40b}$$

with $h = \max_{1 \leqslant i \leqslant l} h_i$*. Further, assume that the time derivative of the state, i.e.* $\dot{y}(t;c)$*, is approximated with uniform global order* r_2*.*

Then it holds that

$$\|\mathbf{W}_{l+1} - \mathbf{U}_{l+1}(t_{l+1})\| = \mathscr{O}(h_{l+1}^{p'+1}) \tag{8.41a}$$

$$\max_{t_l \leqslant t \leqslant t_{l+1}} \|\mathbf{E}_{l+1}(t) - \mathbf{U}_{l+1}(t)\| = \mathscr{O}(h_{l+1}^{q'+1}) \tag{8.41b}$$

where $p' = \min(p, r_1, r_2)$*,* $q' = \min(q, r_1, r_2)$*.*

Proof
Follows directly from Theorem 5.21 applied to the CRK scheme for sensitivity computation. ■

8.2.8. Error Control

Computing the approximations $\eta(t)$ of the nominal solution and $\mathbf{E}(t)$ of the derivatives on the same discretization mesh offers the opportunity to select the stepsizes in such a way that the error in y_{l+1} and $\eta_{l+1}(t_l + \theta h_{l+1})$, $\theta \in [0,1]$, and in $\mathbf{W}_{l+1}$ and $\mathbf{E}_{l+1}(t_l + \theta h_{l+1})$ is controlled.

Error estimation and control for the computation of sensitivities can in principle be done by the same techniques as those that are used for error estimation and control for the nominal solution (recall Section 5.5). This means to employ two methods whose local errors have different orders $p_1' \neq p_2'$ and $q_1' \neq q_2'$. For the practical design of error estimators it is thereby convenient to ensure that the orders of the local errors are equal to the discrete and uniform local orders of the methods, i.e. $p_i' = p_i$, $q_i' = q_i$.

For the nominal solution, it was already shown that these conditions may not hold because the orders of the local errors may be compromised by the order of the error in the computed past states. For the computation of sensitivities it turns out, see Corollary 8.4, that both the approximation order of the past Wronskian and the approximation order of the past time derivative of the nominal solution have to be taken into account.

8.2.9. Generalization: Sensitivities of IHDDE-IVP Solutions

The generalization of the main results of the previous subsections, Corollaries 8.1 and 8.4, to IHDDE-IVPs is obtained by making the following modifications. For the application of the idealized variant of the modified standard approach, it is required (in addition to previously-made assumptions) that the switching function signs $\zeta(t)$ along the exact solution are known, and that the mesh is chosen in such a way that $\zeta(t)$ is constant between two mesh points.

In the time points of the root discontinuities, the jump in the Wronskian matrix is taken into account. This jump is given by equation (7.108), with the total derivative of the discontinuity time point given by equation (7.105). In the framework of the idealized variant of the modified standard approach, these expressions are approximated numerically by replacing $y^{\pm}(s_k;c) \to y_l^{\pm}$, $\mathbf{W}^{\pm}(s_k;c) \to \mathbf{W}_l^{\pm}$. Further, some method of approximating the time derivative of the state $\dot{y}$ and the Wronskian $\mathbf{W}$ at past time points is needed that is based on the evaluation of deduced functions.

8.2.10. Practical Variant of the Modified Standard Approach

The idealized variant of the modified standard approach is based on a number of irrealistic assumptions (see Definition 5.16). For example, it assumes that the points of discontinuity up to order p in the exact solution $y(t;c)$ are known. It further assumes that the discontinuity interval indicator $\xi^{\alpha}(t)$ is a piecewise constant function with (componentwise) finitely many discontinuities, and that all its points of discontinuity are included into the mesh.

The practical variant for solving DDE-IVPs was formally defined and discussed in Section 5.4 (cf., in particular, Definition 5.22). The key aspects of the practical variant were as follows:

- The selection of the mesh is done such that the discontinuity interval indicator for the numerical solution, denoted by $\hat{\xi}^{\alpha}(t)$, is constant between two mesh points.

- There is a unique consistent choice for the discontinuity interval indicator.

Under certain regularity conditions on the behavior of the switching functions, it was possible to transfer the convergence result of the idealized variant to the practical variant.

For the purpose of sensitivity computation, the practical variant of the modified standard approach has to be modified. Motivated by Corollary 8.1, it is necessary to include all those discontinuity points into the mesh for which the "practically determined discontinuity order" in either y or $\mathbf{W}$ is less than or equal to $p+1$. This requires, of course, to determine the order of the discontinuity in the Wronskian. This can be done as follows: Assume that all initial discontinuities are of order 0 in $\mathbf{W}$. Then, if o_j^y and o_j^W denote the order of the discontinuity in the nominal solution and in the derivatives at the time point s_j, respectively, the order of discontinuity in $\mathbf{W}$ at s_i is set to $o_i^W = \min(o_j^W, o_j^y - 1) + 1$. For the determination of the order of discontinuity in the nominal solution, denoted by o_j^y, it is referred to the answer to question Q1 in Section 5.4.

By using the practical variant instead of the idealized variant, only the underlying assumption on the choice of the mesh changes. Therefore, all equations of the Subsections 8.2.2 and 8.2.3 remain formally valid. The same holds for the computation of the jumps in the Wronskian, which can be computed by equation (8.31) whenever a mesh points t_l is identified as an approximation of the location of a propagated discontinuity. An according statement holds in the time points of root discontinuities in IHDDE-IVP solutions (cf. Subsection 8.2.9).

If certain regularity assumptions on the behavior of the propagation switching functions are fulfilled for the exact nominal DDE-IVP solution $y(t;c)$, then the convergence result of Corollary 8.1 can be transferred to the practical variant of the modified standard approach. The arguments are thereby the same as in the answers to Question Q2 and Q3 in Section 5.4.

Further, the differentiability result given in Theorem 7.11 suggests to make the following safeguard checks: Children of discontinuities of order 0 in y should not coincide, and the propagation switching functions that correspond to critical discontinuities should have a non-zero time derivative in their zeros (both to the left and to the right of the time point of the child discontinuity). In addition, if IHDDE-IVPs are solved numerically, the findings of Chapter 7 suggest to exclude that a zero of a switching function coincided with the zero of a different switching function or with the zero of a propagation switching functions that corresponds to a critical discontinuity.

The safeguard checks that are practically used in Colsol-DDE are presented in Subsection 9.1.11.

8.3. Adjoint Sensitivity Computation

8.3.1. Motivation

The derivative $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$ of an IVP solution with respect to the parameters is a matrix-valued function. It contains in column j the derivatives of all components of y with respect to the j-th parameter, and in row i the derivative of the state vector component y_i with respect to all parameters.

Consider the situation that only the i-th row of the Wronskian matrix is of practical interest for a specific problem. Since the forward approach for sensitivity computation computes the sensitivities columnwise, it becomes very inefficient if there are many parameters. In this case, it is well-known that adjoint methods for sensitivity computation are much more efficient, because the sensitivities are computed rowwise and the computational effort is indepedent of the number of parameters, see e.g. Bock [39], page 210, and Bock, Schlöder, and Schulz [45].

In a slightly more general setting, consider the issue of computing the derivative of a scalar function $\Omega(t^{fin}(c), y(t^{fin}(c);c), c)$, where $y(t;c)$ is the solution of a DDE-IVP. Its total derivative with respect to c can be approximated numerically by

$$\begin{aligned}\frac{d\Omega(t_{n_m}, y_{n_m}, c)}{dc} =& \frac{\partial \Omega}{\partial t}(t_{n_m}, y_{n_m}, c)\frac{dt_{n_m}}{dc} + \frac{\partial \Omega}{\partial c}(t_{n_m}, y_{n_m}, c) \\ &+ \frac{\partial \Omega}{\partial y}(t_{n_m}, y_{n_m}, c)\left[\frac{\partial y_{n_m}}{\partial c} + f^{fin}(c)\frac{dt_{n_m}}{dc}\right]. \end{aligned} \tag{8.42}$$

Herein, $dt_{n_m}/dc = dt^{fin}(c)/dc$ and

$$f^{fin}(c) := f(t_{n_m}, y_{n_m}, c, v_{n_m}), \tag{8.43}$$

and v_{n_m} is an approximation of $y(t_{n_m} - \tau_1(t_{n_m}, y_{n_m}, c); c)$, which is obtained by an evaluation of a deduced function, as usual in the framework of the modified standard approach.

If an approximation to the nominal DDE-IVP solution at the final time has been computed, then the first, the second, and the fourth term in equation (8.42) can directly be computed. However, the third term requires a numerical approximation of the derivative of the IVP solution with respect to parameters, $\mathbf{W}_{n_m} = \partial y_{n_m}/\partial c$. The basic idea of the adjoint approach for sensitivity computation is to set

$$\Lambda_{n_m} := \frac{\partial \Omega}{\partial y}(t_{n_m}, y_{n_m}, c) \in \mathbb{R}^{n_y} \tag{8.44}$$

and to compute the product $\Lambda_{n_m}\mathbf{W}_{n_m}$ without computing the two factors individually. In particular, this avoids the computation of the full Wronskian $\mathbf{W}_{n_m}$.

8.3.2. Sensitivity Computation by Discrete Adjoints

In the following, a discrete adjoint scheme of the CRK method (8.3), (8.8) is derived. More precisely, the goal is to derive the discrete adjoint scheme that fits exactly to the Internal Numerical Differentiation scheme for sensitivity computation given by equations (8.21), (8.30). For notational simplicity the case of a DDE-IVP with a single delay τ_1 is considered. The generalization to the multiple delay case is straightforward.

It is assumed here that there are no discontinuities of order 0 in the Wronskian. This means that for all mesh points t_l that are identified as propagated discontinuities the following equation holds: $\mathbf{W}_l^+ = \mathbf{W}_l^-$. The treatment of problems with discontinuities of order 0 in the Wronskian is discussed later in Subsection 8.3.4.

An idea of Bock [39] is used as basis for the development of the new discrete adjoint scheme that fits exactly to the scheme (8.21), (8.30): Take the equations (8.21), (8.30) for the computation of forward derivatives, multiply them by newly introduced prefactors (the *adjoint variables*), and reorder the obtained expressions such that the quantities $\mathbf{W}_l$, $\mathbf{G}_{l+1}^j$, $\mathbf{W}_{l+1}^j$, and dv_{l+1}^j/dc need not be computed because they are multiplied by zeros. More precisely, the following variational ansatz is made that employs the adjoint variables Λ_{l+1}, Λ_l^j, Γ_l^j, and Π_l^j, all of which are row vectors of dimension n_y (the dimension of the state vector).

$$\begin{aligned}
0 = \sum_{l=0}^{n_m-1} \Bigg\{ & \Lambda_{l+1}\left[-\mathbf{W}_{l+1} + \mathbf{W}_l + h_{l+1}\sum_{j=1}^{\nu}\beta_j\mathbf{G}_{l+1}^j\right] \\
& + h_{l+1}\left(\sum_{j=1}^{\nu}\Lambda_l^j\left[-\mathbf{G}_{l+1}^j + \left(\frac{\partial f}{\partial y}\right)_{l+1}^j\mathbf{W}_{l+1}^j + \left(\frac{\partial f}{\partial c}\right)_{l+1}^j + \left(\frac{\partial f}{\partial v}\right)_{l+1}^j\frac{dv_{l+1}^j}{dc}\right]\right) \\
& + h_{l+1}\left(\sum_{j=1}^{\nu}\Gamma_l^j\left[\mathbf{W}_{l+1}^j - \mathbf{W}_l - h_{l+1}\sum_{i=1}^{\nu}a_{j,i}\mathbf{G}_{l+1}^i\right]\right) \\
& + h_{l+1}\Bigg(\sum_{j=1}^{\nu}\Pi_l^j H_{l+1}\Bigg[-\frac{dv_{l+1}^j}{dc} + \mathbf{W}_{l'} + h_{l'+1}\sum_{i=1}^{\nu}b_i(\theta_{l,j})\mathbf{G}_{l'+1}^i \\
& \qquad -(f_{past})_{l+1}^j\left(\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j\mathbf{W}_{l+1}^j + \left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j\right)\Bigg]\Bigg) \\
& + h_{l+1}\Bigg(\sum_{j=1}^{\nu}\Pi_l^j(1-H_{l+1})\Bigg[-\frac{dv_{l+1}^j}{dc} + \left(\frac{\partial\phi_{\xi[l+1]}}{\partial c}\right)_{l+1}^j \\
& \qquad -\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\left(\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j\mathbf{W}_{l+1}^j + \left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j\right)\Bigg]\Bigg)\Bigg\}.
\end{aligned} \tag{8.45}$$

Herein, $(f_{past})_{l+1}^j = f(t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c), v_{l+1}^j, c, u_{l+1}^j)$, where u_{l+1}^j is, as discussed in the context of equation (8.29), an approximation of the state at a time point even further in the past. Further, H_{l+1} is an integer that characterizes whether or not the past states v_{l+1}^j were, in the integration step $t_l \to t_{l+1}$, computed from a smooth branch of the initial function. More precisely, if $\xi[l+1]$ denotes the value of the discontinuity interval indicator for the sole delay τ_1 in the integration step

$t_l \to t_{l+1}$, then it holds that $H_{l+1} = 0$ if $\xi[l+1] \leqslant 0$, and $H_{l+1} = 1$ if $\xi[l+1] > 0$.

If $\xi[l+1] > 0$, then the past states in step $t_l \to t_{l+1}$ are computed from the continuous representation in the integration step $t_{l'} \to t_{l'+1}$ with $l' \leqslant l$. It is necessary to recall in this context that the index l' depends on both l and j. This fact becomes important in the following.

It is remarked that the terms in the square brackets in equation (8.45) correspond exactly to the equations of the CRK scheme for forward sensitivity computation, see equations (8.21) and (8.30). It is thus obvious that the term on the right hand side is indeed zero such that the equation holds.

In order to avoid the computation of any of the quantities $\mathbf{W}_l$, $\mathbf{W}_{l+1}^j$, $\mathbf{G}_{l+1}^j$, and dv_{l+1}^j/dc, all terms in equation (8.45) are sorted with respect to these quantities. This yields the following expression:

$$
\begin{aligned}
0 = \sum_{l=0}^{n_m-1} \Bigg\{ & -\Lambda_{l+1}\mathbf{W}_{l+1} + \left[\Lambda_{l+1} - h_{l+1}\sum_{j=1}^{\nu}\Gamma_l^j\right]\mathbf{W}_l + h_{l+1}\sum_{j=1}^{\nu}\left[\Lambda_{l+1}\beta_j - \Lambda_l^j - h_{l+1}\sum_{i=1}^{\nu}a_{i,j}\Gamma_l^i\right]\mathbf{G}_{l+1}^j \\
& + h_{l+1}\sum_{j=1}^{\nu}\Bigg[\Lambda_l^j\left(\frac{\partial f}{\partial y}\right)_{l+1}^j + \Gamma_l^j \\
& \qquad -\Pi_l^j\left(H_{l+1}(f_{past})_{l+1}^j + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\right)\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^j\Bigg]\mathbf{W}_{l+1}^j \\
& + h_{l+1}\sum_{j=1}^{\nu}\left[\Lambda_l^j\left(\frac{\partial f}{\partial v}\right)_{l+1}^j - \Pi_l^j\right]\frac{dv_{l+1}^j}{dc} \\
& + h_{l+1}\sum_{j=1}^{\nu}\Lambda_l^j\left(\frac{\partial f}{\partial c}\right)_{l+1}^j - h_{l+1}\sum_{j=1}^{\nu}\Pi_l^j\left(H_{l+1}(f_{past})_{l+1}^j + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\right)\left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^j \\
& + h_{l+1}\sum_{j=1}^{\nu}\Pi_l^j(1-H_{l+1})\left(\frac{\partial \phi_{\xi[l+1]}}{\partial c}\right)_{l+1}^j \\
& + h_{l+1}\sum_{j=1}^{\nu}\Pi_l^j H_{l+1}\left[\mathbf{W}_{l'} + h_{l'+1}\sum_{i=1}^{\nu}b_i(\theta_{l,j})\mathbf{G}_{l'+1}^i\right]\Bigg\}. \qquad (8.46)
\end{aligned}
$$

At this point, the goal to factor out all quantities of the CRK scheme for forward sensitivity computation is not yet reached because the expression in the last line contains terms $\mathbf{W}_{l'}$ and $\mathbf{G}_{l'+1}^j$. These terms can locally, in the step $t_l \to t_{l+1}$, only be taken into account if overlapping occurs, i.e. if $l' = l$.

However, since the summation goes over all step indices l, it is always possible to move these terms to other integration steps. In order to do this, recall that the index l' actually depends on $l+1$ and j, i.e. $l'(l+1, j)$ is a more appropriate notation. In order to rearrange the terms in the last row of the equation (8.46), let $\mathcal{M}_l$ for $l \geqslant 0$ be a set of all those pairs of indices (μ, ρ) such that $l'(\mu+1, \rho) = l$. Further, let $\mathcal{M}_l$, $-n_s^\phi - 1 \leqslant l \leqslant -1$ be the sets containing those indices μ for which $\xi[\mu+1] = l+1$.

With this notation the following equation is obtained:

$$
\begin{aligned}
0 = \sum_{l=0}^{n_m-1} \Bigg\{ & -\Lambda_{l+1}\mathbf{W}_{l+1} + \left[\Lambda_{l+1} - h_{l+1}\sum_{j=1}^{\nu}\Gamma_l^j + \sum_{(\mu,\rho)\in\mathcal{M}_l} h_{\mu+1}\Pi_\mu^\rho\right]\mathbf{W}_l \\
& + h_{l+1}\sum_{j=1}^{\nu}\left[\Lambda_{l+1}\beta_j - \Lambda_l^j - h_{l+1}\sum_{i=1}^{\nu} a_{i,j}\Gamma_l^i + \sum_{(\mu,\rho)\in\mathcal{M}_l} h_{\mu+1}\Pi_\mu^\rho b_j(\theta_{\mu,\rho})\right]\mathbf{G}_{l+1}^j \\
& + h_{l+1}\sum_{j=1}^{\nu}\Bigg[\Lambda_l^j\left(\frac{\partial f}{\partial y}\right)_{l+1}^j + \Gamma_l^j - \Pi_l^j\Bigg(H_{l+1}(f_{past})_{l+1}^j \\
& \qquad + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\Bigg)\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j\Bigg]\mathbf{W}_{l+1}^j \\
& + h_{l+1}\sum_{j=1}^{\nu}\left[\Lambda_l^j\left(\frac{\partial f}{\partial v}\right)_{l+1}^j - \Pi_l^j\right]\frac{dv_{l+1}^j}{dc} \\
& + h_{l+1}\sum_{j=1}^{\nu}\Lambda_l^j\left(\frac{\partial f}{\partial c}\right)_{l+1}^j - h_{l+1}\sum_{j=1}^{\nu}\Pi_l^j\Bigg(H_{l+1}(f_{past})_{l+1}^j \\
& \qquad + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\Bigg)\left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j\Bigg\} \\
& + \sum_{l=-n_s-1}^{-1}\sum_{(\mu,\rho)\in\mathcal{M}_l} h_{\mu+1}\Pi_\mu^\rho\left(\frac{\partial\phi_l}{\partial c}\right)_{\mu+1}^\rho .
\end{aligned}
\tag{8.47}
$$

This equation motivates the following definition of a discrete adjoint of the CRK scheme (8.3), (8.8).

Definition 8.5 (Discrete Adjoint Scheme of a CRK Scheme Applied to DDE-IVPs)

The equations

$$\Lambda_l = \Lambda_{l+1} - h_{l+1}\sum_{j=1}^{\nu}\Gamma_l^j + \sum_{(\mu,\rho)\in\mathcal{M}_l} h_{\mu+1}\Pi_\mu^\rho \tag{8.48a}$$

$$\Lambda_l^j = \Lambda_{l+1}\beta_j - h_{l+1}\sum_{i=1}^{\nu} a_{i,j}\Gamma_l^i + \sum_{(\mu,\rho)\in\mathcal{M}_l} h_{\mu+1}\Pi_\mu^\rho b_j(\theta_{\mu,\rho}) \tag{8.48b}$$

$$\Gamma_l^j = -\Lambda_l^j\left(\frac{\partial f}{\partial y}\right)_{l+1}^j + \Pi_l^j\left(H_{l+1}(f_{past})_{l+1}^j + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\right)\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j \tag{8.48c}$$

$$\Pi_l^j = \Lambda_l^j\left(\frac{\partial f}{\partial v}\right)_{l+1}^j . \tag{8.48d}$$

constitute a discrete adjoint scheme of the CRK scheme (8.3), (8.8).

Remark 8.6 (Adjoint Internal Numerical Differentiation)

It is clear, from the way in which Definition 8.5 has been motivated, that the discrete adjoint scheme (8.48) "fits exactly" to the forward method for sensitivity computation given by the CRK scheme (8.21), (8.30). Thus, in view of Remark 8.3, it constitutes an *adjoint Internal Numerical Differentiation method*.

By solving the above-defined discrete adjoint scheme backward for $l = n_m - 1, \ldots, 0$, the terms in the square brackets in the second, third, and fourth line of equation (8.47) vanish for all l according to equations (8.48b), (8.48c), and (8.48d). Further, by using equation (8.48a), only two terms remain after summation over all l of the terms in the first line of equation (8.47). The two remaining terms are $-\Lambda_{n_m}\mathbf{W}_{n_m}$ and $\Lambda_0\mathbf{W}_0$.

Hence, equation (8.47) simplifies to

$$\begin{aligned}\Lambda_{n_m}\mathbf{W}_{n_m} = &+ \Lambda_0\mathbf{W}_0 + \sum_{l=0}^{n_m-1}\left\{h_{l+1}\sum_{j=1}^{\nu}\Lambda_l^j\left(\frac{\partial f}{\partial c}\right)_{l+1}^j\right.\\ &- h_{l+1}\sum_{j=1}^{\nu}\Pi_l^j\left(H_{l+1}(f_{past})_{l+1}^j + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\right)\left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j\Bigg\}\\ &+ \sum_{l=-n_s-1}^{-1}\sum_{(\mu,\rho)\in\mathcal{M}_l}h_{\mu+1}\Pi_\mu^\rho\left(\frac{\partial\phi_l}{\partial c}\right)_{\mu+1}^\rho. \end{aligned} \tag{8.49}$$

Herein, $\mathbf{W}_0$ is given by equation (8.20), which requires only the right-hand-side function evaluation at the initial time, the derivatives of the initial time $t_0 = t^{ini}(c)$ and the initial value $y_0 = y^{ini}(c)$ with respect to parameters.

Obviously, expression (8.49) allows to compute the product $\Lambda_{n_m}\mathbf{W}_{n_m}$ only in terms of the adjoint variables, which are given by solving the discrete adjoint scheme (8.48). Additional contributions in the equations (8.48), (8.49) that occur for problems with discontinuities in the Wronskian are presented in Subsection 8.3.4.

Subsection (8.3.5) below discusses the relation of adjoint sensitivities to forward sensitivities and the convergence of adjoint derivatives to the exact derivative.

8.3.3. Sensitivities of the States at Inner Time Points

The discrete adjoint approach for sensitivity computation can also be used for computing the derivatives of states at inner time points of the considered interval, i.e. for the computation of $dy(\bar{t};c)/dc$ for $\bar{t} \in (t^{ini}(c), t^{fin}(c))$. In particular, this is also possible if $\bar{t}$ is not part of the mesh that was used for the forward solution of the nominal DDE-IVP.

For illustration, consider a scalar function $\Omega(y(\bar{t};c))$ whose derivative with respect to the parameters c should be computed. For this purpose, set

$$\Lambda_{\bar{y}} := \frac{\partial\Omega}{\partial y}(y(\bar{t};c)) \tag{8.50}$$

and observe that

$$\frac{d\Omega}{dc}(y(\bar{t};c)) = \Lambda_{\bar{y}}\frac{dy(\bar{t};c))}{dc}. \tag{8.51}$$

Let $\bar{y}$ be the numerical approximation of $y(\bar{t};c)$ obtained by a continuous Runge-Kutta method applied to the nominal DDE-IVP, and let $\bar{n}$ be the integer number such that $t_{\bar{n}} < \bar{t} \leqslant t_{\bar{n}+1}$. Then a numerical approximation of $dy(\bar{t};c))/dc$ by a CRK scheme for forward sensitivity computation is given by

$$\frac{d\bar{y}}{dc} = \mathbf{W}_{\bar{n}} + h_{\bar{n}+1}\sum_{j=1}^{\nu}b_j(\bar{\theta})\mathbf{G}_{\bar{n}+1}^j. \tag{8.52}$$

Therein $\mathbf{W}_{\bar{n}}$ represents $\partial y_{\bar{n}}/\partial c$ and $\mathbf{G}_{\bar{n}+1}^j$ represents $\partial g_{\bar{n}+1}^j/\partial c$. Further, $\bar{\theta} = (\bar{t}-t_{\bar{n}})/h_{\bar{n}+1} \in (0,1]$ denotes the relative position of $\bar{t}$ on the interval $[t_{\bar{n}}, t_{\bar{n}+1}]$.

In order to avoid the computation of forward sensitivities, a variational approach is used:

$$\Lambda_{\bar{y}}\left(\frac{d\bar{y}}{dc} - \mathbf{W}_{\bar{n}} - h_{\bar{n}+1}\sum_{j=1}^{\nu}b_j(\bar{\theta})\mathbf{G}_{\bar{n}+1}^j\right) = 0. \tag{8.53}$$

This equation, in combination with the variational approach for the continuous Runge-Kutta method for the steps up to the mesh point $t_{\bar{n}+1}$ suggests to make the following initialization step for the discrete adjoint scheme:

$$\Lambda_{\bar{n}} = \Lambda_{\bar{y}} - h_{\bar{n}+1}\sum_{j=1}^{\nu}\Gamma_{\bar{n}}^j + \sum_{(\mu,\rho)\in\mathcal{M}_{\bar{n}},\mu=\bar{n}}h_{\mu+1}\Pi_\mu^\rho \tag{8.54a}$$

$$\Lambda_{\bar{n}}^{j} = \Lambda_{\bar{y}} b_j(\bar{\theta}) - h_{\bar{n}+1} \sum_{i=1}^{\nu} a_{i,j} \Gamma_{\bar{n}}^{i} + \sum_{(\mu,\rho)\in\mathcal{M}_{\bar{n}},\mu=\bar{n}} h_{\mu+1} \Pi_{\mu}^{\rho} b_j(\theta_{\mu,\rho}) \tag{8.54b}$$

$$\Gamma_{\bar{n}}^{j} = -\Lambda_{\bar{n}}^{j} \left(\frac{\partial f}{\partial y}\right)_{\bar{n}+1}^{j} + \Pi_{\bar{n}}^{j} \left(H_{\bar{n}+1} (f_{past})_{\bar{n}+1}^{j} + (1 - H_{\bar{n}+1}) \left(\frac{d\phi_{\xi(\bar{n}+1)}}{dt}\right)_{\bar{n}+1}^{j} \right) \left(\frac{\partial \tau_1}{\partial y}\right)_{\bar{n}+1}^{j} \tag{8.54c}$$

$$\Pi_{\bar{n}}^{j} = \Lambda_{\bar{n}}^{j} \left(\frac{\partial f}{\partial v}\right)_{\bar{n}+1}^{j}. \tag{8.54d}$$

This initialization step differs from a usual step in the discrete adjoint scheme (8.48) in two respects. On the one hand, the first terms in the equation (8.54b) contains a factor $b_j(\bar{\theta})$ instead of β_j. On the other hand, the summation in the equations (8.54a) and (8.54b) goes only over index pairs (μ, ρ) for which overlapping occured in the step $t_{\bar{n}} \to t_{\bar{n}+1}$.

After the initialization step (8.54), the discrete adjoint scheme (8.48) can be applied for $l = \bar{n} - 1, \ldots, 0$. Thereby, it has also be taken into account that the summations in the equations (8.48a) and (8.48b) only go over those pairs (μ, ρ) for which $\mu \leqslant \bar{n}$. As a result, the sought derivative can be computed by

$$\begin{aligned} \Lambda_{\bar{y}} \frac{d\bar{y}}{dc} = &+ \Lambda_0 \mathbf{W}_0 + \sum_{l=0}^{\bar{n}} \Bigg[h_{l+1} \sum_{j=1}^{\nu} \Lambda_l^j \left(\frac{\partial f}{\partial c}\right)_{l+1}^{j} \\ &- h_{l+1} \sum_{j=1}^{\nu} \Pi_l^j \left(H_{l+1} (f_{past})_{l+1}^{j} + (1 - H_{l+1}) \left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^{j} \right) \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{j} \Bigg] \\ &+ \sum_{l=-n_s-1}^{-1} \sum_{(\mu,\rho)\in\mathcal{M}_l,\mu\leqslant\bar{n}} h_{\mu+1} \Pi_{\mu}^{\rho} \left(\frac{\partial \phi_l}{\partial c}\right)_{\mu+1}^{\rho}. \end{aligned} \tag{8.55}$$

8.3.4. Discontinuities of Order 0 in the Sensitivities

According to the differentiability theorem for general DDE-IVPs (Theorem 7.11) the exact derivative $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$ has discontinuities that are potentially of order 0. For example, let s_i be the time point of a propagated discontinuity whose parent discontinuity at $s_j(c)$ is of order 0 in y. Further, let t_l be the corresponding mesh point (i.e. $t_l = s_i$). Then the jump in the Wronskian $\mathbf{W}$ was taken into account in the idealized variant of the modified standard approach by equation (8.31). For the adjoint sensitivity computation, it is appropriate to express this equation as

$$\mathbf{W}_l^+ = \mathbf{A}_l^1 \mathbf{W}_l^- + \mathbf{A}_l^2. \tag{8.56}$$

Herein, the matrices $\mathbf{A}_l^1$ and $\mathbf{A}_l^2$ are given by

$$\mathbf{A}_l^1 = \mathbf{1}_{n_y} + (f_l^-(c) - f_l^+(c)) \left(\frac{\frac{\partial \tau_1(t_l, y_l, c)}{\partial y}}{1 - \frac{\partial \tau_1(t_l, y_l, c)}{\partial t} - \frac{\partial \tau_1(t_l, y_l, c)}{\partial y} f_l^-(c)} \right) \tag{8.57a}$$

$$\mathbf{A}_l^2 = +(f_l^-(c) - f_l^+(c)) \left(\frac{\frac{\partial \tau_1(t_l, y_l, c)}{\partial c} + \frac{ds_j}{dc}}{1 - \frac{\partial \tau_1(t_l, y_l, c)}{\partial t} - \frac{\partial \tau_1(t_l, y_l, c)}{\partial y} f_l^-(c)} \right), \tag{8.57b}$$

and $\mathbf{1}_{n_y}$ is the identity matrix of dimension $n_y \times n_y$.

Note that discontinuities of order 0 in y arise, for DDE-IVPs, only because of discontinuities in the initial function. It can therefore be assumed that ds_j/dc in the expression for $\mathbf{A}_l^2$ is in practice available as part of the problem formulation.

Consider the following variational approach for equation (8.56):

$$\Lambda_l^+ (\mathbf{W}_l^+ - \mathbf{A}_l^1 \mathbf{W}_l^- - \mathbf{A}_l^2) = 0. \tag{8.58}$$

Hence, discontinuities of order 0 can be taken into account in the discrete adjoint approach as follows: During the application of the discrete adjoint scheme, stop at all mesh points t_l that are identified as approximations of propagated discontinuity points. Then compute $\Lambda_l^- := \Lambda_l^+ \mathbf{A}_l^1$, and continue the application of the discrete adjoint scheme with Λ_l^-.

Further, let $\mathcal{N}_j$ be a set that contains all indices l of the mesh points where the child discon-

tinuities of s_j are located (for all $-n_s^\phi \leqslant j \leqslant 0$). Formally: an index l is contained in $\mathcal{N}_j$ if $t_l - \tau_1(t_l, y_l, c) = s_j$. With this definition, a term

$$\sum_{j=-n_s^\phi}^{0} \sum_{l \in \mathcal{N}_j} \Lambda_l^+ \mathbf{A}_l^2$$

is added to equation (8.49).

8.3.5. Equivalence of Forward Approach and Adjoint Approach

The discrete adjoint scheme was derived by taking the equations for the forward approach, multiplying these expressions by adjoint variables, and reordering the terms in the obtained equation in such a way that the forward sensitivities do not need to be computed because they are multiplied by zeros. As a result, the equation (8.49) is obtained. Hence, computation of the sought derivative with the forward approach (left hand side of that equation) or by the adjoint approach (right hand side of that equation) yields, theoretically, exactly the same result, i.e. the two approaches are fully equivalent. In particular, it is also possible to compute the full Wronskian matrix $\mathbf{W}_{n_m}$ by starting the discrete adjoint scheme with the n_y unit vectors of $\mathbb{R}^{n_y}$.

The exact equivalence of the forward and adjoint Internal Numerical Differentiation approaches are also known from ODE- and DAE-theory, see e.g. Bock [39], Albersmeyer [2]. Due to the equivalence, it is immediately clear that the adjoint sensitivities converge, for $h \to 0$, $h := \max_{1 \leqslant l \leqslant n_m} h_l$, with the same convergence order to the exact derivative as the forward sensitivities, provided that the assumptions of Corollary 8.1 are fulfilled. Moreover, convergence is also obtained in the framework of the practical variant of the modified standard approach if the propagation switching functions have a regular behavior (see Section 5.4 and Subsection 8.2.10).

In the numerical practice it happens frequently that forward sensitivities and adjoint sensitivities do not yield exactly the same result. The reason is the use of floating point numbers and operations. Accordingly, equation (8.49) holds exactly in theory, but due to the fact that a different algorithm is used small deviations may occur. Typically, these deviations are very small, such that forward and adjoint sensitivities are often identical in the leading 10 or more digits.

Please note that an excellent agreement between forward and adjoint sensitivities is not a measure for the accuracy of the computed sensitivities. A good agreement is also obtained if only one or two digits are correct as compared to the exact derivative.

8.3.6. Error Control

The adjoint sensitivities are obtained by computing the discrete adjoints that fit exactly to one specific forward computation method for sensitivity computation and for the specific mesh that was used for the solution of the nominal DDE-IVP. Hence, the mesh has to be considered as fixed. The discrete adjoint approach for sensitivity computation does therefore not allow to adapt the stepsizes in order to control the error of the approximation.

8.3.7. Generalization: Sensitivities of IHDDE-IVP Solutions

This subsection deals with the generalization of the discrete adjoint approach for sensitivity computation to the general IHDDE-IVP case. For this purpose it is necessary to take into account the jumps in the Wronskian that may occur at the time points of the root discontinuities.

In order to discuss this issue in detail, consider the time point s_k of a root discontinuity, and let $I(k)$ denote the index of the switching function that is zero at s_k. Further, consider the idealized variant of the modified standard approach and let l be the index of the mesh point that corresponds to the time point of the root discontinuity, i.e. $t_l = s_k$. Assuming that the sufficient differentiability conditions of Section 7.6 are fulfilled (which implies, in particular, a non-zero time derivative of the switching function $\sigma_{I(k)}$ at its zero s_k), the exact derivative ds_k/dc of the time point of the root discontinuity is given by equation (7.105).

The discrete analogue of this equation can, for the special case of a single delay, be expressed as

$$\frac{ds_k}{dc} = \mathbf{B}_l^1 \mathbf{W}_l^- + \mathbf{B}_l^2 \tag{8.59}$$

with matrices $\mathbf{B}_l^1$ and $\mathbf{B}_l^2$ that are given by

$$\mathbf{B}_l^1 = -\frac{\frac{\partial \sigma_{I(k)}}{\partial y} - \frac{\partial \sigma_{I(k)}}{\partial v_1}\dot{y}_{past}^{k,1}\frac{\partial \tau_1}{\partial y}}{\frac{\partial \sigma_{I(k)}}{\partial t} + \frac{\partial \sigma_{I(k)}}{\partial y}f_l^-(c) + \frac{\partial \sigma_{I(k)}}{\partial v_1}\dot{y}_{past}^{k,1}\left[1 - \frac{\partial \tau_1}{\partial t} - \frac{\partial \tau_1}{\partial y}f_l^-(c)\right]} \tag{8.60a}$$

$$\mathbf{B}_l^2 = -\frac{\frac{\partial \sigma_{I(k)}}{\partial c} - \frac{\partial \sigma_{I(k)}}{\partial v_1}\dot{y}_{past}^{k,1}\frac{\partial \tau_1}{\partial c} + \frac{\partial \sigma_{I(k)}}{\partial v_1}\mathbf{W}_{past}^{k,1}}{\frac{\partial \sigma_{I(k)}}{\partial t} + \frac{\partial \sigma_{I(k)}}{\partial y}f_l^-(c) + \frac{\partial \sigma_{I(k)}}{\partial v_1}\dot{y}_{past}^{k,1}\left[1 - \frac{\partial \tau_1}{\partial t} - \frac{\partial \tau_1}{\partial y}f_l^-(c)\right]}. \tag{8.60b}$$

Herein, the partial derivatives of the switching function $\sigma_{I(k)}$ are evaluated at (t_l, y_l^-, c, v_l^-), where y_l^- represents the numerical approximation of $y^-(t_l; c)$ obtained by using the idealized variant of the modified standard approach. Further, v_l^- is an approximation of the past state $y(t_l - \tau_1(t_l, y_l^-, c); c)$, which is obtained by an evaluation of a deduced function.

The partial derivatives of the delay function τ_1 in equation (8.60) are evaluated at the arguments (t_l, y_l^-, c). The symbol $f_l^-(c)$ represents the evaluation of the right-hand-side function to the left of the mesh point t_l, i.e. $f_l^-(c) = f(t_l, y_l^-, c, v_l^-, \zeta^-)$, where ζ^- represents the signs of the switching functions to the left of t_l. Moreover, $\dot{y}_{past}^{k,1}$ represents a numerical approximation of $\dot{y}(t_l - \tau_1(t_l, y_l^-, c); c)$, which, for the discrete adjoint of the CRK scheme (8.21), (8.30), is given by an evaluation of the right-hand-side function at the past time point. Eventually, $\mathbf{W}_{past}^{k,1}$ represents a numerical approximation of $\mathbf{W}(t_l - \tau_1(t_l, y_l^-, c); c)$. This approximation has to be a obtained by an evaluation of a deduced function as usual in the context of the modified standard approach.

The exact jump in the Wronskian matrix $\mathbf{W}(t; c)$ at the discontinuity point s_k is given by equation (7.108). The discrete analogue of this equation is given by

$$\mathbf{W}_l^+ = \mathbf{C}_l^1\mathbf{W}_l^- + \mathbf{C}_l^2\frac{ds_k}{dc} + \mathbf{C}_l^3. \tag{8.61}$$

Therein, ds_k/dc is given by equation (8.59), and the matrices $\mathbf{C}_l^1$, $\mathbf{C}_l^2$, and $\mathbf{C}_l^3$ are given by

$$\mathbf{C}_l^1 = \mathbf{1}_{n_y} + \frac{\partial \omega_{I(k)}}{\partial y} - \frac{\partial \omega_{I(k)}}{\partial v_1}\dot{y}_{past}^{k,1}\frac{\partial \tau_1}{\partial y} \tag{8.62a}$$

$$\mathbf{C}_l^2 = \left(\mathbf{1}_{n_y} + \frac{\partial \omega_{I(k)}}{\partial y}\right)f_l^-(c) + \frac{\partial \omega_{I(k)}}{\partial t} - f_l^+(c) + \frac{\partial \omega_{I(k)}}{\partial v_1}\dot{y}_{past}^{k,1}\left(1 - \frac{\partial \tau_1}{\partial t} - \frac{\partial \tau_1}{\partial y}f_l^-(c)\right) \tag{8.62b}$$

$$\mathbf{C}_l^3 = \frac{\partial \omega_{I(k)}}{\partial c} - \frac{\partial \omega_{I(k)}}{\partial v_1}\dot{y}_{past}^{k,1}\frac{\partial \tau_1}{\partial c} + \frac{\partial \omega_{I(k)}}{\partial v_1}\mathbf{W}_{past}^{k,1}, \tag{8.62c}$$

where $\mathbf{1}_{n_y}$ is the $n_y \times n_y$ dimensional identity matrix. Further, the impulse function $\omega_{I(k)}$ is evaluated at the arguments (t_l, y_l^-, c, v_l^-), and $f_k^+(c)$ is given by $f_l^+(c) = f(t_l, y_l^+, c, v_l^+, \zeta^+)$. Thereby, y_l^+ is the numerical approximation of the state after the impulse, i.e.

$$y_l^+ = y_l^- + \omega_{I(k)}(t_l, y_l^-, c, v_l^-), \tag{8.63}$$

and v_l^+ is an approximation of $y(t_l - \tau_1(t_l, y_l^+, c); c)$ that is computed by an evaluation of a deduced function. Eventually, ζ^+ represents the switching function signs to the right of t_l.

Inserting equation (8.59) into (8.61) yields

$$\mathbf{W}_l^+ = \mathbf{D}_l^1\mathbf{W}_l^- + \mathbf{D}_l^2 \tag{8.64}$$

with $\mathbf{D}_l^1 = \mathbf{C}_l^1 + \mathbf{C}_l^2\mathbf{B}_l^1$ and $\mathbf{D}_l^2 = \mathbf{C}_l^3 + \mathbf{C}_l^2\mathbf{B}_l^2$.

The general idea is to use the procedure that was used in the context of children of discontinuities of order 0 in y, see Subsection 8.3.4. Hence, consider the variational approach

$$\Lambda_l^+(\mathbf{W}_l^+ - \mathbf{D}_l^1\mathbf{W}_l^- - \mathbf{D}_l^2) = 0. \tag{8.65}$$

Accordingly, during the application of the discrete adjoint scheme, it is necessary to stop at the time points of root discontinuities, and to compute the jump in the adjoint variables, i.e. $\Lambda_l^- := \Lambda_l^+\mathbf{D}_l^1$. Subsequently, the application of the discrete adjoint scheme is continued with Λ_l^-.

It then remains to deal with the contribution of the matrix $\mathbf{D}_l^2$. This matrix contains, in particular, terms that depend on the Wronskian $\mathbf{W}_{past}^{k,1}$ at the past time point. If the discontinuity

interval indicator $\xi[l]$ for the sole delay τ_1 is the step $t_{l-1} \to t_l$ is such that $1 \leqslant \xi[l] \leqslant n_s$, then this Wronskian matrix is, in the forward approach, approximated by the continuous representation $E_{l'+1}(t)$, with $l'+1 \leqslant l$. In order to avoid the computation of this quantity in the adjoint approach, the terms $W_{l'}$ and $G^j_{l'+1}$ have to be factored in the integration step $t_{l'} \to t_{l'+1}$. This leads to a modification of the discrete adjoint scheme (8.48).

Only the result is given in the following. Let the set $\mathcal{P}_l$ contain all those indices k of the time points of root discontinuities for which the deviating argument is located in $[t_l, t_{l+1})$. Then the following modified versions of equation (8.48a) and (8.48b) are obtained:

$$\Lambda_l = \Lambda_{l+1} - h_{l+1} \sum_{j=1}^{\nu} \Gamma_l^j + \sum_{(\mu,\rho)\in\mathcal{M}_l} h_{\mu+1}\Pi_\mu^\rho + \sum_{k\in\mathcal{P}_l} \left(\frac{\partial \omega_{I(k)}}{\partial v_1} - \frac{\mathbf{C}^2_{J(k)} \frac{\partial \sigma_{I(k)}}{\partial v_1}}{\dot{\sigma}^-_{I(k)}} \right) \tag{8.66a}$$

$$\begin{aligned} \Lambda_l^j =& \Lambda_{l+1}\beta_j - h_{l+1} \sum_{i=1}^{\nu} a_{i,j}\Gamma_l^i + \sum_{(\mu,\rho)\in\mathcal{M}_l} h_{\mu+1}\Pi_\mu^\rho b_j(\theta_{\mu,\rho}) \\ &+ \sum_{k\in\mathcal{P}_l} \left(\frac{\partial \omega_{I(k)}}{\partial v_1} - \frac{\mathbf{C}^2_{J(k)} \frac{\partial \sigma_{I(k)}}{\partial v_1}}{\dot{\sigma}^-_{I(k)}} \right) b_j(\theta_{I(k)}) \end{aligned} \tag{8.66b}$$

where $J(k)$ is the function that attributes to the discontinuity point s_k the mesh point $t_{J(k)}$ at which it occurs and

$$\dot{\sigma}^-_{I(k)} := \frac{\partial \sigma_{I(k)}}{\partial t} + \frac{\partial \sigma_{I(k)}}{\partial y} f_k^-(c) + \frac{\partial \sigma_{I(k)}}{\partial v_1} \dot{y}^{k,1}_{past} \left[1 - \frac{\partial \tau_1}{\partial t} - \frac{\partial \tau_1}{\partial y} f_k^-(c) \right]. \tag{8.67}$$

Further, $\theta_{I(k)}$ is the relative position of the past time point $t_{J(k)} - \tau_1(t_{J(k)}, y_{J(k)}, c)$ in the interval $[t_l, t_{l+1})$.

The remaining terms in $\mathbf{D}_l^2$, i.e. those that do not depend on the Wronskian at past time points, have to be taken into account by additional terms in equation (8.49). In order to do this, let $\mathcal{Q}$ denote the set of those indices $k \in \{1, \ldots, n_s\}$ such that s_k is the time point of a root discontinuity. This leads to the following contributions in equation (8.49):

$$\sum_{k\in\mathcal{Q}} \frac{\partial \omega_{I(k)}}{\partial c} - \frac{\partial \omega_{I(k)}}{\partial v_1} \dot{y}^{k,1}_{past} \frac{\partial \tau_1}{\partial c} - \frac{\mathbf{C}^2_{J(k)} \left[\frac{\partial \sigma_{I(k)}}{\partial c} - \frac{\partial \sigma_{I(k)}}{\partial v_1} \dot{y}^{k,1}_{past} \frac{\partial \tau_1}{\partial c} \right]}{\dot{\sigma}^-_{I(k)}}.$$

Of course, the treatment of IHDDE-IVPs still requires to take into account possible jumps in the sensitivities that originate from propagations of discontinuities of order 0 in y as discussed in Subsection 8.3.4.

The combined presence of non-zero impulse functions and delays may also leads to a situation where a root discontinuity of order 0 in y has a child discontinuity within the considered time interval. Since the right hand side of equation (8.57b) contains a term that represents the total derivative of the time point of the parent discontinuity, this term can be taken into account in the adjoint sensitivity computation when the time point of the parent discontinuity (i.e. the time point of the root discontinuity of order 0) is reached later (during the backward recursion). More precisely, this leads to an additional contribution in the matrix $\mathbf{C}_l^2$.

8.3.8. Practical Variant of the Modified Standard Approach

In Subsection 8.3.2 a discrete adjoint scheme was derived that fits exactly to the forward Internal Numerical Differentiation method (8.21), (8.30). This forward scheme was originally derived in Subsections 8.2.2 and 8.2.3 in the context of the idealized variant of the modified standard approach. Further, the jumps in the Wronskian matrix in the time points of propagated discontinuities (see Subsection 8.3.4) and in the time points of root discontinuities (see Subsection 8.3.7) have also been derived in the framework of the idealized variant.

However, as discussed in Subsection 8.2.10, the practical variant differs from the idealized variant only with regard to the underlying assumption on the construction of the mesh. Therefore, all equations derived in the Subsection 8.3.2, 8.3.4, and 8.3.7 formally remain valid also in the context of the practical variant.

9. Sensitivity Computation in Colsol-DDE

Ableitungen liegen nicht auf der Straße.

Heard in a lecture by Utz Wever (TU München and Siemens AG).

In Chapter 8 a generalization of Internal Numerical Differentiation (IND) for delay differential equations (DDEs) has been proposed. Further, forward and adjoint methods for sensitivity computation have been presented, which realize IND for DDEs and which allow for an accurate computation of the sensitivities. This chapter discusses how these forward and adjoint IND methods are realized in practice in the newly developed solver Colsol-DDE.

Survey of Existing Solvers with Sensitivity Computation

Computation of sensitivities of initial value problem solutions in ordinary differential equations (ODEs) can be done by a variety of existing programs. It is therefore sufficient to direct the reader to the following codes: The variants of DIFSYS and METAN1 as described in Bock [36], DAESOL by Bauer [20] and Bauer, Bock, and Schlöder [21], DASPK by Li and Petzold [176, 177], and CVODES by Hindmarsh and Serban [147]. Without going into the details, it should be mentioned that these codes differ significantly with respect to the approaches that are realized (forward vs. adjoint), their capabilities (first, second, or higher order sensitivities), and to the extent to which structures are exploited. For a recent code that features efficient arbitrary order forward and adjoint sensitivity computation in differential-algebraic equations, it is referred to DAESOL-II by Albersmeyer [2].

Sensitivity computation in hybrid discrete-continuous ordinary differential equations (HODEs) and impulsive hybrid discrete-continuous ordinary differential equations (IHODEs) is rarely available. For two codes that provide this feature, see DAEPACK by Tolsma and Barton [249] and RKFSWT by Kirches [160]. Both programs compute the sensitivities only in forward mode. To the knowledge of the author, there is presently no solver that computes adjoint sensitivities of initial value problem (IVP) solutions in HODEs and IHODEs.

For problems with time delays, only one program has been developed so far that features sensitivity computation, DDEM by ZivariPiran [271] and ZivariPiran and Enright [273]. The approach taken by DDEM is to solve the variational DDE-IVP and to stop at the time points of the propagated discontinuities in order to apply the jumps to the Wronskian matrix. DDEM is not able to compute adjoint sensitivities.

Features of the New Solver Colsol-DDE

Colsol-DDE is the first solver that features an accurate and efficient computation of first order forward and adjoint sensitivities by means of Internal Numerical Differentiation in the general case of IHDDE-IVPs. Since IHODE-IVPs and DDE-IVPs are included as special cases, Colsol-DDE is thus also the first solver to provide adjoint sensitivities for these simpler classes of IVPs.

Further, with regard to forward sensitivity computation for DDE-IVPs, the methods in Colsol-DDE differ in many respects to the ones implemented in DDEM. In particular, Colsol-DDE strictly follows the IND concept (see Definition 8.2), i.e. the realized methods for sensitivity computation are obtained by applying IND to the numerical methods that were presented in Chapter 6. An analysis of the resulting equation systems reveals structures that are exploited in Colsol-DDE. This is particularly important because Colsol-DDE is based on implicit methods.

Another key feature of Colsol-DDE is that error-controlled sensitivities can be computed. This feature is non-standard even among sensitivity-generating ODE-IVP solvers.

Eventually, it is pointed out that Colsol-DDE is – in contrast to DDEM – designed to be used in conjunction with an Automatic Differentiation tool, which provides (up to machine precision) the exact derivatives of the model functions. This improves significantly the achievable precision for the computed sensitivities. At the same time, the use of Automatic Differentiation for the computation

of the model function derivatives makes the usage of Colsol-DDE entirely derivative-free for the user.

Organization of This Chapter

The chapter is divided into two sections. In Section 9.1, the practically implemented methods for the computation of forward sensitivities are presented. Section 9.2 deals with the realization of adjoint sensitivity computation.

9.1. Practical Computation of Forward Sensitivities

9.1.1. Collocation Method

A first discrete and continuous approximation of a DDE-IVP solution $y(t)$ in the step $t_l \to t_{l+1}$ is, in Colsol-DDE, obtained by the application of a collocation method, see Section 6.1. This provides a value $y_{l+1,p}$ and a polynomial continuous representation $\eta_{l+1,q}(t_l + \theta h_{l+1})$, $\theta \in [0,1]$, which approximate the exact local solution in that integration step with discrete local order p and with uniform local order q, respectively. Note that, in agreement with Section 6.4, the local orders of the approximations are indicated by subscripts.

More precisely, the employed collocation methods in Colsol-DDE are the one-stage Gauss method, the two-stage Radau IIA method, and the three-stage Lobatto IIIA method. Their abscissae γ_i, coefficients $a_{i,j}$, weights β_j, and continuous weight functions b_j were given in Subsection 6.1.2 and in Table 6.1. Recall that the pairs of discrete local orders and uniform local orders for these methods are $(p,q) = (2,1)$ for the one-stage Gauss method, $(p,q) = (3,2)$ for the two-stage Radau IIA method, and $(p,q) = (4,3)$ for the three-stage Lobatto IIIA method.

The starting point of the discussion is the $\nu \times n_y$-dimensional equation system ($(\nu-1) \times n_y$-dimensional in the case of the Lobatto IIIA method) that needs to be solved for the collocation method. For notational simplicity, the case of a DDE-IVP with a single delay τ_1 is regarded in the following. In this case, the system that needs to be solved was given in equation (6.53), which is recalled here:

$$g_{l+1}^j = f(t_{l+1}^j, y_{l+1}^j, c, v_{l+1}^j) \quad \text{for} \quad 1 \leqslant j \leqslant \nu. \tag{9.1}$$

Therein, y_{l+1}^j is given by

$$y_{l+1}^j = y_l + h_{l+1} \sum_{i=1}^{\nu} a_{j,i} g_{l+1}^i. \tag{9.2}$$

After the system (9.1) has been solved, the discrete and continuous numerical approximations of the DDE-IVP solution in the interval $[t_l, t_{l+1}]$ are given by

$$y_{l+1,p} = y_l + h_{l+1} \sum_{j=1}^{\nu} \beta_j g_{l+1}^j \tag{9.3a}$$

$$\eta_{l+1,q}(t_l + \theta h_{l+1}) = y_l + h_{l+1} \sum_{j=1}^{\nu} b_j(\theta) g_{l+1}^j. \tag{9.3b}$$

The approximations v_{l+1}^j of the past states are, in Colsol-DDE, computed by the practical variant of the modified standard approach (see Definition 5.22, and also Subsection 8.2.10). This means that a deduced function is evaluated for the computation of v_{l+1}^j, and the selection of the deduced function depends on the value of the numerically determined discontinuity interval indicator $\hat{\xi}_1^\alpha(t)$ for the sole delay τ_1 and for $t \in (t_l, t_{l+1})$.

Since only a single integration step $t_l \to t_{l+1}$ is considered in the following, the notation ξ is used as an abbreviation for $\hat{\xi}_1^\alpha(t')$, $t' \in (t_l, t_{l+1})$ arbitrary.

Let n_s^ϕ denote the total number of discontinuities to left of the initial time, and let n_s denote the total number of propagated discontinuities[1] up to order $p+1$ to the right of the initial time

[1] In Colsol-DDE, the time points of all children of discontinuities up to order $p+1$ need to be included into the mesh, because the discrete error-estimating method is of order $p+1$. Further, when making the decision on the

that were found until the mesh point t_l. Then, according to Subsection 6.4.1, the computation of past states is done in the following way:

- If $-n_s^\phi \leqslant \xi \leqslant 0$, then the past states are computed by an evaluation of a smooth branch ϕ_ξ of the initial function ϕ.
- If $1 \leqslant \xi \leqslant n_s + 1$, then the past states are computed from the continuous representation in the step $[t_{l'}, t_{l'+1}]$. The index l' is thereby different for every stage j (and for different indices $l+1$ of the integration step, of course). For the practical choice of l', see Subsection 6.4.1.

  If overlapping does not occur, i.e. $l' < l$, then the past states are computed by an evaluation of the continuous representation $\eta_{l'+1,p}$ of uniform local order p in the past integration step $t_{l'} \to t_{l'+1}$.

  If overlapping occurs, i.e. $l' = l$, then the past states are computed by an evaluation of the continuous representation $\eta_{l+1,q}$ of uniform local order q in the current integration step $t_l \to t_{l+1}$ (because the higher order continuous representation $\eta_{l+1,p}$, obtained from the implicit uniform correction, is not yet available).

The application of Internal Numerical Differentiation to the equations (9.1) and (9.2) then yields, at first,

$$\mathbf{G}_{l+1}^j = \left(\frac{\partial f}{\partial y}\right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial f}{\partial c}\right)_{l+1}^j + \left(\frac{\partial f}{\partial v}\right)_{l+1}^j \frac{dv_{l+1}^j}{dc} \tag{9.4}$$

and

$$\mathbf{W}_{l+1}^j = \mathbf{W}_l + h_{l+1} \sum_{i=1}^{\nu} a_{j,i} \mathbf{G}_{l+1}^i. \tag{9.5}$$

The discrete and continuous approximations for the sensitivities are then given

$$\mathbf{W}_{l+1,p} = \mathbf{W}_l + h_{l+1} \sum_{j=1}^{\nu} \beta_j \mathbf{G}_{l+1}^j \tag{9.6a}$$

$$\mathbf{E}_{l+1,q}(t_l + \theta h_{l+1}) = \mathbf{W}_l + h_{l+1} \sum_{j=1}^{\nu} b_j(\theta) \mathbf{G}_{l+1}^j. \tag{9.6b}$$

Note that these equations have also been found in Section 8.2 (see equation (8.21)), and that an (almost) identical notation has been used. For example, compare

$$\mathbf{W}_{l+1,p} := \frac{\partial y_{l+1,p}}{\partial c}, \quad \mathbf{W}_{l+1}^j := \frac{\partial y_{l+1}^j}{\partial c}, \quad \mathbf{G}_{l+1}^j := \frac{\partial g_{l+1}^j}{\partial c} \tag{9.7}$$

with equation (8.19), and

$$\left(\frac{\partial f}{\partial y}\right)_{l+1}^j := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(t_{l+1}^j, y_{l+1}^j, c, v_{l+1}^j)} \tag{9.8a}$$

$$\left(\frac{\partial f}{\partial c}\right)_{l+1}^j := \left.\frac{\partial f(t,y,c,v)}{\partial c}\right|_{(t_{l+1}^j, y_{l+1}^j, c, v_{l+1}^j)} \tag{9.8b}$$

$$\left(\frac{\partial f}{\partial v}\right)_{l+1}^j := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(t_{l+1}^j, y_{l+1}^j, c, v_{l+1}^j)} \tag{9.8c}$$

with equation (8.22).

For the computation of the derivatives of the past states, dv_{l+1}^j/dc, one option is the use of equation (8.30). As discussed in Section 8.2, this way of computing dv_{l+1}^j/dc is compatible with the concept of Internal Numerical Differentiation for DDEs (Definition 8.2). This ensures that

further propagation of a discontinuity, the orders in the approximations of both y and $\mathbf{W}$ are taken into account, because the error control strategy optionally applies to both the nominal solution and to the sensitivities.

the orders of the discrete and uniform local errors of $\mathbf{W}_{l+1,p}$ and $\mathbf{E}_{l+1,q}$ are indeed p and q (see Corollary 8.4), as indicated by the subscripts.

For the practical realization in Colsol-DDE, it is taken into account that different continuous representations are used in the overlapping and non-overlapping case. This leads to

$$\frac{dv_{l+1}^j}{dc} = \left(\frac{\partial\phi_\xi}{\partial c}\right)_{l+1}^j - \left(\frac{d\phi_\xi}{dt}\right)_{l+1}^j \left[\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j\right] \quad \text{if} \quad -n_s^\phi \leqslant \xi \leqslant 0, \tag{9.9}$$

to

$$\begin{aligned}\frac{dv_{l+1}^j}{dc} =& \mathbf{E}_{l'+1,p}(t_{l'} + \theta_{l,j} h_{l'+1}) - (f_{past})_{l+1}^j \cdot \left[\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j\right] \\ & \text{if } 1 \leqslant \xi \leqslant n_s + 1 \text{ and } l' < l,\end{aligned} \tag{9.10}$$

and to

$$\begin{aligned}\frac{dv_{l+1}^j}{dc} =& \mathbf{E}_{l+1,q}(t_l + \theta_{l,j} h_{l+1}) - (f_{past})_{l+1}^j \cdot \left[\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j\right] \\ & \text{if } \xi = n_s + 1 \text{ and } l' = l.\end{aligned} \tag{9.11}$$

In accordance with the notation of Section 8.2, the partial derivatives of the initial function ϕ and of the delay function τ_1 have thereby been abbreviated as

$$\left(\frac{d\phi_\xi}{dt}\right)_{l+1}^j := \left.\frac{d\phi_\xi(t,c)}{dt}\right|_{(t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c), c)} \tag{9.12a}$$

$$\left(\frac{\partial\phi_\xi}{\partial c}\right)_{l+1}^j := \left.\frac{\partial\phi_\xi(t,c)}{\partial c}\right|_{(t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c), c)} \tag{9.12b}$$

$$\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j := \left.\frac{\partial\tau_1(t,y,c)}{\partial y}\right|_{(t_{l+1}^j, y_{l+1}^j, c)} \tag{9.12c}$$

$$\left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j := \left.\frac{\partial\tau_1(t,y,c)}{\partial c}\right|_{(t_{l+1}^j, y_{l+1}^j, c)}. \tag{9.12d}$$

Further, the symbol $\theta_{l,j}$ was used in the equations (9.10) and (9.11) in order to represent the relative position of the past time point $t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c)$ in the interval $[t_{l'}, t_{l'+1}]$:

$$\theta_{l,j} = \frac{t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c) - t_{l'}}{h_{l'+1}}. \tag{9.13}$$

Eventually,

$$(f_{past})_{l+1}^j = f(t_{l'} + \theta_{l,j} h_{l'+1}, v_{l+1}^j, c, u_{l+1}^j) \tag{9.14}$$

is an abbreviating notation for an evaluation of the right-hand-side function at the past time point $(t_{past})_{l+1}^j := t_{l+1}^j - \tau_1(t_{l+1}^j, y_{l+1}^j, c)$. The argument u_{l+1}^j in this right-hand-side function evaluation is thereby

$$u_{l+1}^j \approx y((t_{past})_{l+1}^j - \tau_1((t_{past})_{l+1}^j, v_{l+1}^j, c); c), \tag{9.15}$$

meaning that it is an approximation of a state at a time point even further in the past, compare Subsection 8.2.3. The computation of u_{l+1} is thereby done depending on the value ξ_{past} of the discontinuity interval indicator in the past integration step $t_{l'} \to t_{l'+1}$.

9.1.2. Implicit Uniform Correction

Colsol-DDE uses the implicit uniform correction procedure as described in Section 6.2 to obtain a continuous representation $\eta_{l+1,p}(t_l + \theta h_{l+1})$ whose uniform local order is equal to p, i.e. equal to

the discrete local order of the collocation method. For this purpose, the system given in equation (6.60) is solved, i.e.

$$g^*_{l+1} = f(t^*_{l+1}, y^*_{l+1}, c, v^*_{l+1}) - \dot{\eta}_{l+1,q}(t^*_{l+1}), \tag{9.16}$$

where

$$y^*_{l+1} = y_l + h_{l+1}\left(\sum_{i=1}^{\nu} b_i(\theta^*) g^i_{l+1} + b_*(\theta^*) g^*_{l+1}\right) \tag{9.17a}$$

$$\dot{\eta}_{l+1,q}(t^*_{l+1}) = \sum_{i=1}^{\nu} b_i(\theta^*) g^i_{l+1}. \tag{9.17b}$$

After the system (9.16) has been solved, the higher order continuous representation in the interval $[t_l, t_{l+1}]$ is given by

$$\eta_{l+1,p}(t_l + \theta h_{l+1}) = y_l + h_{l+1}\left(\sum_{i=1}^{\nu} b_i(\theta) g^i_{l+1} + b_*(\theta) g^*_{l+1}\right). \tag{9.18}$$

The function $b_*(\theta)$ and the additional abscissa θ^* for the methods that are practically used in Colsol-DDE are defined in Subsection 6.2.2.

The approximation v^*_{l+1} of the past state is, in Colsol-DDE, computed as described in Subsection 6.4.2. Due to the use of the practical variant of the modified standard approach, the computation depends on the value of the discontinuity interval indicator ξ for the sole deviating argument as follows:

- If $-n^\phi_s \leqslant \xi \leqslant 0$, then the past state is computed by an evaluation of a smooth branch ϕ_ξ of the initial function ϕ.

- If $1 \leqslant \xi \leqslant n_s + 1$, then the past state is computed by an evaluation of the continuous representation $\eta_{l'+1,p}$ of uniform local order p in the step $t_{l'} \to t_{l'+1}$, where $l' \leqslant l$, i.e. regardless of whether or not overlapping occurs. For the practical choice of l', see Subsection 6.4.2.

In order to compute a higher order continuous representation of the sensitivities, the principle of Internal Numerical Differentiation is applied to equations (9.16) and (9.17). This yields

$$\mathbf{G}^*_{l+1} = \left(\frac{\partial f}{\partial y}\right)^*_{l+1} \mathbf{W}^*_{l+1} + \left(\frac{\partial f}{\partial c}\right)^*_{l+1} + \left(\frac{\partial f}{\partial v}\right)^*_{l+1} \frac{dv^*_{l+1}}{dc} - \dot{\mathbf{E}}_{l+1,q}(t^*_{l+1}) \tag{9.19}$$

and

$$\mathbf{W}^*_{l+1} = \mathbf{W}_l + h_{l+1}\left(\sum_{i=1}^{\nu} b_i(\theta^*) \mathbf{G}^i_{l+1} + b_*(\theta^*) \mathbf{G}^*_{l+1}\right) \tag{9.20a}$$

$$\dot{\mathbf{E}}_{l+1,q}(t^*_{l+1}) = \sum_{i=1}^{\nu} \dot{b}_i(\theta^*) \mathbf{G}^i_{l+1}. \tag{9.20b}$$

The higher order continuous representation for the sensitivities is then given by

$$\mathbf{E}_{l+1,p}(t_l + \theta h_{l+1}) = \mathbf{W}_l + h_{l+1}\left(\sum_{i=1}^{\nu} b_i(\theta) \mathbf{G}^i_{l+1} + b_*(\theta) \mathbf{G}^*_{l+1}\right). \tag{9.21}$$

In the above equations, the notations

$$\mathbf{W}^*_{l+1} := \frac{\partial y^*_{l+1}}{\partial c}, \quad \mathbf{G}^*_{l+1} := \frac{\partial g^*_{l+1}}{\partial c}, \tag{9.22}$$

and

$$\left(\frac{\partial f}{\partial y}\right)^*_{l+1} := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(t^*_{l+1},y^*_{l+1},c,v^*_{l+1})} \tag{9.23a}$$

$$\left(\frac{\partial f}{\partial c}\right)^*_{l+1} := \left.\frac{\partial f(t,y,c,v)}{\partial c}\right|_{(t^*_{l+1},y^*_{l+1},c,v^*_{l+1})} \tag{9.23b}$$

$$\left(\frac{\partial f}{\partial v}\right)^*_{l+1} := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(t^*_{l+1},y^*_{l+1},c,v^*_{l+1})}, \tag{9.23c}$$

have been used.

Further, in agreement with Definition 8.2 of Internal Numerical Differentiation for DDEs, a suitable expression for the derivative of the past state, dv^*_{l+1}/dc, is given by

$$\frac{dv^*_{l+1}}{dc} = \left(\frac{\partial \phi_\xi}{\partial c}\right)^*_{l+1} - \left(\frac{d\phi_\xi}{dt}\right)^*_{l+1}\left[\left(\frac{\partial \tau_1}{\partial y}\right)^*_{l+1}\mathbf{W}^*_{l+1} + \left(\frac{\partial \tau_1}{\partial c}\right)^*_{l+1}\right] \text{ if } -n_s^\phi \leqslant \xi \leqslant 0 \tag{9.24}$$

and by

$$\begin{aligned}\frac{dv^*_{l+1}}{dc} =& \mathbf{E}_{l'+1,p}(t_{l'} + \theta_{l,*}h_{l'+1}) - (f_{past})^*_{l+1}\cdot\left[\left(\frac{\partial \tau_1}{\partial y}\right)^*_{l+1}\mathbf{W}^*_{l+1} + \left(\frac{\partial \tau_1}{\partial c}\right)^*_{l+1}\right] \\ & \text{if } 1 \leqslant \xi \leqslant n_s+1, \quad l' \leqslant l.\end{aligned} \tag{9.25}$$

The partial derivatives of the initial function ϕ and of the delay function τ_1 have thereby been abbreviated as

$$\left(\frac{d\phi_\xi}{dt}\right)^*_{l+1} := \left.\frac{d\phi_\xi(t,c)}{dt}\right|_{(t^*_{l+1}-\tau_1(t^*_{l+1},y^*_{l+1},c),c)} \tag{9.26a}$$

$$\left(\frac{\partial \phi_\xi}{\partial c}\right)^*_{l+1} := \left.\frac{\partial \phi_\xi(t,c)}{\partial c}\right|_{(t^*_{l+1}-\tau_1(t^*_{l+1},y^*_{l+1},c),c)} \tag{9.26b}$$

$$\left(\frac{\partial \tau_1}{\partial y}\right)^*_{l+1} := \left.\frac{\partial \tau_1(t,y,c)}{\partial y}\right|_{(t^*_{l+1},y^*_{l+1},c)} \tag{9.26c}$$

$$\left(\frac{\partial \tau_1}{\partial c}\right)^*_{l+1} := \left.\frac{\partial \tau_1(t,y,c)}{\partial c}\right|_{(t^*_{l+1},y^*_{l+1},c)}. \tag{9.26d}$$

Further, the symbol $\theta_{l,*}$ was used in equation (9.25) in order to represent the relative position of the past time point $t^*_{l+1} - \tau_1(t^*_{l+1}, y^*_{l+1}, c)$ in the interval $[t_{l'}, t_{l'+1}]$:

$$\theta_{l,*} = \frac{t^*_{l+1} - \tau_1(t^*_{l+1}, y^*_{l+1}, c) - t_{l'}}{h_{l'+1}}. \tag{9.27}$$

Eventually,

$$(f_{past})^*_{l+1} = f(t_{l'} + \theta_{l,*}h_{l'+1}, v^*_{l+1}, c, u^*_{l+1}) \tag{9.28}$$

is an abbreviating notation for an evaluation of the right-hand-side function at the time point $(t_{past})^*_{l+1} := t^*_{l+1} - \tau_1(t^*_{l+1}, y^*_{l+1}, c)$. The argument u^*_{l+1} in this right-hand-side function evaluation is thereby

$$u^*_{l+1} \approx y((t_{past})^*_{l+1} - \tau_1((t_{past})^*_{l+1}, v^*_{l+1}, c); c), \tag{9.29}$$

meaning that it is an approximation of a state at a time point even further in the past, compare Subsection 8.2.3. The computation of u^*_{l+1} is thereby done depending on the value ξ_{past} of the discontinuity interval indicator in the past integration step $t_{l'} \to t_{l'+1}$,

9.1.3. Implicit Quadrature Rule

Colsol-DDE uses the implicit quadrature rule as described in Section 6.3 in order to obtain a discrete approximation $y_{l+1,p+1}$ whose discrete local order is $p+1$. For this purpose, the equation system given in equation (6.66) is solved, i.e.

$$g_{l+1}^{\diamond} = f(t_{l+1}, y_{l+1,p+1}, c, v_{l+1}^{\diamond}), \tag{9.30}$$

where

$$y_{l+1,p+1} = y_l + h_{l+1}\left(\sum_{i=1}^{\mu-1} B_i f(\bar{t}_{l+1}^i, \bar{y}_{l+1}^i, c, \bar{v}_{l+1}^i) + B_\mu g_{l+1}^{\diamond}\right) \tag{9.31}$$

denotes the sought approximation of discrete local order $p+1$.

The approximation $v_{l+1}^{\diamond}$ of the past state is, in Colsol-DDE, computed as described in Subsection 6.4.3. Since the implementation of Colsol-DDE closely follows the practical variant of the modified standard approach, the computation depends on the value of the discontinuity interval indicator ξ for the sole deviating argument as follows:

- If $-n_s^\phi \leqslant \xi \leqslant 0$, then the past state is computed by an evaluation of a smooth branch ϕ_ξ of the initial function ϕ.
- If $1 \leqslant \xi \leqslant n_s + 1$, then the past state is computed by an evaluation of the continuous representation $\eta_{l'+1,p}$ of uniform local order p in the integration step $t_{l'} \to t_{l'+1}$, where $l' \leqslant l$, i.e. regardless of whether or not overlapping occurs. For the practical choice of l', see Subsection 6.4.3.

The quadrature rules that are practically used in Colsol-DDE are discussed in Subsection 6.3.2. This defines, in particular, the abscissae $\bar{t}_{l+1}^i$ and the weights B_i. Moreover, $\bar{y}_{l+1}^i$ are given by evaluations of the higher order continuous representation in the current integration step, i.e. by $\eta_{l+1,p}(\bar{t}_{l+1}^i)$. Further, $\bar{v}_{l+1}^i$ are the corresponding past states, which are computed in a way analogous to $v_{l+1}^{\diamond}$, i.e. either by an evaluation of a smooth branch of the initial function, or by an evaluation of the continuous representation in the integration step $t_{l'} \to t_{l'+1}$ (where l' is dependent on the stage i of the quadrature rule and, of course, on the integration step l).

For the application of the implicit quadrature rule, the states $\bar{y}_{l+1}^i$ and $\bar{v}_{l+1}^i$ are known and fixed values.

In order to compute a higher order discrete approximation for the sensitivities, the principle of Internal Numerical Differentiation for DDEs, Definition 8.2, is applied to the equations (9.30) and (9.31). This gives

$$\mathbf{G}_{l+1}^{\diamond} = \left(\frac{\partial f}{\partial y}\right)_{l+1}^{\diamond} \mathbf{W}_{l+1,p+1} + \left(\frac{\partial f}{\partial c}\right)_{l+1}^{\diamond} + \left(\frac{\partial f}{\partial v}\right)_{l+1}^{\diamond} \frac{dv_{l+1}^{\diamond}}{dc}. \tag{9.32}$$

and

$$\begin{aligned}\mathbf{W}_{l+1,p+1} = \mathbf{W}_l + h_{l+1}\Bigg[&\sum_{i=1}^{\mu-1} B_i \left(\left(\frac{\overline{\partial f}}{\partial y}\right)_{l+1}^{i} \bar{\mathbf{W}}_{l+1}^i + \left(\frac{\overline{\partial f}}{\partial c}\right)_{l+1}^{i} + \left(\frac{\overline{\partial f}}{\partial v}\right)_{l+1}^{i} \frac{d\bar{v}_{l+1}^i}{dc}\right) \\ &+ B_\mu \mathbf{G}_{l+1}^{\diamond}\Bigg].\end{aligned} \tag{9.33}$$

This quantity represents the higher order discrete approximation of the sensitivities.

In the above equations, the notations

$$\mathbf{W}_{l+1,p+1} := \frac{\partial y_{l+1,p+1}}{\partial c}, \quad \mathbf{G}_{l+1}^{\diamond} := \frac{\partial g_{l+1}^{\diamond}}{\partial c}. \tag{9.34}$$

and

$$\left(\frac{\partial f}{\partial y}\right)_{l+1}^{\diamond} := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(t_{l+1}, y_{l+1,p+1}, c, v_{l+1}^{\diamond})} \tag{9.35a}$$

$$\left(\frac{\partial f}{\partial c}\right)_{l+1}^{\diamondsuit} := \left.\frac{\partial f(t,y,c,v)}{\partial c}\right|_{(t_{l+1},y_{l+1,p+1},c,v_{l+1}^{\diamondsuit})} \tag{9.35b}$$

$$\left(\frac{\partial f}{\partial v}\right)_{l+1}^{\diamondsuit} := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(t_{l+1},y_{l+1,p+1},c,v_{l+1}^{\diamondsuit})}. \tag{9.35c}$$

have been used.

Further, in agreement with Definition 8.2 of Internal Numerical Differentiation for DDEs, a suitable expression for the derivative of the past state, $dv_{l+1}^{\diamondsuit}/dc$, is given by

$$\frac{dv_{l+1}^{\diamondsuit}}{dc} = \left(\frac{\partial \phi_\xi}{\partial c}\right)_{l+1}^{\diamondsuit} - \left(\frac{d\phi_\xi}{dt}\right)_{l+1}^{\diamondsuit}\left[\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{\diamondsuit}\mathbf{W}_{l+1}^{\diamondsuit} + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{\diamondsuit}\right] \text{ if } -n_s^\phi \leqslant \xi \leqslant 0 \tag{9.36}$$

and by

$$\begin{aligned}\frac{dv_{l+1}^{\diamondsuit}}{dc} =& \mathbf{E}_{l'+1,p}(t_{l'} + \theta_{l,\diamondsuit}h_{l'+1}) - (f_{past})_{l+1}^{\diamondsuit}\cdot\left[\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{\diamondsuit}\mathbf{W}_{l+1}^{\diamondsuit} + \left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{\diamondsuit}\right] \\ & \text{if } 1 \leqslant \xi \leqslant n_s+1, \quad l' \leqslant l.\end{aligned} \tag{9.37}$$

The partial derivatives of the initial function ϕ and of the delay function τ_1 have thereby been abbreviated as

$$\left(\frac{d\phi_\xi}{dt}\right)_{l+1}^{\diamondsuit} := \left.\frac{d\phi_\xi(t,c)}{dt}\right|_{(t_{l+1}-\tau_1(t_{l+1},y_{l+1,p+1},c),c)} \tag{9.38a}$$

$$\left(\frac{\partial \phi_\xi}{\partial c}\right)_{l+1}^{\diamondsuit} := \left.\frac{\partial \phi_\xi(t,c)}{\partial c}\right|_{(t_{l+1}-\tau_1(t_{l+1},y_{l+1,p+1},c),c)} \tag{9.38b}$$

$$\left(\frac{\partial \tau_1}{\partial y}\right)_{l+1}^{\diamondsuit} := \left.\frac{\partial \tau_1(t,y,c)}{\partial y}\right|_{(t_{l+1},y_{l+1,p+1},c)} \tag{9.38c}$$

$$\left(\frac{\partial \tau_1}{\partial c}\right)_{l+1}^{\diamondsuit} := \left.\frac{\partial \tau_1(t,y,c)}{\partial c}\right|_{(t_{l+1},y_{l+1,p+1},c)}. \tag{9.38d}$$

Further, the symbol $\theta_{l,\diamondsuit}$ was used in equation (9.37) in order to represent the relative position of the past time point $t_{l+1} - \tau_1(t_{l+1}, y_{l+1,p+1}, c)$ in the interval $[t_{l'}, t_{l'+1}]$:

$$\theta_{l,\diamondsuit} = \frac{t_{l+1} - \tau_1(t_{l+1}, y_{l+1,p+1}, c) - t_{l'}}{h_{l'+1}}. \tag{9.39}$$

Eventually,

$$(f_{past})_{l+1}^{\diamondsuit} = f(t_{l'} + \theta_{l,\diamondsuit}h_{l'+1}, v_{l+1}^{\diamondsuit}, c, u_{l+1}^{\diamondsuit}) \tag{9.40}$$

is an abbreviating notation for an evaluation of the right-hand-side function at the past time point $(t_{past})_{l+1}^{\diamondsuit} := t_{l+1} - \tau_1(t_{l+1}, y_{l+1,p+1}, c)$. The argument $u_{l+1}^{\diamondsuit}$ in the right-hand-side function evaluation is thereby

$$u_{l+1}^{\diamondsuit} \approx y((t_{past})_{l+1}^{\diamondsuit} - \tau_1((t_{past})_{l+1}^{\diamondsuit}, v_{l+1}^{\diamondsuit}, c); c), \tag{9.41}$$

meaning that it is an approximation of a state at a time point even further in the past, compare Subsection 8.2.3. The computation of $u_{l+1}^{\diamondsuit}$ is thereby done depending on the value ξ_{past} of the discontinuity interval indicator in the past integration step $t_{l'} \to t_{l'+1}$.

It is further mentioned that equation (9.33) makes use of the abbreviating notations

$$\bar{\mathbf{W}}_{l+1}^{i} := \frac{\partial \bar{y}_{l+1}^{i}}{\partial c} \tag{9.42}$$

and

$$\left(\overline{\frac{\partial f}{\partial y}}\right)_{l+1}^{i} := \left.\frac{\partial f(t,y,c,v)}{\partial y}\right|_{(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c,\bar{v}_{l+1}^{i})} \tag{9.43a}$$

$$\left(\overline{\frac{\partial f}{\partial c}}\right)_{l+1}^{i} := \left.\frac{\partial f(t,y,c,v)}{\partial c}\right|_{(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c,\bar{v}_{l+1}^{i})} \tag{9.43b}$$

$$\left(\overline{\frac{\partial f}{\partial v}}\right)_{l+1}^{i} := \left.\frac{\partial f(t,y,c,v)}{\partial v}\right|_{(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c,\bar{v}_{l+1}^{i})}. \tag{9.43c}$$

The derivatives of the past states at the first $\mu-1$ stages of the quadrature rule, i.e. $d\bar{v}_{l+1}^{i}/dc$, are computed by

$$\frac{d\bar{v}_{l+1}^{i}}{dc} = \left(\overline{\frac{\partial\phi_\xi}{\partial c}}\right)_{l+1}^{i} - \left(\overline{\frac{d\phi_\xi}{dt}}\right)_{l+1}^{i}\left[\left(\overline{\frac{\partial\tau_1}{\partial y}}\right)_{l+1}^{i}\bar{\mathbf{W}}_{l+1}^{i} + \left(\overline{\frac{\partial\tau_1}{\partial c}}\right)_{l+1}^{i}\right] \text{ if } -n_s^{\phi} \leqslant \xi \leqslant 0 \tag{9.44}$$

and by

$$\begin{aligned}\frac{d\bar{v}_{l+1}^{i}}{dc} =& \mathbf{E}_{l'+1,p}(t_{l'}+\bar{\theta}_{l,i}h_{l'+1}) - (\overline{f_{past}})_{l+1}^{i}\cdot\left[\left(\overline{\frac{\partial\tau_1}{\partial y}}\right)_{l+1}^{i}\bar{\mathbf{W}}_{l+1}^{i} + \left(\overline{\frac{\partial\tau_1}{\partial c}}\right)_{l+1}^{i}\right] \\ & \text{if } 1 \leqslant \xi \leqslant n_s+1, \quad l' \leqslant l.\end{aligned} \tag{9.45}$$

Herein, the index l' depends of course on the stage i of the quadrature rule. Furthermore, the equations (9.44), (9.45) make use of the following abbreviations for the partial derivatives of the initial function and of the delay function τ_1:

$$\left(\overline{\frac{d\phi_\xi}{dt}}\right)_{l+1}^{i} := \left.\frac{d\phi_\xi(t,c)}{dt}\right|_{(\bar{t}_{l+1}^{i}-\tau_1(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c),c)} \tag{9.46a}$$

$$\left(\overline{\frac{\partial\phi_\xi}{\partial c}}\right)_{l+1}^{i} := \left.\frac{\partial\phi_\xi(t,c)}{\partial c}\right|_{(\bar{t}_{l+1}^{i}-\tau_1(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c),c)} \tag{9.46b}$$

$$\left(\overline{\frac{\partial\tau_1}{\partial y}}\right)_{l+1}^{i} := \left.\frac{\partial\tau_1(t,y,c)}{\partial y}\right|_{(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c)} \tag{9.46c}$$

$$\left(\overline{\frac{\partial\tau_1}{\partial c}}\right)_{l+1}^{i} := \left.\frac{\partial\tau_1(t,y,c)}{\partial c}\right|_{(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c)}. \tag{9.46d}$$

The symbol $\bar{\theta}_{l,i}$ denotes the relative position of the past time point $\bar{t}_{l+1}^{i}-\tau_1(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c)$ in the interval $[t_{l'}, t_{l'+1}]$

$$\bar{\theta}_{l,i} = \frac{\bar{t}_{l+1}^{i}-\tau_1(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c)-t_{l'}}{h_{l'+1}}. \tag{9.47}$$

Eventually,

$$(\overline{f_{past}})_{l+1}^{i} = f(t_{l'}+\bar{\theta}_{l,i}h_{l'+1},\bar{v}_{l+1}^{i},c,\bar{u}_{l+1}^{i}) \tag{9.48}$$

represents an evaluation of the right-hand-side function at the past time point $(\overline{t_{past}})_{l+1}^{i} := \bar{t}_{l+1}^{i} - \tau_1(\bar{t}_{l+1}^{i},\bar{y}_{l+1}^{i},c)$. The argument $\bar{u}_{l+1}^{i}$ in the right-hand-side function evaluation is thereby

$$\bar{u}_{l+1}^{i} \approx y((\overline{t_{past}})_{l+1}^{i} - \tau_1((\overline{t_{past}})_{l+1}^{i},\bar{v}_{l+1}^{i},c);c), \tag{9.49}$$

meaning that it is an approximation of a state at a time point even further in the past. The computation of $\bar{u}_{l+1}^{i}$ is thereby done depending on the value of the discontinuity interval indicator that was used for the past integration step $t_{l'} \to t_{l'+1}$, compare Subsection 8.2.3.

9.1.4. Equation Systems for Sensitivity Computation

In the previous subsections, Internal Numerical Differentiation for DDEs (Definition 8.2) was applied to the nonlinear equation systems that have to be solved in each integration step for the collocation method, for the implicit uniform correction, and for the implicit quadrature rule. As a result, the following was obtained

- For the collocation method, the equation system (9.4) was obtained, with $\mathbf{W}_{l+1}^{j}$ being given by equation (9.5) and dv_{l+1}^{j}/dc being given by either equation (9.9), by equation (9.10), or by equation (9.11). If the equations are inserted into each other, it is possible to obtain an equation system in which only $\mathbf{G}_{l+1}^{j}$ occur as derivative variables from the current step.

- For the implicit uniform correction, the equation system (9.19) was obtained, with $\mathbf{W}_{l+1}^{*}$ and $\dot{\mathbf{E}}_{l+1}^{q}(t)$ being given by equations (9.20a) and (9.20b), respectively, and with dv_{l+1}^{*}/dc being given by equation (9.24) or by equation (9.25). If the equations are inserted into each other, it is possible to obtain an equation system in which only $\mathbf{G}_{l+1}^{*}$ occurs as derivative variable from the current step.

- For the implicit quadrature formula, the equation system (9.32) was obtained, with $\mathbf{W}_{l+1,p+1}$ being given by equation (9.33) and $dv_{l+1}^{\diamondsuit}/dc$ being given either by equation (9.36) or by equation (9.37). If the equations are inserted into each other, it is possible to obtain an equation system in which only $\mathbf{G}_{l+1}^{\diamondsuit}$ occurs as derivative variable from the current step.

9.1.5. Properties of the Equation Systems

A first important observation is as follows: If the equation system for $\mathbf{G}_{l+1}^{j}$ is considered together with the corresponding equation system for the nominal solution for g_{l+1}^{j}, then the combined system is of dimension $\nu \times n_y \times (1 + n_c)$. Furthermore, this combined system is coupled, because the equations for $\mathbf{G}_{l+1}^{j}$ depend on g_{l+1}^{j}, e.g. through the evaluations of the partial derivatives of the right-hand-side function f. The resulting equation system is generally also nonlinear in g_{l+1}^{j}, again because of the evaluations of the partial derivatives of the right-hand-side function f

Approaching the task of sensitivity computation naively by applying a Newton or Newton-type method (see Section 6.5.2) to this potentially large nonlinear system is computationally expensive. It is pointed out that this is exactly what happens if the user of a standard DDE-solver decides to implement the variational equations manually.

Contrariwise, if the equation systems for the nominal solution are solved first, then the variables g_{l+1}^{j} can be regarded as fixed for the sensitivity computation. Accordingly, the three equation systems for the sensitivities contain only the quantities $\mathbf{G}_{l+1}^{j}$, $\mathbf{G}_{l+1}^{*}$, and $\mathbf{G}_{l+1}^{\diamondsuit}$ as unknowns. Furthermore, the equations for the sensitivities with respect to all n_c parameters decouple, such that only $n_c + 1$ systems of dimension $\nu \times n_y$ need to be solved.

A second important observation is that, in the decoupled case, the equation systems that need to be solved for sensitivity computation are linear in their unknowns $\mathbf{G}_{l+1}^{j}$, $\mathbf{G}_{l+1}^{*}$, and $\mathbf{G}_{l+1}^{\diamondsuit}$. This means that they can be represented as

$$\mathbf{A}\mathbf{G} = \mathbf{B}, \tag{9.50}$$

where $\mathbf{G}$ represents the unknowns for the respective system.

More precisely, for each of the three methods – collocation method, implicit uniform correction, and implicit quadrature rule – the matrices $\mathbf{A}$ are very similar to those derived in Subsection 6.5.4. The sole difference is that the time derivative of the past states is, in the Jacobians, approximated by the time derivative of the continuous representation, whereas for the computation of the sensitivities they are approximated by an evaluation of the right-hand-side function f in order to comply with the principle of IND.

9.1.6. Practical Solution of Equation Systems

The above findings suggest to use one of the following two approaches for an efficient computation of the unknowns $\mathbf{G}$:

- The first option is to exploit the fact that the equation systems are linear in their unknowns. Hence, the equation system can be solved by explicit construction and decomposition of the matrix $\mathbf{A}$. This approach is referred to *direct Internal Numerical Differentiation*.

- The second approach is to exploit the fact that an approximate inverse of the matrix $\mathbf{A}$ is available from the solution of the corresponding nominal equation system. Hence, a Newton-type method (see Definition 6.13) can be used to compute the unknowns. This approach is referred to *iterative Internal Numerical Differentiation*.

The terms "direct" and "iterative" Internal Numerical Differentiation have been established in the literature in the context of other integration methods and of other classes of differential equations, see e.g. Albersmeyer and Bock [3], Albersmeyer [2], and Beigel [23].

In Colsol-DDE, both direct and iterative Internal Numerical Differentiation are realized.

For direct IND, the matrix $\mathbf{A}$ is computed and decomposed by a singular value decomposition. After that, the unknowns $\mathbf{G}$ are computed by using the decomposition for the computation of $\mathbf{A}^{-1}\mathbf{B}$.

For iterative IND, the computation of the initial guesses for the unknowns is done completely analogous to the computation of initial guesses for the nominal solution, see Subsection 6.5.3. With regard to the termination criterion, iterative Internal Numerical Differentiation is realized in two variants, both of them being well-justified as follows.

On the one hand, the user may choose to do, for the computation of each directional derivative (i.e. for each column of $\mathbf{G}$), exactly as many iterations as were needed to solve the corresponding nominal equation system. The idea behind this is that the obtained sensitivities are, in some sense, closer to the "exact" derivative of the discrete mapping that is used for solving the nominal initial value problem, see Albersmeyer [2].

On the other hand, the user may choose to iterate until convergence is obtained, which may occur after a different number of iterations for each directional derivative. This approach is motivated by the fact that Colsol-DDE allows to compute error-controlled sensitivities (see Subsection 9.1.8 below). For this purpose, it is important that the equation systems are solved with sufficient accuracy so that the quantities that are used for error control have at least two valid digits.

Clearly, for the latter approach, a termination criterion for the computation of sensitivities is needed. The convergence criterion employed in Colsol-DDE for this purpose is constructed in complete analogy to the convergence criterion for the nominal solution, see Subsections 6.5.7 and 6.6.4. In particular, the termination criterion involves scaling factors for the sensitivities, which are computed analogous to Subsection 6.5.8.

9.1.7. Computation of Model Function Derivatives

An issue that has not yet been addressed is the practical computation of the partial derivatives of the right-hand-side function f, of the initial function ϕ (or smooth branches thereof), and of the delay functions τ_i, $1 \leqslant i \leqslant n_\tau$. Colsol-DDE provides several options for this purpose, which are exemplarily demonstrated for the computation of $(\partial f/\partial y)^j_{l+1}$:

- *One-Sided Finite Differences*: Here, the k-th column of the derivative $(\partial f/\partial y)^j_{l+1}$ is approximated by

$$\begin{aligned}\left(\frac{\partial f}{\partial y_k}\right)^j_{l+1} &= \left.\frac{\partial f(t,y,c,v)}{\partial y_k}\right|_{(t^j_{l+1},y^j_{l+1},c,v^j_{l+1})} \\ &= \frac{f(t^j_{l+1},y^j_{l+1}+\epsilon_k e_k,c,v^j_{l+1}) - f(t^j_{l+1},y^j_{l+1},c,v^j_{l+1})}{\epsilon_k}\end{aligned} \tag{9.51}$$

  where e_k is the k-th unit vector and ϵ_k is a variational parameter.

- *Two-Sided Finite Differences*: Here, the k-th column of the derivative $(\partial f/\partial y)^j_{l+1}$ is approximated by

$$\left(\frac{\partial f}{\partial y_k}\right)^j_{l+1} = \frac{f(t^j_{l+1},y^j_{l+1}+\frac{1}{2}\epsilon_k e_k,c,v^j_{l+1}) - f(t^j_{l+1},y^j_{l+1}-\frac{1}{2}\epsilon_k e_k,c,v^j_{l+1})}{\epsilon_k}. \tag{9.52}$$

- *Automatic Differentiation*: Colsol-DDE is designed to be used in conjuction with the Automatic Differentiation tool Tapenade [140, 141]. Tapenade provides, by source-to-source transformation, a code that computes with machine precision the matrix-vector product

$$\left(\frac{\partial f}{\partial y}\right)_{l+1}^{j} d$$

for any arbitrary $d \in \mathbb{R}^{n_y}$.

It is pointed out that none of the three approaches requires the user to bother about the computation of derivatives. Hence, Colsol-DDE is, in any case, derivative-free for the user.

The accuracy of both one-sided and two-sided finite differences depends on the choice of the variational parameter ϵ_k. In Colsol-DDE, this variational parameter is selected proportional to the internally computed scaling factor of the corresponding component of the state vector (see Subsection 6.5.8), in order to make the computation independent of a possibly inappropriate user scaling. However, the choice of the variational parameter is essentially heuristic, and even for an optimal choice the obtained derivative approximations have, in general, only half as many valid digits as the evaluations of the right-hand-side function itself.

It is therefore highly recommended to use Colsol-DDE in connection with Tapenade (or another source-to-source Automatic Differentiation tool that provides suitable interfaces for the derivative-computing routines). The obtained derivatives are then accurate to machine precision.

It is remarked that in several terms, e.g. in the first term on the right hand side of equation (9.4), only the product of the partial derivatives with other matrices is needed (e.g. with $\mathbf{W}_{l+1}^{j}$), but not necessarily the full $n_y \times n_y$ matrix $\left(\frac{\partial f}{\partial y}\right)_{l+1}^{j}$.

One approach is to ignore this fact and to compute the full matrix $\left(\frac{\partial f}{\partial y}\right)_{l+1}^{j}$ once. This has the advantage that the product $\left(\frac{\partial f}{\partial y}\right)_{l+1}^{j} \bar{W}_{l+1}^{j}$ can subsequently be computed by matrix-vector products for each derivative direction and for each iteration of the Newton-type method.

An alternative approach is to exploit the fact that only the product $\left(\frac{\partial f}{\partial y}\right)_{l+1}^{j} \bar{W}_{l+1}^{j}$ is needed. This can be done by calling the derivative routine provided by Tapenade for each derivative direction and in each iteration.

It is, in general, not possible to say which approach is more efficient, since this depends on the number of iterations, on the number of derivative directions, and on the cost of an evaluation of the derivative function in relation to the cost of a matrix-vector product. At present, Colsol-DDE always computes the full derivative matrices.

9.1.8. Error Control

In the Subsections 9.1.1-9.1.3, several equation systems were derived. These equation systems allowed to compute discrete approximations of the sensitivities which have discrete local errors of order p and $p+1$, and continuous approximations which have uniform local errors of order q and $q+1$. Hence, the techniques described in Section 6.6 can directly be transferred to construct error estimates for the sensitivities, and, based thereon, also a strategy for error control.

Colsol-DDE allows to control the local errors in both the nominal solution and in the sensitivities. Optionally, local error control for the sensitivities may also be disabled. In this case, for efficiency reasons, the implicit quadrature rule is not applied to the sensitivities because its results are not needed.

9.1.9. Incorporation of Sensitivity Computation into the Main Algorithm

The sensitivity computation is incorporated as follows into the main algorithm of Colsol-DDE (see Algorithm 6.20 in Section 6.8).

- The equation systems for the collocation step in the sensitivities are solved between step 4 and step 5. If direct IND or iterative IND with a fixed number of iterations is applied, the code always proceeds with step 5. If iterative IND is used until convergence is achieved, and n_{itmax}^{1} or more iterations are needed to obtain convergence, then recomputation of the

Jacobian matrix for the next integration step might be triggered. In the case that the method fails to converge, stepsize rejections may occur such that the code proceeds with step 2 of the main algorithm.

- The equation systems for the implicit uniform correction in the sensitivities are solved between step 7 and step 8. The above remarks regarding the use of direct and iterative IND apply accordingly.

- The equation systems for the implicit quadrature rule applied to the sensitivities are solved between step 10 and step 11. The above remarks regarding the use of direct and iterative IND apply accordingly.

9.1.10. Discontinuities of Order 0 in the Sensitivities

Colsol-DDE ensures that the time points of root discontinuities and the time points of propagated discontinuities are included into the mesh as described in Section 6.9.

In the time points of propagated discontinuities, it is checked whether the parent discontinuity is of order 0 in y. If yes, this may lead to a jump of order 0 in the Wronskian, and thus the expression (8.31) is evaluated in Colsol-DDE.

In the time points of root discontinuities, it must generally be expected that the sensitivities are discontinuous. At these points, Colsol-DDE computes the discrete analogue (see Subsection 8.2.9) of the jump in the Wronskian as it is given in Section 7.6.

9.1.11. Differentiability Checks

Motivated by the results of Chapter 7, Colsol-DDE uses numerical checks to ensure that the following conditions are fulfilled:

- Root discontinuities are "sufficiently far away" from other root discontinuities.

- Root discontinuities are "sufficiently far away" from children of critical discontinuities.

- Children of discontinuities of order 0 in y are "sufficiently far away" from other children of discontinuities of order 0 in y.

- Root discontinuities and children of discontinuities of order 0 do not occur at the initial time or at the final time.

- Switching functions and propagation switching functions have, at their zeros, a "sufficiently non-zero" time derivative.

In all cases, the meaning of "suffciently far away" and "sufficiently non-zero" depends on user-given input parameters.

9.2. Practical Computation of Adjoint Sensitivities

9.2.1. Simplified Case: No Discontinuities in the Sensitivities

In Section 8.3 the discrete adjoint scheme of a CRK method applied to DDE-IVPs was defined (Definition 8.5), and it was presented how the discrete adjoints can be used for the computation of the sensitivities, see equation (8.49).

For the practical realization in Colsol-DDE, an additional aspect must be taken into account in order to obtain a scheme that is the "exact" discrete adjoint of the method for forward sensitivity computation. This additional aspect is that continuous representations of different uniform local order are used in the collocation method depending on whether or not overlapping occurs.

Analogously to Section 8.3, the case of a single delay function is considered. Further, it is first assumed that all propagated discontinuities are at least of order 1 in the Wronskian, such that no jump expressions need to be taken into account.

In this setting, consider the following variational ansatz as a modification to equation (8.45):

$$
\begin{aligned}
0 = \sum_{l=0}^{n_m-1} \Bigg\{ & \Lambda_{l+1} \left[-\mathbf{W}_{l+1} + \mathbf{W}_l + h_{l+1} \sum_{j=1}^{\nu} \beta_j \mathbf{G}_{l+1}^j \right] \\
& + h_{l+1} \left(\sum_{j=1}^{\nu} \Lambda_l^j \left[-\mathbf{G}_{l+1}^j + \left(\frac{\partial f}{\partial y} \right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial f}{\partial c} \right)_{l+1}^j + \left(\frac{\partial f}{\partial v} \right)_{l+1}^j \frac{dv_{l+1}^j}{dc} \right] \right) \\
& + h_{l+1} \left(\Lambda_l^* \left[-\mathbf{G}_{l+1}^* + \left(\frac{\partial f}{\partial y} \right)_{l+1}^* \mathbf{W}_{l+1}^* + \left(\frac{\partial f}{\partial c} \right)_{l+1}^* + \left(\frac{\partial f}{\partial v} \right)_{l+1}^* \frac{dv_{l+1}^*}{dc} - \sum_{i=1}^{\nu} \dot{b}_i(\theta^*) \mathbf{G}_{l+1}^i \right] \right) \\
& + h_{l+1} \left(\sum_{j=1}^{\nu} \Gamma_l^j \left[\mathbf{W}_{l+1}^j - \mathbf{W}_l - h_{l+1} \sum_{i=1}^{\nu} a_{j,i} \mathbf{G}_{l+1}^i \right] \right) \\
& + h_{l+1} \left(\Gamma_l^* \left[\mathbf{W}_{l+1}^* - \mathbf{W}_l - h_{l+1} \sum_{i=1}^{\nu} b_i(\theta^*) \mathbf{G}_{l+1}^i - h_{l+1} b_*(\theta^*) \mathbf{G}_{l+1}^* \right] \right) \\
& + h_{l+1} \left(\sum_{j=1}^{\nu} \Pi_l^j H_{l+1} \left[-\frac{dv_{l+1}^j}{dc} + \mathbf{W}_{l'} + h_{l'+1} \sum_{i=1}^{\nu} b_i(\theta_{l,j}) \mathbf{G}_{l'+1}^i + h_{l'+1} \Theta_{l,j} b_*(\theta_{l,j}) \mathbf{G}_{l'+1}^* \right. \right. \\
& \qquad \left. \left. - (f_{past})_{l+1}^j \left(\left(\frac{\partial \tau_1}{\partial y} \right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial \tau_1}{\partial c} \right)_{l+1}^j \right) \right] \right) \\
& + h_{l+1} \left(\Pi_l^* H_{l+1} \left[-\frac{dv_{l+1}^*}{dc} + \mathbf{W}_{l'} + h_{l'+1} \sum_{i=1}^{\nu} b_i(\theta_{l,*}) \mathbf{G}_{l'+1}^i + h_{l'+1} b_*(\theta_{l,*}) \mathbf{G}_{l'+1}^* \right. \right. \\
& \qquad \left. \left. - (f_{past})_{l+1}^* \left(\left(\frac{\partial \tau_1}{\partial y} \right)_{l+1}^* \mathbf{W}_{l+1}^* + \left(\frac{\partial \tau_1}{\partial c} \right)_{l+1}^* \right) \right] \right) \\
& + h_{l+1} \left(\sum_{j=1}^{\nu} \Pi_l^j (1 - H_{l+1}) \left[-\frac{dv_{l+1}^j}{dc} + \left(\frac{\partial \phi_{\xi[l+1]}}{\partial c} \right)_{l+1}^j \right. \right. \\
& \qquad \left. \left. - \left(\frac{d\phi_{\xi[l+1]}}{dt} \right)_{l+1}^j \left(\left(\frac{\partial \tau_1}{\partial y} \right)_{l+1}^j \mathbf{W}_{l+1}^j + \left(\frac{\partial \tau_1}{\partial c} \right)_{l+1}^j \right) \right] \right) \\
& + h_{l+1} \left(\sum_{j=1}^{\nu} \Pi_l^* (1 - H_{l+1}) \left[-\frac{dv_{l+1}^*}{dc} + \left(\frac{\partial \phi_{\xi[l+1]}}{\partial c} \right)_{l+1}^* \right. \right. \\
& \qquad \left. \left. - \left(\frac{d\phi_{\xi[l+1]}}{dt} \right)_{l+1}^* \left(\left(\frac{\partial \tau_1}{\partial y} \right)_{l+1}^* \mathbf{W}_{l+1}^* + \left(\frac{\partial \tau_1}{\partial c} \right)_{l+1}^* \right) \right] \right) \Bigg\}.
\end{aligned}
\tag{9.53}
$$

In this equation, the following notations have been used in agreement with Sections 6.4 and 8.3:

- $\xi[l+1]$ is the value of the numerically determined discontinuity interval indicator in the step $t_l \to t_{l+1}$.
- $H_{l+1} = 0$ if the discontinuity interval indicator is less than or equal to 0, i.e. the past states in integration step $t_l \to t_{l+1}$ are computed from a smooth branch of the initial function. Otherwise, if the discontinuity interval indicator is greater or equal 1, indicating that the past states are computed from the continuous representation in a past or in the current integration step, then $H_{l+1} = 1$.
- $\Theta_{l,j} = 0$ indicates that overlapping occurs for the j-th stage in integration step $t_l \to t_{l+1}$, whereas $\Theta_{l,j} = 1$ indicates that overlapping does not occur.
- Λ_{l+1}, Λ_l^j , Γ_l^j, and Π_l^j are the adjoint variables for the collocation method, and Λ_l^*, Γ_l^*, and Π_l^* are the adjoint variables for the implicit uniform correction.

Note that the implicit quadrature rule does not need to be taken into account, because its results are only used for error estimation.

The goal is exactly the same as in Section 8.3, i.e. the product $\Lambda_{n_m} \mathbf{W}_{n_m}$ should be computed without computing the variables $\mathbf{W}_l$, $\mathbf{W}_{l+1}^j$, $\mathbf{G}_{l+1}^j$, and dv_{l+1}^j/dc. Furthermore, also the procedure

to reach this goal is the same as in Section 8.3, i.e. the terms are sorted in such a way that $\mathbf{W}_l$, $\mathbf{W}_{l+1}^j$ $\mathbf{G}_{l+1}^j$, and dv_{l+1}^j/dc can be factored out.

In order to state the result, the following is defined (compare, once again, to Section 8.3):

- Let M_l for $l \geqslant 0$ contain the set of all those pairs of indices μ and ρ such that $l'(\mu+1,\rho) = l$, i.e. it contains the indices of those steps and stages for which the past state is computed by an evaluation of the continuous representation in step $t_l \to t_{l+1}$.
- Let M_l^* for $l \geqslant 0$ contain those indices μ for which the past state of the implicit uniform correction is computed by an evaluation of the continuous representation in step $t_l \to t_{l+1}$.
- Let M_l for $l < 0$ contain the set of those indices μ for which $\xi[\mu+1] = l+1$, i.e. the discontinuity interval indicator of step $t_\mu \to t_{\mu+1}$ points to a smooth branch of ϕ.
- Let M_l^* for $l < 0$ contain the set of those indices μ for which $\xi[\mu+1] = l+1$, i.e. the discontinuity interval indicator of step $t_\mu \to t_{\mu+1}$ points to a smooth branch of ϕ.

By using these notations, the following is obtained as a generalization of the discrete adjoint scheme (Definition 8.5):

Definition 9.1 (Discrete Adjoint Scheme for the Methods used in Colsol-DDE)
The equations

$$\Lambda_l = \Lambda_{l+1} - h_{l+1}\sum_{j=1}^{\nu}\Gamma_l^j - h_{l+1}\Gamma_l^* + \sum_{(\mu,\rho)\in M_l} h_{\mu+1}\Pi_\mu^\rho + \sum_{\mu\in M_l^*} h_{\mu+1}\Pi_\mu^* \tag{9.54a}$$

$$\Lambda_l^j = \Lambda_{l+1}\beta_j - h_{l+1}\sum_{i=1}^{\nu} a_{i,j}\Gamma_l^i - h_{l+1}b_j(\theta^*)\Gamma_l^* + \sum_{(\mu,\rho)\in M_l} h_{\mu+1}\Pi_\mu^\rho b_j(\theta_{\mu,\rho}) + \sum_{\mu\in M_l^*} h_{\mu+1}\Pi_\mu^* b_j(\theta_{\mu,*}) - \dot{b}_j(\theta^*)\Lambda_l^* \tag{9.54b}$$

$$\Lambda_l^* = -h_{l+1}b_*(\theta^*)\Gamma_l^* + \sum_{(\mu,\rho)\in M_l} h_{\mu+1}\Pi_\mu^\rho\Theta_{\mu,\rho}b_*(\theta_{\mu,\rho}) + \sum_{\mu\in M_l^*} h_{\mu+1}\Pi_\mu^* b_*(\theta_{\mu,*}) \tag{9.54c}$$

$$\Gamma_l^j = -\Lambda_l^j\left(\frac{\partial f}{\partial y}\right)_{l+1}^j + \Pi_l^j\left(H_{l+1}(f_{past})_{l+1}^j + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\right)\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^j \tag{9.54d}$$

$$\Gamma_l^* = -\Lambda_l^*\left(\frac{\partial f}{\partial y}\right)_{l+1}^* + \Pi_l^*\left(H_{l+1}(f_{past})_{l+1}^* + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^*\right)\left(\frac{\partial\tau_1}{\partial y}\right)_{l+1}^* \tag{9.54e}$$

$$\Pi_l^j = \Lambda_l^j\left(\frac{\partial f}{\partial v}\right)_{l+1}^j \tag{9.54f}$$

$$\Pi_l^* = \Lambda_l^*\left(\frac{\partial f}{\partial v}\right)_{l+1}^* . \tag{9.54g}$$

constitute the discrete adjoint scheme of the method for forward sensitivity computation used in Colsol-DDE*.*

By using this discrete adjoint scheme, it is possible to show that the sought sensitivities can be computed by:

$$\begin{aligned}\Lambda_{n_m}\mathbf{W}_{n_m} = {} & +\Lambda_0\mathbf{W}_0 + \sum_{l=0}^{n_m-1}\Bigg[h_{l+1}\sum_{j=1}^{\nu}\Lambda_l^j\left(\frac{\partial f}{\partial c}\right)_{l+1}^j + h_{l+1}\Lambda_l^*\left(\frac{\partial f}{\partial c}\right)_{l+1}^* \\ & - h_{l+1}\sum_{j=1}^{\nu}\Pi_l^j\left(H_{l+1}(f_{past})_{l+1}^j + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^j\right)\left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^j \\ & - h_{l+1}\sum_{j=1}^{\nu}\Pi_l^*\left(H_{l+1}(f_{past})_{l+1}^* + (1-H_{l+1})\left(\frac{d\phi_{\xi[l+1]}}{dt}\right)_{l+1}^*\right)\left(\frac{\partial\tau_1}{\partial c}\right)_{l+1}^*\Bigg] \\ & + \sum_{l=-n_s-1}^{-1}\sum_{(\mu,\rho)\in M_l} h_{\mu+1}\Pi_\mu^\rho\left(\frac{\partial\phi_l}{\partial c}\right)_{\mu+1}^\rho + \sum_{l=-n_s-1}^{-1}\sum_{\mu\in M_l^*} h_{\mu+1}\Pi_\mu^\rho\left(\frac{\partial\phi_l}{\partial c}\right)_{\mu+1}^* .\end{aligned} \tag{9.55}$$

Colsol-DDE features a practical implementation of the discrete adjoint scheme (9.54) and computes adjoint sensitivities by means of equation (9.55).

A subtle but important aspect is that the Lobatto IIIA method is explicit in its first stage. Thus, it plays a role whether overlapping occured for the last stage of the previous integration step. For notational simplicity, the effect of this explicit stage for the adjoint sensitivity computation has not been taken into account in the equations above.

9.2.2. Practical Computation of the Discrete Adjoint Scheme

Inserting the equations (9.54e) and (9.54g) into equation (9.54c), and inserting the equations (9.54d) and (9.54f) into equation (9.54b) yields equation systems of the form

$$\Lambda \mathbf{A} = \tilde{\mathbf{B}}. \tag{9.56}$$

This means that the equations are linear in the unknowns $(\Lambda_l^1, \ldots, \Lambda_l^\nu)$ and λ_l^*, respectively. Some further properties that are theoretically satisfying and useful for the numerical computation are as follows:

- In the equation system for Λ_l^*, the matrices $\mathbf{A}$ and B are independent of all Λ_l^j, i.e. the adjoint variables of the collocation method of the same step. In other words, the two equation systems for λ_l^* and Λ_l^j decouple, and Λ_l^* can be computed by solving an $n_y \times n_y$ dimensional linear system. This decoupling is a direct correspondence of the fact that the computation of the step in the collocation method is independent of the implicit uniform correction (which is applied subsequently in the case of forward sensitivity computation).
- The system for λ_l^j is, generally, of dimension $\nu \cdot n_y \times \nu \cdot n_y$. However, the Lobatto IIIA method is, in the forward mode, explicit in its first stage, i.e. $\mathbf{W}_{l+1}^1 = \mathbf{W}_l$. This has a correspondence in the adjoint mode as well. More precisely, for the three-stage Lobatto IIIA method implemented in Colsol-DDE, Λ_l^2 and Λ_l^3 can be determined by solving a $2n_y \times 2n_y$ system, and Λ_l^1 can subsequently be computed explicitly.
- The matrices $\mathbf{A}$ that occur in the schematic equation (9.56) for the collocation method and for the implicit uniform correction are the same as the matrices $\mathbf{A}$ that occur in the forward mode, see the schematic equation (9.50). The matrix-valued right hand sides of these equations are, however, different.

For the practical solution, it is therefore possible to use the same approaches as those discussed in Subsection 9.1.6, i.e. direct solution of the linear equation systems by matrix decomposition, or iterative solution with an approximate inverse. These approaches can be called *direct adjoint Internal Numerical Differentiation* and *iterative adjoint Internal Numerical Differentiation*. Moreover, if the matrices that were used in the forward nominal solution are stored, they can be reused such that for each integration step the same number of iterations are performed with the same approximate inverse.

At present, however, only the direct adjoint Internal Numerical Differentiation is implemented in Colsol-DDE.

9.2.3. Discontinuities in the Sensitivities and Generalization to IHDDE-IVPs

If there are discontinuities of order 0 in the initial function ϕ of a DDE-IVP, then the Wronskian generally exhibits jumps at the time points of the child discontinuities. These jumps are taken into account in Colsol-DDE as described in Subsection 8.3.4.

Sensitivity computation for IHDDE-IVP solutions requires to take into account a jump in the adjoint sensitivities at the time points of the root discontinuities. Furthermore, dependencies of the switching functions and impulse function on past states lead to additional terms in the discrete adjoint scheme and in the expression that is used for sensitivity computation.

The necessary modifications for CRK methods were described in Subsection 8.3.7. Since Colsol-DDE uses, for forward sensitivity computation, the continuous representation of order p, also a modification in equation (9.54c) for the adjoint variable Λ_l^* of the uniform order correction becomes necessary. This modification is implemented in Colsol-DDE.

Contributions that arise from children of root discontinuities of order 0 in y are also taken into account.

Part IV.

Parameter Estimation

10. Problem Formulation and Theory

> *Sampling experiments (...) have shown, however, that the maximum likelihood method produces acceptable estimates in many situations. Whereas better methods may be available for specific cases, a powerful argument for the use of the maximum likelihood method is the generality and relative ease of application.*
>
> Bard, in his book "Nonlinear Parameter Estimation" [19], motivating the use of maximum likelihood estimation.

A typical situation in the natural and engineering sciences is as follows: On the one hand, there is a real-world dynamic process, and on the other hand there is a mathematical object $y(t; c)$ that describes the state of the process as a function of the time t and of parameters c.

In practice, it is frequently the case that the parameters c cannot be derived from "first principles". Instead, it is necessary to estimate the parameters from measurement data. The formulation of a mathematical problem whose solution provides – in some sense – a "good" estimate of the parameters is the main topic of this chapter.

Organization of This Chapter

Section 10.1 introduces the notions of a dynamic model and a measurement model for an observed process, and makes elementary assumptions for these models. Section 10.2 deals with the issue that practically obtained measurement data are almost always random numbers. Furthermore, the section introduces the likelihood function. Section 10.3 recalls that a maximizer of the likelihood function – a so-called maximum likelihood estimate – is, in the special case of normally distributed measurement errors, obtained by solving an optimization problem with a least-squares objective function. Eventually, Section 10.4 gives the necessary and sufficient conditions for solutions of optimization problems.

10.1. Models and Assumptions

10.1.1. Dynamic Model

Consider the situation that a real-world process is mathematically described by a function $y(t; c)$ of time $t \in [t^a, t^b]$ and of parameters $c \in \mathbb{R}^{n_c}$. The function $y(t; c)$ is typically defined by a *dynamic model*, e.g. by a system of algebraic equations, differential equations, or both. Depending on the knowledge and insight that went into the construction of the dynamic model, the function $y(t; c)$ may be more or less suitable for the description of the real behavior of the system. For this part of the thesis, the fundamental assumption is made that the model is *correct*. This means that there exist parameters $c^\star$, called the *correct parameters*, such that the mathematical object $y(t; c^\star)$ describes the state of the real-world process completely and correctly.

In this thesis, the focus lies on the estimation of parameters in functions $y(t; c)$ that solve impulsive hybrid discrete-continuous delay differential equations (IHDDEs). IHDDEs as dynamic model equations are considered in Chapter 13. In this chapter – and also in Chapters 11 and 12 – the concrete shape of the dynamic model is irrelevant, and it is assumed that the function $y(t; c)$ is available.

10.1.2. Measurement Model(s)

Even in the fortunate case that the dynamic model is correct, some or all of the correct parameters $c^\star$ of the system might be unknown. However, there is typically some device that allows to observe the dynamic process. From this device, or several such devices, a number of measurements

η_i, $1 \leqslant i \leqslant n_h$, is obtained. If a function h_i is available that describes the measurement process, then the measurement η_i can be expressed as

$$\eta_i = h_i(\{y(t_j; c^\star)\}_{j=1}^{n_t}, c^\star) + \epsilon_i. \tag{10.1}$$

The function $h_i : \mathbb{R}^{n_y \times n_t} \times \mathbb{R}^{n_c} \to \mathbb{R}$ is called *measurement model* or *measurement function*. The function h_i is allowed to depend on the system states $y(t_j; c)$ at n_t time points and on the parameters c. The mismatch between the measurements and the evaluation of the measurement function is denoted by ϵ_i.

The dependencies of the measurement function in equation (10.1) are quite general. For many applications, it is sufficient to consider measurement functions h_i that depend only on the state at a single measurement time t_j. More precisely, this means that for each η_i there exists an index j, $1 \leqslant j \leqslant n_t$, such that

$$\eta_i = h_i(y(t_j; c^\star), c^\star) + \epsilon_i. \tag{10.2}$$

It is assumed throughout this part of the thesis that the measurement functions h_i are *correct models* for the observation process. In this case, the quantity ϵ_i in equation (10.1) can be interpreted as the *measurement error* for the i-th observation.

10.1.3. Goal of Parameter Estimation

The goal of parameter estimation is to use the measurement data η_i, $1 \leqslant i \leqslant n_h$, to obtain a "good" estimate $\hat{c}$ of the (unknown) correct parameters $c^\star$.

10.2. Random Measurements

Measurement errors are, in practice, *random*. This has to be understood as follows. If two *experiments* are conducted, for which the dynamic process behaves identically, and in which the same devices are used for taking measurements in exactly same way, then this yields two sets of measurement data $\eta = (\eta_1, \ldots, \eta_{n_h})^T$ and $\tilde{\eta} = (\tilde{\eta}_1, \ldots, \tilde{\eta}_{n_h})^T$. These sets of measurement data correspond to two sets of measurement errors $\epsilon = (\epsilon_1, \ldots, \epsilon_{n_h})^T$ and $\tilde{\epsilon}_i = (\tilde{\epsilon}_1, \ldots, \tilde{\epsilon}_{n_h})^T$.

The errors are thereby such that $\tilde{\epsilon} \neq \epsilon$ (componentwise), and thus also $\tilde{\eta} \neq \eta$, even though the observed process is exactly the same as before. Formally, ϵ_i is called a *specific realization* of the *random number* $\mathfrak{e}_i$.

It is assumed at this point that the reader is familiar with some elementary concepts of *probability theory*, in particular with *probability*, *continuous random variables*, and *independent random variables*.

The next definition introduces the notion of a *probability density function*.

Definition 10.1 (Probability Density Function)

A function $p : \mathbb{R} \to \mathbb{R}$ is called a probability density function *if it has the following properties:*

$$\int_{-\infty}^{\infty} p(x)dx = 1 \tag{10.3a}$$

$$p(x) \geqslant 0 \quad \textit{for all} \quad x \in \mathbb{R}. \tag{10.3b}$$

By assigning, to a continuous random variable $\mathfrak{x}$, the probability density function p_x, the probability for $\mathfrak{x}$ to assume a value in the interval $[x_a, x_b]$ is given by

$$P(\mathfrak{x} \in [x_a, x_b]) = \int_{x_a}^{x_b} p_x(x)dx. \tag{10.4}$$

With regard to the measurement errors ϵ_i, it is assumed that they are realizations of continuous random variables $\mathfrak{e}_i$. To these continuous random variables, probability density functions $p_i : \mathbb{R} \to \mathbb{R}$, $1 \leqslant i \leqslant n_h$, are assigned.

In the following, let $p_{all} : \mathbb{R}^{n_h} \to \mathbb{R}$ be the function that gives, for any $\epsilon \in \mathbb{R}^{n_h}$, the probability density that corresponds to the specific combination ϵ of measurement errors. This function is called

joint probability density function. Some elementary results for the function p_{all} are as follows (cf. Bard [19], page 23):

- *Independent measurement errors*: If the measurement error $\mathfrak{e}_i$ is independent of the measurement error $\mathfrak{e}_j$ for $i \neq j$, then the joint probability density function $p_{all}(\epsilon)$ can be expressed as

$$p_{all}(\epsilon) = \prod_{i=1}^{n_h} p_i(\epsilon_i). \tag{10.5}$$

- *Normally distributed measurement errors*: If the measurement errors $\mathfrak{e}$ are normally distributed with mean 0 and regular covariance matrix $\mathbf{V}_\epsilon$, shortly denoted by $\mathfrak{e} \sim \mathcal{N}(0, \mathbf{V}_\epsilon)$, then the joint probability density function is

$$p_{all}(\epsilon) = (2\pi)^{-\frac{n_h}{2}} (\det \mathbf{V}_\epsilon)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\epsilon^T \mathbf{V}_\epsilon^{-1} \epsilon\right). \tag{10.6}$$

If the measurement errors are independent and normally distributed, $\mathfrak{e}_i \sim \mathcal{N}(0, \sigma_i^2)$, $\sigma_i \in (0, \infty)$, then it follows immediately that

$$p_{all}(\epsilon) = (2\pi)^{-\frac{n_h}{2}} \prod_{i=1}^{n_h} (\sigma_i)^{-1} \exp\left(-\frac{1}{2}\frac{\epsilon_i^2}{\sigma_i^2}\right). \tag{10.7}$$

Consider the situation that measurements η are available, which correspond to one specific realization ϵ of the vector-valued random variable $\mathfrak{e}$. It is possible to define, for any values of the parameters c, the *residuals*

$$r(c; \eta) = \eta - h(\{y(t_j; c)\}_{j=1}^{n_t}, c), \tag{10.8}$$

where h is a vector-valued function whose components are the measurement functions h_i. It is thereby clear that $r(c^\star; \eta) = \epsilon$, i.e. the residuals for the true parameters $c^\star$ are the measurement errors.

The residuals are used in the following definition.

Definition 10.2 (Likelihood Function)

By replacing, in the joint probability distribution function p_{all}, the measurement errors ϵ by the residuals $r(c; \eta)$, the likelihood function *is obtained.*

For example, in the case that the measurement errors are assumed to be normally distributed with covariance matrix $\mathbf{V}_\epsilon$, the joint probability density function is given by equation (10.6). Consequently, the likelihood function is defined as

$$\begin{aligned}\Lambda(c; \eta, \mathbf{V}_\epsilon) :=& p(r(c; \eta)) \\ =& (2\pi)^{-\frac{n_h}{2}} (\det \mathbf{V}_\epsilon)^{-\frac{1}{2}} \exp\left(-\frac{1}{2} r^T(c; \eta) \mathbf{V}_\epsilon^{-1} r(c; \eta)\right).\end{aligned} \tag{10.9}$$

The integral of the likelihood function defined in equation (10.9) over some volume in $\mathbb{R}^{n_c}$ can be interpreted as the probability that parameters in that volume yield a particular set η of measurement data, provided that the measurement data are assumed to be normally distributed with covariance matrix $\mathbf{V}_\epsilon$.

10.3. Maximum Likelihood Estimation

The idea of maximum likelihood estimation is to find those parameters c for which the likelihood function assumes its maximum value. This *maximum likelihood estimate* is denoted by $\hat{c}$. For likelihood functions that are differentiable with respect to c (e.g. in the case of a normal distribution) it follows from standard analysis that a maximum likelihood estimate has to fulfill the following

necessary condition:

$$\left.\frac{d\Lambda(c;\eta,\mathbf{V}_\epsilon)}{dc}\right|_{c=\hat{c}} = 0. \tag{10.10}$$

A maximum $\hat{c}$ of the function $\Lambda(c;\eta,\mathbf{V}_\epsilon)$ is also a maximum of the function $\ln(\Lambda(c;\eta,\mathbf{V}_\epsilon))$, because the logarithm is a strictly increasing function. It is therefore possible to reformulate the above necessary condition as

$$\left.\frac{d\ln(\Lambda(c;\eta,\mathbf{V}_\epsilon))}{dc}\right|_{c=\hat{c}} = 0. \tag{10.11}$$

A very useful result for practical maximum likelihood parameter estimation is formulated in the following theorem.

Theorem 10.3 (Maximum Likelihood Estimation for Normally Distributed Errors)

If the measurement errors $\mathfrak{e}$ are normally distributed with mean zero and a regular covariance matrix $\mathbf{V}_\epsilon$, i.e. $\mathfrak{e} \sim \mathcal{N}(0,\mathbf{V}_\epsilon)$, then maximization of the likelihood function $\Lambda(c;\eta,\mathbf{V}_\epsilon)$ is equivalent to minimization of the function

$$\varphi(c) := r^T(c;\eta)\mathbf{V}_\epsilon^{-1}r(c;\eta). \tag{10.12}$$

Herein, the implicit dependencies of φ on the data η and on the covariance matrix $\mathbf{V}_\epsilon$ of the data are suppressed.

If the measurement errors are, in addition, independent, i.e. $\mathfrak{e}_i \sim \mathcal{N}(0,\sigma_i^2)$, then the covariance matrix $\mathbf{V}_\epsilon$ is a diagonal matrix with entries $\sigma_1^2,\ldots,\sigma_{n_h}^2$, and the function $\varphi(c)$ simplifies to

$$\varphi(c) = \sum_{i=1}^{n_h} \frac{(r_i(c;\eta))^2}{\sigma_i^2}. \tag{10.13}$$

Proof
See Bard [19], page 63. ■

Covariance matrices are real-valued and symmetric. If they are, in addition, assumed to be regular (see Theorem 10.3), then there exists a matrix $\mathbf{V}_\epsilon^{\frac{1}{2}}$ such that $\mathbf{V}_\epsilon^{\frac{1}{2}}\mathbf{V}_\epsilon^{\frac{1}{2}} = \mathbf{V}_\epsilon$. Furthermore, also $\mathbf{V}_\epsilon^{\frac{1}{2}}$ is regular, and its inverse $\mathbf{V}_\epsilon^{-\frac{1}{2}}$ is such that $\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathbf{V}_\epsilon^{-\frac{1}{2}} = \mathbf{V}_\epsilon^{-1}$. Therefore, by defining $\bar{r}(c;\eta) := \mathbf{V}_\epsilon^{-\frac{1}{2}}r(c;\eta)$, the function $\varphi(c)$ can be expressed as

$$\varphi(c) = \|\bar{r}(c;\eta)\|_2^2 = \|\mathbf{V}_\epsilon^{-\frac{1}{2}}r(c;\eta)\|_2^2 = \|\mathbf{V}_\epsilon^{-\frac{1}{2}}\left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t},c)\right)\|_2^2. \tag{10.14}$$

This means that $\varphi(c)$ can be expressed as a sum of squares for any arbitrary but regular covariance matrix $\mathbf{V}_\epsilon$. Hence, maximum likelihood estimation is, under the conditions of Theorem 10.3, equivalent to minimizing the differences between the measurements and the evaluations of the measurement functions in a weighted Euclidean norm.

In general, the minimization of the function $\|\varphi(c)\|_2^2$ may be subject to additional equality constraints. It is natural to allow that the equality constraint functions g_i, $1 \leqslant i \leqslant n_g$, have the same dependencies as the measurement function h_i. This leads to the following *optimization problem*:

$$\min_c \|\mathbf{V}_\epsilon^{-\frac{1}{2}}\left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t},c)\right)\|_2^2 \tag{10.15a}$$

$$\text{subject to } g(\{y(t_j;c)\}_{j=1}^{n_t},c) = 0. \tag{10.15b}$$

Herein, g is a vector-valued functions whose components are given by $g_i(\{y(t_j;c)\}_{j=1}^{n_t},c)$.

Equality constraints of the form (10.15b) may be appropriate to formalize "expert knowledge" that a modeler has about a process, e.g., if a process is known to be periodic or if there are conservation laws that have to be fulfilled. There further exist applications in which some measurements are known to be highly accurate compared to other measurements. In such a case, it may be numerically favorable to define equality constraints rather than least-squares terms.

For compactness and consistency of notation, let now $F_1 : \mathbb{R}^{n_c} \to \mathbb{R}^{n_h}$ be a function whose

components are given by

$$F_{1,i}(c) := \bar{r}_i(c;\eta) \tag{10.16}$$

and let $F_2 : \mathbb{R}^{n_c} \to \mathbb{R}^{n_g}$ be a function whose components are given by

$$F_{2,i}(c) := g_i(\{y(t_j;c)\}_{j=1}^{n_t}, c) \quad \text{for} \quad 1 \leqslant i \leqslant n_{F_2}. \tag{10.17}$$

Further, define $n_{F_1} := n_h$ and $n_{F_2} := n_g$.

With these notations, the problem (10.15) can be stated in a compact form, as it is done in the following definition.

Definition 10.4 (Nonlinear Constrained Least-Squares Problem)

A minimization problem of the form

$$\min_c \|F_1(c)\|_2^2 \tag{10.18a}$$

$$\text{subject to} \quad F_2(c) = 0, \tag{10.18b}$$

with functions $F_i : \mathbb{R}^{n_c} \to \mathbb{R}^{n_{F_i}}$, as it arises in constrained maximum likelihood parameter estimation for normally distributed measurement errors with known covariance matrix, is called a nonlinear constrained least-squares problem*.*

If the functions F_1 and F_2 are differentiable with respect to c, then their derivatives are denoted by

$$\mathbf{J}_i(c) := \left.\frac{dF_i(c')}{dc'}\right|_{c'=c}, \quad \text{for} \quad i = 1, 2, \tag{10.19}$$

and are called the *Jacobian matrices*.

The following definitions are helpful in order to establish an understanding of a solution of the nonlinear constrained least-squares problem (10.18).

Definition 10.5 (Feasible Point)

A point $c \in \mathbb{R}^{n_c}$ is called a feasible point *if the equality constraints are satisfied:*

$$F_2(c) = 0. \tag{10.20}$$

Definition 10.6 (Feasible Set)

The set that contains all feasible points,

$$\mathcal{F} := \{c \in \mathbb{R}^{n_c} \mid F_2(c) = 0\} \tag{10.21}$$

is called the feasible set*.*

A *(strict) global solution* of the constrained nonlinear least-squares problem (10.18) is defined as follows.

Definition 10.7 (Global Solution, Strict Global Solution)

Let $\hat{c}$ be a feasible point such that $\|F_1(c)\| \geqslant \|F_1(\hat{c})\|$ for all $c \in \mathcal{F}$. Then $\hat{c}$ is called a global solution*. If the inequality is strict, $\|F_1(c)\| > \|F_1(\hat{c})\|$, then $\hat{c}$ is called a* strict global solution*.*

Contrariwise, *(strict) local solutions* fulfill the inequalities only in a neighborhood of $\hat{c}$.

Definition 10.8 (Local Solution, Strict Local Solution)

Let $\hat{c}$ be a feasible point and let $\mathcal{U}^c$ be a neighborhood of $\hat{c}$. If $\|F_1(c)\| \geqslant \|F_1(\hat{c})\|$ holds for all $c \in \mathcal{F} \cap \mathcal{U}^c$, then $\hat{c}$ is called a local solution*. If the inequality is strict, $\|F_1(c)\| > \|F_1(\hat{c})\|$ for all $c \in \mathcal{F} \cap \mathcal{U}^c$, then $\hat{c}$ is called a* strict local solution*.*

In the context of parameter estimation problems it of course typical $n_{F_2} < n_c$, because $n_{F_2} = n_c$ may lead to a situation in which only one point $\hat{c}$ in $\mathbb{R}^{n_c}$ fulfills all equality constraints. Hence, there would be no remaining degrees of freedom in order to minimize the function $\|F_1(c)\|_2^2$.

It is further typical for parameter estimation problems that $n_{F_1} > n_{F_2} - n_c$, meaning that there are more measurements than degrees of freedom. In fact it is desirable to have $n_{F_1} \gg n_{F_2} - n_c$ in order to estimate the parameters reliably in the sense of *confidence intervals*. Confidence intervals are an important aspect in the statistical analysis of parameter estimates and are discussed in detail in Subsection 12.2.4.

10.4. Characterization of Solutions by Optimality Conditions

The presentation of this section is based upon Bock [39], pages 45ff and Nocedal and Wright [195], pages 304ff.

Even if a point c is known to be a (strict) local solution it is, in general, very hard to decide whether or not it is also a (strict) global solution. Typically, the knowledge on the global shape of $\|F_1(c)\|_2^2$ and $F_2(c)$ that would be required for such a characterization is not available. However, the set of necessary conditions for a local solution and the set of sufficient condition for a strict local solution are well-known and recalled in the following.

As a first step toward the formulation of the necessary and sufficient optimality conditions, the *tangent space of the feasible set* and the *linearized feasible direction set* are defined.

Definition 10.9 (Tangent to the Feasible Set, Tangent Space)

A vector $\Delta c \in \mathbb{R}^{n_c}$ is called a tangent to the feasible set $\mathcal{F}$ *at the point $c \in \mathcal{F}$ if there is a sequence ξ_k of feasible points, $\lim_{k\to\infty} \xi_k = c$ and a sequence of positive scalars α_k, $\lim_{k\to\infty} \alpha_k = 0$, such that*

$$\Delta c = \lim_{k\to\infty} \frac{\xi_k - c}{\alpha_k}. \tag{10.22}$$

The set of all possible tangents Δc is called the tangent space *and denoted by $\mathcal{Z}(c)$.*

Definition 10.10 (Linearized Feasible Direction Set)

Consider a feasible point, i.e. $c \in \mathcal{F}$. Then, if the constraint function F_2 is differentiable in c, the set

$$\mathcal{A}(c) := \{\Delta c \in \mathbb{R}^{n_c} \mid \mathbf{J}_2(c)\Delta c = 0,\ \Delta c \neq 0\} \tag{10.23}$$

is defined and called the linearized feasible direction set*.*

The linearized feasible direction set contains all those directions for which the equality constraints remain fulfilled in the sense of a first order Taylor expansion. Contrariwise, the tangent space defines the shape of the feasible set in the vicinity of a point c in terms of other feasible points, and not by the mathematical description of the feasible set in terms of the constraint function $F_2(c)$.

It is natural to ask for the relation between the tangent space and the linearized feasible direction set. An important property of the constraint function $F_2(c)$ in this context is defined as follows.

Definition 10.11 (Linear Independence Constraint Qualification)

A point $c \in \mathbb{R}^{n_c}$ satisfies the linear independence constraint qualification *if F_2 is differentiable and if the rows of the Jacobian matrix $\mathbf{J}_2(c)$ are linearly independent. For the typical case $n_{F_2} < n_c$, it follows that*

$$\operatorname{rank}(\mathbf{J}_2(c)) = n_{F_2}. \tag{10.24}$$

Consider next the following lemma, which clarifies the relation between $\mathcal{A}(c)$ and $\mathcal{Z}(c)$.

Lemma 10.12 (Relationship Between Tangent Space and Linearized Feasible Direction Set)

For all $c \in \mathcal{F}$ it holds that $\mathcal{Z}(c) \subseteq \mathcal{A}(c)$. Further, if linear independence constraint qualification (equation (10.24)) holds in c, then it follows that $\mathcal{Z}(c) = \mathcal{A}(c)$.

Proof
See Nocedal and Wright [195], page 323f. ■

In short, the linear independence constraint qualification ensures that the linearization of the equality constraints (in the point c) yields an adequate description of the shape of the feasible set (in the vicinity of c).

Two further crucial definitions for the characterization of solutions by optimality conditions are as follows.

Definition 10.13 (Lagrange Function, Lagrangian)

The function

$$L(c, \lambda) = \|F_1(c)\|_2^2 + \lambda^T F_2(c) \tag{10.25}$$

is called the Lagrange function *or the* Lagrangian *of the nonlinear constrained least-squares problem (10.18).*

Definition 10.14 (Hessian Matrix of the Lagrange Function, Hessian)

If the Lagrange function $L(c, \lambda)$ is twice differentiable with respect to the parameters c, then the second derivative

$$\mathbf{H}(c, \lambda) = \left. \frac{\partial^2}{\partial c'^2} L(c', \lambda) \right|_{c'=c} \tag{10.26}$$

is called the Hessian matrix of the Lagrange function*, or, in short, the* Hessian*.*

With the above-established definitions and notations, it is possible to formulate the following theorem on the necessary optimality conditions of first and second order.

Theorem 10.15 (Necessary Optimality Conditions of First and Second Order)

Let F_i, $i = 1, 2$, be twice continuously differentiable functions, i.e. $F_i \in \mathscr{C}^2(\mathbb{R}^{n_c}, \mathbb{R}^{n_{F_i}})$, and let $\hat{c}$ be a local solution of problem (10.18) that satisfies the linear independence constraint qualification (Definition 10.11). Then the following holds:

- *$\hat{c}$ is a feasible point,*
- *the* necessary optimality condition of first order *is fulfilled: there exists a unique $\hat{\lambda}$ such that*

$$\left. \frac{d}{dc} L(c, \lambda) \right|_{(c,\lambda)=(\hat{c},\hat{\lambda})} = 2F_1^T(\hat{c})\mathbf{J}_1(\hat{c}) - \hat{\lambda}^T \mathbf{J}_2(\hat{c}) = 0, \tag{10.27}$$

- *and the* necessary optimality condition of second order *is fulfilled: the Hessian matrix of the Lagrangian is positive semi-definite on the tangent space of the feasible set:*

$$\Delta c^T \mathbf{H}(\hat{c}, \hat{\lambda}) \Delta c \geqslant 0 \quad \forall \Delta c \in \mathcal{Z}(\hat{c}) \tag{10.28}$$

Proof
See Nocedal and Wright [195], pages 321 and 332. ■

Definition 10.16 (Karush-Kuhn-Tucker Conditions, Karush-Kuhn-Tucker Points)

The equations (10.20) and (10.27) together are called the Karush-Kuhn-Tucker conditions *(in short:* KKT conditions*) for the nonlinear constrained least-squares problem (10.18). A pair $(\hat{c}, \hat{\lambda})$ that fulfills these conditions is called a* Karush-Kuhn-Tucker point *(in short:* KKT point*).*

The following theorem allows to conclude, from a set of *sufficient conditions of first and second order*, that a given point in parameter space is a strict local solution.

Theorem 10.17 (Sufficient Optimality Conditions of First and Second Order)

Let $F_i \in \mathscr{C}^2(\mathbb{R}^{n_c}, \mathbb{R}^{n_{F_i}})$ for $i = 1, 2$, and let the following sufficient optimality condition of first order *be fulfilled:*

- *$(\hat{c}, \hat{\lambda})$ is a KKT point of the minimization problem (10.18).*

Assume further that the following sufficient optimality condition of second order *is fulfilled:*

- *The Hessian matrix* $\mathbf{H}(c, \lambda)$ *is positive definite on the linearized feasible direction set* $\mathcal{A}(\hat{c})$*, i.e. for all* $\Delta c \in \mathcal{A}(\hat{c})$ *it holds that*

$$\Delta c^T \mathbf{H}(\hat{c}, \hat{\lambda}) \Delta c > 0. \tag{10.29}$$

Then $\hat{c}$ *is a strict local solution.*

Proof
See Nocedal and Wright [195], page 333f. ■

11. Numerical Methods for Parameter Estimation

Vielmehr erscheint es gerade als Vorzug des Gauß-Newton Verfahrens, nicht gegen stationäre Punkte mit $\kappa > 1$ zu konvergieren.

Bock, in his PhD Dissertation [39], pointing out a favorable convergence property of a Gauss-Newton method for parameter estimation.

In Chapter 10 the concept of maximum likelihood estimation was recalled. For the special case that the measurement errors are assumed to be normally distributed with known covariance matrix, maximum likelihood estimation is equal to the minimization of the residuals in a weighted Euclidean norm. This minimization is, in general, subject to additional equality constraints. Hence, a nonlinear constrained least-squares problem was obtained.

The topic of this chapter is to recall numerical methods for the solution of this optimization problem. More precisely, a *Generalized Gauss-Newton method* is discussed along with a modification for the treatment of ill-conditioned and singular problems. Furthermore, the *restrictive monotonicity test* is recalled as a strategy to select the stepsizes in a *damped Generalized Gauss-Newton method*.

Damped Generalized Gauss-Newton methods such as those presented in this chapter have successfully been used for solving parameter estimation problems in various disciplines. For examples from automotive engineering, biology, chemical engineering, and space flight, see Bock [36, 38, 39], Schlöder and Bock [223], Schlöder [222], Baake and Schlöder [7], Körkel [164], Kirches [160], Bock, Kostina, and Schlöder [42], Lenz [171], Lenz et al. [172], and Binder et al. [31].

Organization of This Chapter

Section 11.1 contains the basic algorithm of a Generalized Gauss-Newton Method and recalls a theorem that establishes the local convergence of this method. In Section 11.2, a modification of the Generalized Gauss-Newton method for the treatment of singular and ill-conditioned parameter estimation problems is discussed. Eventually, Section 11.3 motivates the use of a damped Generalized Gauss-Newton method and presents a practical algorithm that employs the so-called restrictive monotonicity test.

The presentation of parts of this chapter relies on Bock [39].

11.1. Generalized Gauss-Newton Method

11.1.1. Definition of the Method

As a starting point for this section, the nonlinear constrained least-squares problem (10.18) is recalled:

$$\min_c \|F_1(c)\|_2^2 \tag{11.1a}$$

$$\text{subject to} \quad F_2(c) = 0. \tag{11.1b}$$

As discussed in Section 10.4, a problem of this form arises in constrained maximum likelihood estimation for normally distributed measurement errors.

In the following, also the term *linear constrained least-squares problem* is defined.

Definition 11.1 (Linear Constrained Least-Squares Problem)

A minimization problem of the form

$$\min_c \|a_1 + \mathbf{A}_1 c\|_2^2 \tag{11.2a}$$

$$\text{subject to} \quad a_2 + \mathbf{A}_2 c = 0, \tag{11.2b}$$

with vectors $a_i \in \mathbb{R}^{n_{A_i}}$ *and matrices* $\mathbf{A}_i \in \mathbb{R}^{n_{A_i} \times m_{A_i}}$*, is called a* linear constrained least-squares problem.

The solution method that is regarded here is a so-called *Generalized Gauss-Newton Method* (cf. Bock [36], [38], and Bock [39], page 47).

Algorithm 11.2 (Gauss-Newton Method, Generalized Gauss-Newton Method)

Assume that the functions F_i *are continuously differentiable, i.e.* $F_i \in \mathscr{C}^1(\mathbb{R}^{n_c}, \mathbb{R}^{n_{F_i}})$.

1. *Start with* $k = 0$ *and with an initial guess* c^0 *for the unknowns.*
2. *Determine the solution* Δc^k *of the following linear constrained least-squares problem*

$$\min_{\Delta c} \|F_1(c^k) + \mathbf{J}_1(c^k)\Delta c\|_2^2 \tag{11.3a}$$

$$\text{subject to} \quad F_2(c^k) + \mathbf{J}_2(c^k)\Delta c = 0, \tag{11.3b}$$

   where $\mathbf{J}_i(c^k) := dF_i(c)/dc|_{c=c^k}$.
3. *Set* $c^{k+1} = c^k + \Delta c^k$.
4. *If* $\|\Delta c^k\| \leqslant \epsilon_{term}$*, with* ϵ_{term} *being a given threshold, then terminate the algorithm. Otherwise set* $k = k + 1$ *and go back to step 2.*

In step 4 of the algorithm, any norm in $\mathbb{R}^{n_c}$ *may be used for computing* $\|\Delta c^k\|$.

The above algorithm is called Generalized Gauss-Newton method*. If there are no equality constraints, i.e.* $n_{F_2} = 0$*, then the algorithm is called* Gauss-Newton method.

As a next step, the notion of a *generalized inverse* is defined.

Definition 11.3 (Generalized Inverse)

Consider a matrix $\mathbf{A} \in \mathbb{R}^{n_A \times m_A}$*. Then a matrix* $\mathbf{A}^+ \in \mathbb{R}^{m_A \times n_A}$ *that fulfills the relation* $\mathbf{A}^+\mathbf{A}\mathbf{A}^+ = \mathbf{A}^+$ *is called a* generalized inverse *of* $\mathbf{A}$.

The Generalized Gauss-Newton method requires, in each iteration, the solution of a linear constrained least-squares problem. The next theorem gives sufficient conditions for the existence and uniqueness of the solution of the linear problem. In order to formulate the theorem, the following convenient notations are introduced:

$$F(c^k) := \begin{pmatrix} F_1(c^k) \\ F_2(c^k) \end{pmatrix} \quad \text{and} \quad \mathbf{J}(c^k) := \begin{pmatrix} \mathbf{J}_1(c^k) \\ \mathbf{J}_2(c^k) \end{pmatrix}. \tag{11.4}$$

Theorem 11.4 (Existence and Uniqueness of Solutions of Linear Constrained Least-Squares Problems)

Consider the linear constrained least-squares problem (11.3) and assume that the rank conditions $\operatorname{rank}(\mathbf{J}_2(c^k)) = n_{F_2} \leqslant n_c$ *and* $\operatorname{rank}(\mathbf{J}(c^k)) = n_c$ *are fulfilled. Then the following holds:*

1. *For arbitrary* $F_1(c^k) \in \mathbb{R}^{n_{F_1}}$ *and arbitrary* $F_2(c^k) \in \mathbb{R}^{n_{F_2}}$ *there exists exactly one KKT point* $(\Delta c^k, \lambda^k)$ *of the minimization problem (11.3). The point* Δc^k *is a strict local minimum.*
2. *There exists a generalized inverse* $\mathbf{J}^+(c^k)$ *of the matrix* $\mathbf{J}(c^k)$*, and the strict local minimum* Δc^k *can be represented as*

$$\Delta c^k = -\mathbf{J}^+(c^k)F(c^k). \tag{11.5}$$

3. The KKT point $(\Delta c^k, \lambda^k)$ can be expressed as

$$\begin{pmatrix} \Delta c^k \\ -\frac{1}{2}\lambda^k \end{pmatrix} = -\begin{pmatrix} \mathbf{J}_1^T(c^k)\mathbf{J}_1(c^k) & \mathbf{J}_2^T(c^k) \\ \mathbf{J}_2(c^k) & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{J}_1^T(c^k) & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \begin{pmatrix} F_1(c^k) \\ F_2(c^k) \end{pmatrix}. \tag{11.6}$$

Proof
See Bock [39], page 56. ■

The generalized inverse $\mathbf{J}^+(c^k)$ thus takes the form

$$\mathbf{J}^+(c^k) = \begin{pmatrix} \mathbf{1} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{J}_1^T(c^k)\mathbf{J}_1(c^k) & \mathbf{J}_2^T(c^k) \\ \mathbf{J}_2(c^k) & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{J}_1^T(c^k) & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}. \tag{11.7}$$

The explicit construction of the generalized inverse $\mathbf{J}^+(c^k)$ by means of equation (11.7) is in practice not necessary. Instead, it is numerically favorable to compute the solution of the linear constrained least-squares problem by means of matrix decompositions. This issue is discussed in Section 13.3.

11.1.2. Convergence Properties

The justification for approaching nonlinear constrained least-squares problems by a Generalized Gauss-Newton method is provided by the next lemma and the subsequent theorem.

Lemma 11.5 (KKT points of Linear and Nonlinear Constrained Least-Squares Problems)

Let $F_i \in \mathscr{C}^1(\mathbb{R}^{n_c}, \mathbb{R}^{n_{F_i}})$. Then $(c, \lambda) = (\hat{c}, \hat{\lambda})$ is a KKT point of the nonlinear constrained least-squares problem (11.1) if and only if $(\Delta c, \lambda) = (0, \hat{\lambda})$ is a KKT point of the linear constrained least-squares problem

$$\min_{\Delta c} \|F_1(\hat{c}) + \mathbf{J}_1(\hat{c})\Delta c\|_2^2 \tag{11.8a}$$

$$\text{subject to} \quad F_2(\hat{c}) + \mathbf{J}_2(\hat{c})\Delta c = 0. \tag{11.8b}$$

Proof
Follows directly by considering the KKT conditions (Definition 10.16) of the problems (11.1) and (11.8). ■

Theorem 11.6 (Local Contraction (of Generalized Gauss-Newton Methods))

Let $F_i \in \mathscr{C}^1(\mathcal{D}^c, \mathbb{R}^{n_{F_i}})$, where $\mathcal{D}^c \subset \mathbb{R}^{n_c}$. Consider $y \in \mathcal{D}^c$, $z \in \mathcal{D}^c$, $z = y + \Delta y$, where $\Delta y = -\mathbf{J}^+(y)F(y)$ is the increment of a Generalized Gauss-Newton method and $\mathbf{J}^+(y)$ is the generalized inverse of $\mathbf{J}(y)$. Let further $\tilde{z} \in \mathcal{D}^c$. Assume that the following conditions are fulfilled for all y, z, $\tilde{z}$, and $\vartheta \in [0, 1]$:

1. $\|\mathbf{J}^+(z)\left(\mathbf{J}(y + \vartheta\Delta y) - \mathbf{J}(y)\right)\Delta y\| \leqslant \omega\vartheta\|\Delta y\|^2$ *with* $\omega < \infty$,
2. $\|\mathbf{J}^+(\tilde{z})R(y)\| \leqslant \kappa(x)\|\tilde{z} - y\| \leqslant \kappa\|\tilde{z} - y\|$, *where* $R(y) := F(y) + \mathbf{J}(y)\Delta y$ *and* $\kappa < 1$.

Let $c^0 \in \mathcal{D}^c$ be an initial guess such that

3. $\delta_0 := \kappa + \frac{\omega}{2}\|\Delta c^0\| < 1$, *where* $\|\Delta c^0\| = \|\mathbf{J}^+(c^0)F(c^0)\|$,
4. *the ball centered at c^0 defined by* $\mathcal{B}_{c^0} := \left\{c \mid \|c - c^0\| \leqslant \frac{\|\Delta c^0\|}{1-\delta_0}\right\}$ *is contained in $\mathcal{D}^c$.*

Then it holds that

(I) the iterates c^k are within $\mathcal{B}_{c^0}$,

(II) there exists $\hat{c} \in \mathcal{B}_{c^0}$ such that $c^k \to \hat{c}$ and $\|\Delta c^k\| \to 0$ for $k \to \infty$,

(III) $\|\Delta c^{k+1}\| \leqslant \delta_k\|\Delta c^k\|$, with $\delta_k := \kappa + \frac{\omega}{2}\|\Delta c^k\|$.

(IV) $\|c^k - \hat{c}\| \leqslant \delta_k \frac{\|\Delta c^k\|}{1-\delta_k}$.

Proof
See Bock [39], page 59. ■

This is a generalization of Theorem 6.14, which establishes the local convergence of Newton-type methods applied to systems of nonlinear equations (see Section 6.5). Both theorems are valid for any norm on finite-dimensional spaces.

The local contraction theorem (Theorem 11.6) guarantees, under certain conditions on the derivatives $\mathbf{J}(c)$, on the generalized inverse $\mathbf{J}^+(c)$, and on the initial guess c^0, that the iterates c^k converge to a point $\hat{c}$ and that the increments Δc^k converge to 0. Thus, there exists $\hat{\lambda}$ such that $(\Delta c, \lambda) = (0, \hat{\lambda})$ is a KKT point of the problem (11.8). By using Lemma 11.5, it further follows that $(\hat{c}, \hat{\lambda})$ is a KKT point of the nonlinear constrained least-squares problem (11.1).

The question that remains is whether $(\Delta c, \lambda) = (0, \hat{\lambda})$ and $(c, \lambda) = (\hat{c}, \hat{\lambda})$ are also (strict) local minima of the linear and nonlinear constrained least-squares problems, respectively.

In order to answer this question, recall Theorem 10.17. This theorem states that a sufficient condition for a minimum is that the Hessian matrix of the Lagrange function is positive definite on the linearized feasible direction set.

For the linear constrained least-squares problem (11.8), this condition is equivalent to the positive definiteness of $\mathbf{J}_1^T(\hat{c})\mathbf{J}_1(\hat{c})$ on the kernel of $\mathbf{J}_2(\hat{c})$. Such a condition can easily be checked in practice, see Section 13.3.

For the nonlinear constrained least-squares problem (11.1) the Hessian matrix can be expressed as

$$\mathbf{H}(\hat{c}, \hat{\lambda}) = 2\mathbf{J}_1^T(\hat{c})\mathbf{J}_1(\hat{c}) + 2F_1^T(\hat{c})\frac{\partial \mathbf{J}_1(\hat{c})}{\partial c} - \hat{\lambda}^T \frac{\partial \mathbf{J}_2(\hat{c})}{\partial c}, \tag{11.9}$$

where the vector-tensor products in the second term and in the third term are to be understood as

$$F_1^T(\hat{c})\frac{\partial \mathbf{J}_1(\hat{c})}{\partial c} := \sum_{i=1}^{n_{F_1}} (F_1(\hat{c}))_i \frac{\partial F_{1,i}(\hat{c})}{\partial c_j \partial c_k} \tag{11.10a}$$

$$\hat{\lambda}^T \frac{\partial \mathbf{J}_2(\hat{c})}{\partial c} := \sum_{i=1}^{n_{F_2}} \hat{\lambda}_i \frac{\partial F_{2,i}(\hat{c})}{\partial c_j \partial c_k}. \tag{11.10b}$$

Due to the presence of the second and the third term in equation (11.9), positive definiteness of $\mathbf{J}_1^T(\hat{c})\mathbf{J}_1(\hat{c})$ is not sufficient for positive definiteness of the Hessian matrix $\mathbf{H}(\hat{c}, \hat{\lambda})$; the Hessian may become indefinite or negative definite if the "second order terms" are sufficiently large. Thus, a minimum of the linear problem may turn out to be a saddle point or even a maximum of the nonlinear problem.

Importantly, the following results can be proven for the obtained solution $\hat{c}$.

(i) Let condition 2 in Theorem 11.6 be fulfilled for some $\kappa < 1$ for all points y, z in a neighborhood of the KKT point $\hat{c}$. Then it follows that the Hessian $\mathbf{H}(\hat{c}, \hat{\lambda})$ is positive definite. Thus, $\hat{c}$ is a strict local minimum of the nonlinear constrained least-squares problem (11.1).

(ii) Let $\hat{c}$ be a KKT point such that condition 2 does not hold for some $\kappa < 1$. Let further ξ be a perturbation of the measurement data – which enter the function F_1 – whose order of magnitude is equal to the size of the residuals $R(\hat{c})$. Then there exists a pair $(\hat{c}(\xi), \hat{\lambda}(\xi))$, which is a KKT point of the perturbed problem. However, $\hat{c}(\xi)$ is not a minimum of the perturbed problem. The solution $\hat{c}$ of the unperturbed problem can thus be characterized as statistically instable.

The key argument for these results is that $\kappa < 1$ can, on the one hand, be interpreted as a condition on the quality of the measurement data (i.e. the measurement errors should be sufficiently small). On the other hand, it holds that the condition $\kappa < 1$ limits the size of $F_1^T(\hat{c})$ and $\hat{\lambda}$ such that the second order terms given in the equations (11.10) remain sufficiently small compared to the Hessian $2\mathbf{J}_1^T(\hat{c})\mathbf{J}_1(\hat{c})$ of the linear problem (which is positive definite under the given rank conditions).

For a further discussion of the two results (i) and (ii), it is referred to Bock [39], page 63ff.

11.2. Regularization Strategy for Singular and Ill-Conditioned Problems

11.2.1. Motivation and Goal

In step 2 of the Generalized Gauss-Newton method (Algorithm 11.2) the linear constrained least-squares problem (11.3) has to be solved. According to Theorem 11.4, a unique solution of this linear problem exists provided that two rank conditions are fulfilled: $\operatorname{rank}(\mathbf{J}_2(c^k)) = n_{F_2}$ and $\operatorname{rank}(\mathbf{J}(c^k)) = n_c$. The topic of this section is a modification of the Generalized Gauss-Newton method for the case that the second of the two rank conditions is violated.

In the following, a linear constrained least-squares problem that violates the rank condition $\operatorname{rank}(\mathbf{J}(c^k)) = n_c$ is called a *singular problem*. Further, if the rank condition is fulfilled but the matrix $\mathbf{J}(c^k)$ is ill-conditioned, then the linear constrained least-squares problem is called an *ill-conditioned problem*. Please note that in finite precision arithmetic, the rank of a matrix is ill-defined in the sense that a tiny round-off error may alter a singular matrix in such a way that it has formally full rank but a very poor condition.

Ill-conditioning of $\mathbf{J}(c^k)$ implies ill-conditioning of the generalized inverse $\mathbf{J}^+(c^k)$ and thus a high sensitivity of the increment $\Delta c^k = -\mathbf{J}^+(c^k)F(c^k)$ toward changes in $F(c^k)$. This is particularly critical because $F_1(c^k)$ contains differences between evaluations of measurement functions and measurements, and thus quantities that depend on the specific realization ϵ of the random measurement errors. Therefore, in a practical realization, it is reasonable to modify the Generalized Gauss-Newton method in both the singular and in the ill-conditioned case.

In view of the above discussion, it is appropriate to mention that the underlying assumption for this section is that $\operatorname{rank}(\mathbf{J}_2(c^k)) = n_{F_2}$ and that $\mathbf{J}_2(c^k)$ has a small or moderate condition number.

11.2.2. Reduced Form of the Linear Constrained Least-Squares Problem

For the following discussion, it is convenient to define the short notation

$$\mathbf{J} := \mathbf{J}(c^k) = \begin{pmatrix} \mathbf{J}_1(c^k) \\ \mathbf{J}_2(c^k) \end{pmatrix} \tag{11.11}$$

and to assume that $\mathbf{J}$ is given in *reduced form*.

Definition 11.7 (Jacobian Matrix in Reduced Form)

Let $\mathbf{J} \in \mathbb{R}^{(n_{F_1}+n_{F_2})\times n_c}$ be the Jacobian matrix of a linear constrained least-squares problem. Assume that $\mathbf{J}$ takes the form

$$\mathbf{J} := \begin{pmatrix} \mathbf{A} & \mathbf{S} \\ \mathbf{L} & \mathbf{0} \end{pmatrix} \tag{11.12}$$

with $\mathbf{L} \in \mathbb{R}^{n_{F_2}\times n_{F_2}}$ being a regular lower triangular matrix, and $\mathbf{S} \in \mathbb{R}^{n_{F_1}\times(n_c-n_{F_2})}$ being a diagonal matrix with entries s_i, $1 \leqslant i \leqslant n_c - n_{F_2}$, such that

$$s_1 \geqslant s_2 \geqslant \cdots \geqslant s_{n_c-n_{F_2}} \geqslant 0. \tag{11.13}$$

Further, let $\mathbf{A} \in \mathbb{R}^{n_{F_1}\times n_{F_2}}$.

Then the matrix $\mathbf{J}$ is called a Jacobian matrix in reduced form.

The Jacobian matrix $\mathbf{J}$ (equation (11.11)) of any linear constrained least-squares problem (11.3) can – under the given assumption $\operatorname{rank}(\mathbf{J}_2(c^k)) = n_{F_2}$ – be brought into this reduced form (see Section 13.3 for details) by orthogonal transformations.

Definition 11.8 (Linear Constrained Least-Squares Problem in Reduced Form)

Consider a linear constrained least-squares problem of the form (11.3) and assume that $\mathbf{J} := \mathbf{J}(c^k)$ is given in reduced form. Then the problem can be rewritten as

$$\min_{\Delta c} \|\tilde{F}_1 + \mathbf{A}\Delta c_1 + \mathbf{S}\Delta c_2\|_2^2 \tag{11.14a}$$

$$\text{subject to} \quad \tilde{F}_2 + \mathbf{L}\Delta c_1 = 0, \tag{11.14b}$$

with $\tilde{F}_1 := F_1(c^k)$, $\tilde{F}_2 := F_2(c^k)$. *Further* $\Delta c^k = (\Delta c_1^T, \Delta c_2^T)^T$ *with* $\Delta c_1 \in \mathbb{R}^{n_{F_2}}$ *and* $\Delta c_2 \in \mathbb{R}^{n_c - n_{F_2}}$.

The linear constrained least-squares problem in reduced form is singular if and only if one or several of the diagonal entries s_i of the matrix $\mathbf{S}$ are 0. Ill-conditioning of the matrix $\mathbf{J}$ arises if $s_1/s_{n_c-n_{F_2}} \gg 1$, independent of the absolute size of s_1 and $s_{n_c-n_{F_2}}$.

11.2.3. Modification for Singular and Ill-Conditioned Problems

For the practical treatment of singular and ill-conditioned problems, Bock [39], page 144f., suggests to solve the following *modified linear problem* (instead of problem (11.14)):

$$\min_{\Delta c} \left\| \tilde{F}_1 + \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \mathbf{P}_r \begin{pmatrix} \Delta c_1 \\ \Delta c_2 \end{pmatrix} \right\|_2^2 + \eta^2 \left(\|\Delta c_1\|_2^2 + \|\Delta c_2\|_2^2 \right) \tag{11.15a}$$

$$\text{subject to} \quad \tilde{F}_2 + \mathbf{L}\Delta c_1 = 0. \tag{11.15b}$$

Herein, $\eta > 0$ is a regularization factor, and $\mathbf{P}_r \in \mathbb{R}^{n_c \times n_c}$ is an orthogonal projection matrix of rank r, $n_{F_2} \leqslant r \leqslant n_c$. The projection matrix is explicitly given by

$$\mathbf{P}_r = \begin{pmatrix} \mathbf{1}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{11.16}$$

with $\mathbf{1}_r$ being the $r \times r$-dimensional identity matrix. An analogous notation is used in the following for identity matrices of other dimensions.

It is helpful to define $\mathbf{B} = \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix}$, $\tilde{\mathbf{L}} = \begin{pmatrix} \mathbf{L} & \mathbf{0} \end{pmatrix}$ and to reformulate problem (11.15) as follows:

$$\min_{\Delta c} \left\| \begin{pmatrix} \tilde{F}_1 \\ 0 \end{pmatrix} + \begin{pmatrix} \mathbf{B}\mathbf{P}_r \\ \eta \cdot \mathbf{1}_{n_c} \end{pmatrix} \Delta c \right\|_2^2 \tag{11.17a}$$

$$\text{subject to} \quad \tilde{F}_2 + \tilde{\mathbf{L}}\Delta c = 0. \tag{11.17b}$$

In this representation of the modified linear problem, and with the above assumptions that $\mathbf{L}$ is regular and $\eta > 0$, it is obvious that $\operatorname{rank}(\tilde{\mathbf{L}}) = n_{F_2}$, and that

$$\operatorname{rank} \left(\begin{pmatrix} \mathbf{B}\mathbf{P}_r \\ \eta \cdot \mathbf{1}_{n_c} \\ \tilde{\mathbf{L}} \end{pmatrix} \right) = n_c. \tag{11.18}$$

As a consequence, Theorem 11.4 can directly be applied to the modified problem (11.17), which yields the following corollary (cf. Bock [39], page 144).

Corollary 11.9 (Existence and Uniqueness of Solutions of the Modified Linear Constrained Least-Squares Problem)

Consider the modified constrained least-squares problem (11.15). Then it holds that

1. *For arbitrary* $\tilde{F}_1 \in \mathbb{R}^{n_{F_1}}$ *and arbitrary* $\tilde{F}_2 \in \mathbb{R}^{n_{F_2}}$ *there exists exactly one KKT point* $(\Delta c, \lambda)$ *of the minimization problem (11.15). The point* Δc *is a strict local minimum.*

2. *There exists a generalized inverse of the matrix* $\begin{pmatrix} \mathbf{B}\mathbf{P}_r \\ \eta \cdot \mathbf{1}_{n_c} \\ \tilde{\mathbf{L}} \end{pmatrix}$, *and the strict local minimum* Δc *can be represented as*

$$\Delta c = - \begin{pmatrix} \mathbf{B}\mathbf{P}_r \\ \eta \cdot \mathbf{1}_{n_c} \\ \tilde{\mathbf{L}} \end{pmatrix}^+ \begin{pmatrix} \tilde{F}_1 \\ 0 \\ \tilde{F}_2 \end{pmatrix}. \tag{11.19}$$

3. *The KKT point* $(\Delta c, \lambda)$ *can be expressed as*

$$\begin{pmatrix} \Delta c \\ -\frac{1}{2}\lambda \end{pmatrix} = - \begin{pmatrix} \mathbf{P}_r^T \mathbf{B}^T \mathbf{B} \mathbf{P}_r + \eta^2 \cdot \mathbf{1}_{n_c} & \tilde{\mathbf{L}}^T \\ \tilde{\mathbf{L}} & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{P}_r^T \mathbf{B}^T & \eta \cdot \mathbf{1}_{n_c} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{1}_{n_{F_2}} \end{pmatrix} \begin{pmatrix} \tilde{F}_1 \\ 0 \\ \tilde{F}_2 \end{pmatrix}. \tag{11.20}$$

Furthermore, it is possible to derive – from equation (11.20) – the following explicit expression for Δc_1 and Δc_2:

$$\Delta c_1 = -\mathbf{L}^{-1}\tilde{F}_2 \tag{11.21a}$$

$$(\Delta c_2)_i = \begin{cases} -\frac{s_i}{s_i^2+\eta^2} G_i & \text{for} \quad 1 \leqslant i \leqslant \tilde{r} \\ 0 & \text{for} \quad \tilde{r}+1 \leqslant i \leqslant n_c - n_{F_2}, \end{cases} \tag{11.21b}$$

with $\tilde{r} = r - n_{F_2}$ and $G = \tilde{F}_1 + \mathbf{A}\Delta c_1$.

Hence, the modified linear constrained least-squares problem (11.15) has, for any $\eta > 0$, a unique solution. Of particular interest is the limit $\eta \to 0$, i.e. the limit of the solution (see equation (11.21)) for vanishing regularization term in (11.15). This limit is given by

$$\Delta c_1 = -\mathbf{L}^{-1}\tilde{F}_2 \tag{11.22a}$$

$$(\Delta c_2)_i = \begin{cases} -\frac{G_i}{s_i} & \text{for} \quad 1 \leqslant i \leqslant \tilde{r} \\ 0 & \text{for} \quad \tilde{r}+1 \leqslant i \leqslant n_c - n_{F_2}. \end{cases} \tag{11.22b}$$

In order to express this solution conveniently in matrix notation, let $\tilde{\mathbf{S}} \in \mathbb{R}^{n_{F_1} \times (n_c - n_{F_2})}$ be a diagonal matrix with entries

$$s_1 \geqslant s_2 \geqslant \cdots \geqslant s_{\tilde{r}} > s_{\tilde{r}+1} = s_{\tilde{r}+2} = \cdots = s_{n_c - n_{F_2}} = 0. \tag{11.23}$$

That is, $\tilde{\mathbf{S}}$ is obtained from $\mathbf{S}$ by setting the $n_c - n_{F_2} - \tilde{r}$ smallest diagonal elements to zero.

Let further $\tilde{\mathbf{S}}^\dagger \in \mathbb{R}^{(n_c - n_{F_2}) \times n_{F_1}}$ denote the Moore-Penrose pseudoinverse of the matrix $\tilde{\mathbf{S}}$. The matrix $\tilde{\mathbf{S}}^\dagger$ is such that the first $\tilde{r}$ diagonal elements are given by $s_1^{-1}, \ldots, s_{\tilde{r}}^{-1}$, and the remaining elements are 0.

With this notation, the solution (11.22) can be expressed as

$$\Delta c = -\begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \tilde{\mathbf{S}}^\dagger & -\tilde{\mathbf{S}}^\dagger \mathbf{A} \mathbf{L}^{-1} \end{pmatrix} \begin{pmatrix} \tilde{F}_1 \\ \tilde{F}_2 \end{pmatrix}. \tag{11.24}$$

It can easily be verified that the matrix in equation (11.24) is a *rank-deficient generalized inverse* of the matrix $\mathbf{J}$ as given in equation (11.12), i.e. for

$$\mathbf{J}_{[r]}^+ := \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \tilde{\mathbf{S}}^\dagger & -\tilde{\mathbf{S}}^\dagger \mathbf{A} \mathbf{L}^{-1} \end{pmatrix} \tag{11.25}$$

it holds that $\mathbf{J}_{[r]}^+ \mathbf{J} \mathbf{J}_{[r]}^+ = \mathbf{J}_{[r]}^+$. It is noted that the subscript r indicating the rank has been enclosed in square brackets in order to avoid confusion with the Jacobian matrices $\mathbf{J}_1$ and $\mathbf{J}_2$ of the functions F_1 and F_2, respectively.

Since $\mathbf{J}_{[r]}^+$ possesses the property to be a generalized inverse of $\mathbf{J}$, the modified version of the Generalized Gauss-Newton method based on a rank-deficient generalized inverse is also locally convergent (cf. Theorem 11.6 and Bock [39], page 148 and page 150). An analysis of the obtained *rank-deficient solution* is given in Section 12.3.

11.2.4. Practical Rank Decision

It remains to discuss the practical choice of the rank r (or, equivalently, the practical choice of $\tilde{r}$) in the modified version of the Generalized Gauss-Newton method. One aim of using a rank $\tilde{r} < n_c - n_{F_2}$ is to ensure that the matrix $\mathbf{S}$ is well-conditioned on the space $\{\mathbf{P}_r \Delta c \mid \Delta c \in \mathbb{R}^{n_c}\}$. More precisely, if γ_{max} is a given upper bound to the acceptable condition of $\mathbf{S}$ in the spectral norm, then set $s_{min} = s_1/\gamma_{max}$ and choose $\tilde{r}$ such that

$$s_1 \geqslant \cdots \geqslant s_{\tilde{r}} > s_{min} \geqslant s_{\tilde{r}+1} = \cdots = s_{n_c - n_{F_2}} = 0. \tag{11.26}$$

In addition to this regularization based on the condition number, it is useful in the context of parameter estimation to define a lower bound for the diagonal elements s_i of the matrix $\mathbf{S}$. As seen later in Section 12.2, the quantity s_i^{-2} is – in the solution of the least-squares problem – a measure

for the variance of the corresponding parameter estimate as a function of the data, i.e. for the uncertainty in the parameter estimate. Given an appropriate scaling of the parameters, the choice $s_{min} = 1/\sigma_{max}$ guarantees that only those parameters are estimated whose relative variance, as a function of the data, is less than σ^2_{max}.

The combination of both regularization strategies thus leads to the choice

$$s_{min} = \max\left(\frac{s_1}{\gamma_{max}}, \frac{1}{\sigma_{max}}\right). \tag{11.27}$$

11.3. Damped Generalized Gauss-Newton Methods

11.3.1. Definition of Damped Method

Theorem 11.6 ensures – under certain conditions – local convergence of the Generalized Gauss-Newton method (Algorithm 11.2) to a KKT point of problem (11.1). One of the conditions is that the "initial increment" Δc^0 is such that $\delta_0 = \kappa + \omega\|\Delta c^0\| < 1$. In practical situations, it may be difficult to find an initial guess c^0 such that this assumption is fulfilled, and the Generalized Gauss-Newton method presented in Algorithm 11.2 may fail to converge.

One approach for constructing a method with improved convergence properties is the use of *damped methods*.

Definition 11.10 (Damped Generalized Gauss-Newton Method)

If the iteration $c^{k+1} = c^k + \Delta c^k$ *(step 3 in the Generalized Gauss-Newton method (Algorithm 11.2) is replaced by the following* damped iteration step*:*

$$c^{k+1} = c^k + \alpha^k \Delta c^k, \quad \textit{with} \quad \alpha^k \in (0, 1], \tag{11.28}$$

then the resulting method is called a damped (Generalized Gauss-Newton) method*.*

In contrast, the original Generalized Gauss-Newton method as stated in Algorithm 11.2 is in the following referred to as *full-step (Generalized Gauss-Newton) method*.

11.3.2. Level Functions

The basic idea of the damped method is to choose the stepsize α^k in such a way that the new iterate c^{k+1} is in some sense "better" then the old iterate c^k. The use of damped iterations therefore requires a definition of a "good" iterate. In view of the fact that a minimum of the Lagrange function (equation (10.25)) is sought, the use of the following *level function* as a measure for the quality of an iterate seems appropriate:

$$T(c) := \|F_1(c)\|_2^2 + \sum_{i=1}^{n_{F_2}} \beta_i |F_{2,i}(c)|. \tag{11.29}$$

Herein, $\beta_i > |\lambda_i|$, which attributes a higher weight to the equality constraints.

Indeed, it is possible to show – under mild assumptions – that the sequence of iterates of a damped Generalized Gauss-Newton method converges for any arbitrary initial guess c^0 to a KKT point of the nonlinear constrained least-squares problem (11.1) if α^k is determined by

$$\alpha^k = \text{argmin}_{\alpha \in [0,1]}(T(c^k + \alpha \Delta c^k)), \tag{11.30}$$

see Bock [39], page 77. Such a "global convergence" result also holds if α^k is given by an approximation of the minimum of the level function along the increment Δc^k.

Unfortunately, this damped Generalized Gauss-Newton method may converge – in contrast to the full-step method – to "stastically instable" minima with $\kappa > 1$ (Bock [39], page 79).

Moreover, already for mildly ill-conditioned problems (e.g. condition $\gamma \approx 10^2$) the stepsizes α^k as determined by equation (11.30) become very small, see Bock [39], page 80ff. Hence, the convergence becomes very slow, even in cases where the obtained KKT point is a minimum with $\kappa < 1$.

This motivates the use of alternative level functions and alternative strategies for the determination of the stepsizes α^k. In the following, *natural level functions* of the form

$$T_{nat}(c; c^k) := \|\mathbf{J}^+(c^k)F(c)\|_2^2 \tag{11.31}$$

are discussed. This level function can – for $c^k \to \hat{c}$ – be interpreted as an approximation of $\|\mathbf{J}^+(\hat{c})F(c)\|$, which, in turn, approximates the distance of the parameters c to the solution $\hat{c}$, i.e. $\|c - \hat{c}\|_2^2$, see Bock [39], page 83f.

The natural level function is "compatible" with the Generalized Gauss-Newton method in the sense that it decreases into the direction of the increment Δc^k. More rigorously, the following upper bound holds (see also Bock [39], page 85ff).

Lemma 11.11 (Descent of Natural Level Function along Generalized Gauss-Newton Increment)

Let $\mathcal{D}^c \subset \mathbb{R}^{n_c}$ be open, $F_i \in \mathscr{C}^2(\mathcal{D}^c, \mathbb{R}^{n_{F_i}})$. Consider $c \in \mathcal{D}^c$, let Δc be the corresponding Generalized Gauss-Newton increment, and assume that $c + \alpha\Delta c \in \mathcal{D}^c$ for all $\alpha \in [0,1]$. Further, define

$$w(c,\alpha) := \sup_{\beta\in(0,\alpha]} \frac{\|\mathbf{J}^+(c)\left[\mathbf{J}(c+\beta\Delta c) - \mathbf{J}(c)\right]\Delta c\|_2}{\beta\|\Delta c\|_2^2} \tag{11.32}$$

and assume $w(c,\alpha) \leqslant w_{max} < \infty$.

Then it holds that

$$T_{nat}(c+\alpha\Delta c; c) \leqslant \left(1 - \alpha + \frac{\alpha^2}{2} w(c,\alpha)\|\Delta c\|_2\right)^2 T_{nat}(c;c). \tag{11.33}$$

Proof (cf. Bock [39], page 85ff)
At first, the triangular inequality $\|u_1\| - \|u_2\| \leqslant |\|u_1\| - \|u_2\|| \leqslant \|u_1 - u_2\|$ is used with $u_1 = \mathbf{J}^+(c)F(c+\alpha\Delta c)$ and $u_2 = (1-\alpha)\mathbf{J}^+(c)F(c)$. This yields

$$\begin{aligned}
&\left\|\mathbf{J}^+(c)F(c+\alpha\Delta c)\right\|_2 - (1-\alpha)\left\|\mathbf{J}^+(c)F(c)\right\|_2 \\
\leqslant &\left\|\mathbf{J}^+(c)\Big[F(c+\alpha\Delta c) - (1-\alpha)F(c)\Big]\right\|_2 \\
\leqslant &\left\|\mathbf{J}^+(c)\Big[F(c+\alpha\Delta c) - F(c) + \alpha\mathbf{J}^+(c)\mathbf{J}(c)F(c) + \alpha\Big(\mathbf{1} - \mathbf{J}^+(c)\mathbf{J}(c)\Big)F(c)\Big]\right\|_2
\end{aligned} \tag{11.34}$$

Using the triangular inequality $\|u_1 + u_2\| \leqslant \|u_1\| + \|u_2\|$ leads to

$$\begin{aligned}
\left\|\mathbf{J}^+(c)F(c+\alpha\Delta c)\right\|_2 - (1-\alpha)\left\|\mathbf{J}^+(c)F(c)\right\|_2 \leqslant &\left\|\mathbf{J}^+(c)\Big[F(c+\alpha\Delta c) - F(c) + \alpha\mathbf{J}(c)\mathbf{J}^+(c)F(c)\Big]\right\|_2 \\
&+ \left\|\mathbf{J}^+(c)\Big[\alpha\Big(\mathbf{1} - \mathbf{J}(c)\mathbf{J}^+(c)\Big)F(c)\Big]\right\|_2
\end{aligned} \tag{11.35}$$

The second summand on the right hand side vanishes because of the defining property of generalized inverses (see Definition 11.3). With regard to the first term, it holds that

$$F(c+\alpha\Delta c) = F(c) + \int_0^\alpha \mathbf{J}(c+\beta\Delta c)\Delta c d\beta. \tag{11.36}$$

From this and with $\Delta c = -\mathbf{J}^+(c)F(c)$ it follows that

$$\begin{aligned}
\left\|\mathbf{J}^+(c)F(c+\alpha\Delta c)\right\|_2 - (1-\alpha)\left\|\mathbf{J}^+(c)F(c)\right\|_2 \leqslant &\int_0^\alpha \left\|\mathbf{J}^+(c)\left[\mathbf{J}(c+\beta\Delta c) - \mathbf{J}(c)\right]\Delta c\right\|_2 d\beta \\
\leqslant &w(c,\alpha)\int_0^\alpha \beta\|\Delta c\|_2^2 d\beta \\
\leqslant &\frac{1}{2}\alpha^2 w(c,\alpha)\|\Delta c\|_2^2,
\end{aligned} \tag{11.37}$$

and thus

$$\left\|\mathbf{J}^{+}(c)F(c+\alpha\Delta c)\right\|_2 \leqslant \left[1-\alpha+\frac{1}{2}\alpha^2 w(c,\alpha)\|\Delta c\|_2\right]\left\|\mathbf{J}^{+}(c)F(c)\right\|_2. \tag{11.38}$$

The assertion then easily follows from the definition of the natural level function. ■

11.3.3. Restrictive Monotonicity Condition

Define, for given c and Δc, the pre-factor in equation (11.38) as the function

$$u(\alpha) := 1-\alpha+\frac{1}{2}\alpha^2 w(c,\alpha)\|\Delta c\|_2. \tag{11.39}$$

Further, define for every $0<\eta<2$ the following linear function

$$v(\alpha;\eta) := 1-\alpha\left(1-\frac{1}{2}\eta\right) \tag{11.40}$$

For any fixed η it is possible to find a stepsize $\bar{\alpha}(\eta)$ such that the condition

$$u(\alpha) \leqslant v(\alpha;\eta) \tag{11.41}$$

is fulfilled for all $0<\alpha\leqslant\bar{\alpha}(\eta)$. More precisely, the maximum stepsize $\bar{\alpha}(\eta)$ for which the relation holds is given by

$$\bar{\alpha}(\eta) = \frac{\eta}{w(c,\alpha)\|\Delta c\|_2}. \tag{11.42}$$

The natural level function thus fulfills both the "standard" monotonicity condition $T_{nat}(c^k+\alpha\Delta c^k;c^k) < T_{nat}(c^k;c^k)$, and the *restrictive monotonicity condition*

$$T_{nat}(c^k+\alpha\Delta c^k;c^k) \leqslant (v(\alpha;\eta))^2 T_{nat}(c^k;c^k) \tag{11.43}$$

for all $0<\alpha\leqslant\bar{\alpha}(\eta)$. Of particular interest is the special choice $\eta=1$, because the stepsize $\bar{\alpha}(1)$ corresponds to the minimum of the function $u(\alpha)$.

Define further

$$\bar{\bar{\alpha}}(\eta) := \min(1,\bar{\alpha}(\eta)) \tag{11.44}$$

and consider a damped Generalized Gauss-Newton method, whose stepsizes are such that the relation

$$\alpha^k \in [\bar{\bar{\alpha}}(\eta_1),\bar{\bar{\alpha}}(\eta_2)] \quad \text{with} \quad 0<\eta_1\leqslant\eta_2<2. \tag{11.45}$$

holds. Then, whenever the sequence of iterates c^k converges to a point $\hat{c}$, it follows that full steps ($\alpha^k=1$) are taken in a neighborhood of $\hat{c}$, see Bock [39], page 89. Hence, it also follows that the damped method preserves the property of the full-step method to avoid convergence to statistically instable minima with $\kappa>1$.

In a practical situation, the quantity $w(c,\alpha)$ is typically unknown. Therefore, in order to obtain a practical stepsize selection strategy from the equations (11.42)-(11.45), it is necessary to find a numerical approximation of $w(c,\alpha)$. Bock [39], page 90, makes the following suggestion.

Definition 11.12 (A Posteriori Estimation Formula for $w(c^k,\alpha)$)

Let c^k be an iterate of the Generalized Gauss-Newton method and let Δc^k be the corresponding increment. Then an a posteriori estimation formula for $w(c^k,\alpha)$ *is – for any $\alpha\in(0,1]$ – given by*

$$\hat{w}(c^k,\alpha) = 2\frac{\|-\mathbf{J}^{+}(c^k)F(c^k+\alpha\Delta c^k)-(1-\alpha)\Delta c^k\|_2}{\alpha^2\|\Delta c^k\|_2^2}. \tag{11.46}$$

As shown in Bock [39], page 90f, this formula provides (under certain differentiability conditions) an asymptotically correct estimation of $w(c^k,\alpha)$ for $\alpha\to 0$. Bock [39] further shows: Selecting

$\alpha \in (0,1]$ such that

$$\alpha \leqslant \frac{\eta}{\hat{w}(c^k,\alpha)\|\Delta c^k\|_2} \tag{11.47}$$

holds for $0 < \eta < 2$ is sufficient for the restrictive monotonicity condition (11.43). Accordingly, equation (11.47) is called a practical *restrictive monotonicity test.*

11.3.4. A Practical Algorithm based on the Restrictive Monotonicity Test

The above results justify the use of the following practical stepsize selection algorithm (cf. Bock [39], page 92f):

Algorithm 11.13 (Generalized Gauss-Newton Method with a Practical Stepsize Selection Strategy based on a Restrictive Monotonicity Test)

Start with $k = 0$, $j = 0$, with an initial guess c^0, and with $\alpha^{0,0} = 1$. Further, choose η, η_2 such that $0 < \eta < \eta_2 < 2$.

1. *Determine the increment Δc^k as solution of equation (11.3).*
2. *If $\|\Delta c^k\| \leqslant \epsilon_{term}$, with ϵ_{term} being a given termination criterion, then terminate the algorithm. Otherwise, continue with step 3.*
3. *Set $c^{k+1,j} = c^k + \alpha^{k,j}\Delta c^k$.*
4. *Compute $\hat{w}(c^k,\alpha^{k,j})$ by means of the a posteriori estimation formula (11.46).*
5. *Make the restrictive monotonicity test*

$$\alpha^{k,j}\hat{w}(c^k,\alpha^{k,j})\|\Delta c^k\|_2 \leqslant \eta_2. \tag{11.48}$$

   If the condition is fulfilled, continue with step 7, otherwise continue with step 6.

6. *Propose a new stepsize*

$$\alpha^{k,j+1} = \eta/(\hat{w}(c^k,\alpha^{k,j})\|\Delta c^k\|_2) \tag{11.49}$$

   (which is smaller than $\alpha^{k,j}$), then set $j = j + 1$ and continue with step 3.

7. *Accept the stepsize and the new iterate, i.e. set $\alpha^k = \alpha^{k,j}$ and $c^{k+1} = c^{k+1,j}$.*
8. *Propose a stepsize for the next iteration by*

$$\alpha^{k+1,0} = \min(1, \eta/(\hat{w}(c^k,\alpha^{k,j})\|\Delta c^k\|_2)). \tag{11.50}$$

   Set $k = k + 1$, $j = 0$, and continue with step 1.

A typical choice is $\eta = 1$, because this is the optimal value in the sense that it corresponds to the minimum of the function $u(\alpha)$. This value is used in the predictor stepsize in step 7 and in the corrector stepsize in step 6 of the algorithm. The acceptance test in step 5 employs the larger value η_2 and could, in view of the equations (11.45) and (11.47), be made more restrictive by adding a lower bound η_1.

For the first stepsize in the first iteration, i.e. $\alpha^{0,0}$, Algorithm 11.13 attempts a full step iteration. Alternatively, one may also use

$$\alpha^{0,0} = \min(1, \eta/(\hat{w}^0\|\Delta c^0\|_2)). \tag{11.51}$$

With $\eta = 1$ and $\hat{w}^0 = 100$, this implies $\alpha^{0,0}\|\Delta c^0\|_2 \leqslant 1/100$. Using a suitable scaling of the parameters, the norm of the damped increment is approximately 1% of the norm of c^0.

In theory, it is always possible to fulfill the restrictive monotonicity test in step 5 for a sufficiently small stepsize. However, in practice it is preferable to bound the predictor and corrector stepsizes from below by α_{min} in order to avoid exceedingly small stepsizes. If no sufficient decrease is obtained by using the minimum stepsize α_{min}, the problem may be treated as locally ill-conditioned

(or singular), and another increment can be computed using a rank-deficient generalized inverse, see Bock [39], page 93. An algorithm that uses this modification is given in Section 13.3.

The damped Generalized Gauss-Newton method in Algorithm 11.13 uses full steps $\alpha^k = 1$ if the sequence of iterates c^k converges to a point $\hat{c}$. Accordingly, Algorithm 11.13 avoids convergence to statistically instable minima with $\kappa > 1$, see Bock [39], page 94.

12. Analysis of Solutions

> *It is important for parameter estimation problems to compute not only parameters but also a statistical assessment of the accuracy of these parameter estimates. This can be done by means of the covariance matrix.*
>
> Bock, Kostina, and Kostyukova, in their paper "Covariance Matrices for Parameter Estimates of Constrained Parameter Estimation Problems" [40].

In Chapter 10 it was discussed that the idea of maximum likelihood estimation is to determine those parameters that are the most likely ones to explain the given measurements η. Accordingly, the maximum likelihood estimate is implicitly also a function of η, which motivates to write $\hat{c}(\eta)$.

It was further discussed in Chapter 10 that the measurements η correspond to one specific realization ϵ of the random measurement errors $\mathfrak{e}$. In other words, if the experiment is carried out again, a different realization of the measurement errors is obtained and thus also a different maximum likelihood estimate. Therefore, if the maximum likelihood estimate is regarded as a function of the random counterpart $\mathfrak{h}$ of η, i.e. $\hat{c}(\mathfrak{h})$, it is obvious that the maximum likelihood estimate is also random.

The topic of the chapter is the quantification of the uncertainty in the maximum likelihood estimate by considering the probability distribution of $\hat{c}(\mathfrak{h})$.

Organization of This Chapter

This chapter is divided into three sections. Section 12.1 establishes the necessary notation. Section 12.2 deals with the analysis of solutions for which two regularity assumptions are fulfilled. In particular, it is shown that these regularity assumptions guarantee that the covariance matrix of the estimate $\hat{c}(\mathfrak{h})$ (or an approximation thereof) is a full-rank matrix. Finally, Section 12.3 provides an analysis of solutions under weaker assumptions.

12.1. Preliminaries

As a starting point, recall equation (10.1), i.e.

$$\eta_i = h_i(\{y(t_j; c^\star)\}_{j=1}^{n_t}, c^\star) + \epsilon_i, \tag{12.1}$$

where η_i represents a measurement value, t_j are the measurement times, $c^\star \in \mathbb{R}^{n_c}$ are the correct parameters and $y(t_j; c^\star)$ are the values of the state vector at the measurement times for the correct parameters $c^\star$. If the dynamic model that defines the function $y(t; c)$ and if all measurement models h_i (for $1 \leqslant i \leqslant n_h$) are correct, then ϵ_i, $1 \leqslant i \leqslant n_h$, can be interpreted as measurement errors.

Recall further that parameters are often subject to constraints of the form (10.15b), i.e.

$$g_i(\{y(t_j; c)\}_{j=1}^{n_t}, c) = 0, \tag{12.2}$$

for $1 \leqslant i \leqslant n_g$.

For simplicity of notation, it is convenient to regard h_i and g_i as a function of the parameters alone and to use a vector notation, i.e.

$$\eta = h(c^\star) + \epsilon \tag{12.3}$$

and

$$g(c) = 0. \tag{12.4}$$

In these equations, $\epsilon = (\epsilon_1, \ldots, \epsilon_{n_h})^T$ and $\eta = (\eta_1, \ldots, \eta_{n_h})^T$, and g and h are vector-valued function whose components are the functions g_i and h_i.

The measurement errors $\epsilon = (\epsilon_1, \ldots, \epsilon_{n_h})^T$ – and thus the measurements $\eta = (\eta_1, \ldots, \eta_{n_h})^T$ – denote only one specific realization of the corresponding random variables $\mathfrak{e}$ and $\mathfrak{h}$. For the special case that the measurement errors vanish, i.e. $\epsilon = 0$, it is useful to denote the *correct measurements* by

$$\eta^\star := h(c^\star). \tag{12.5}$$

A maximum likelihood estimate $\hat{c}(\eta)$ for a specific realization η of the measurements is defined by the property that it is a maximizer of the likelihood function, see Section 10.3. If the measurement errors are normally distributed with known and regular covariance matrix $\mathbf{V}_\epsilon$, then $\hat{c}(\eta)$ is a solution of the following constrained least-squares problem:

$$\min_c \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (h(c) - \eta) \right\|_2^2 \tag{12.6a}$$

$$\text{s.t.} \quad g(c) = 0, \tag{12.6b}$$

see Theorem 10.3.

Occasionally, it is useful in this chapter to use the notations

$$F_1(c) := \mathbf{V}_\epsilon^{-\frac{1}{2}} (h(c) - \eta) \quad \text{and} \quad F_2(c) := g(c). \tag{12.7}$$

The Jacobian matrices of the functions F_1 and F_2 are denoted by

$$\mathbf{J}_i(c) := \left. \frac{\partial F_i(c')}{\partial c'} \right|_{c'=c}. \tag{12.8}$$

Further, let

$$F(c) := \begin{pmatrix} F_1(c) \\ F_2(c) \end{pmatrix} \quad \text{and} \quad \mathbf{J}(c) := \begin{pmatrix} \mathbf{J}_1(c) \\ \mathbf{J}_2(c) \end{pmatrix}. \tag{12.9}$$

12.2. Statistical Analysis based on Covariance Matrices

12.2.1. Statistical Distribution of Parameter Estimates

Linear Constrained Least-Squares Problems

A rigorous result for the statistical distribution of a maximum likelihood estimate $\hat{c}(\eta)$ can be obtained under the condition that the functions h_i (and the constraint functions g_i, if present), are both linear in c.

Theorem 12.1 (Covariance of Parameter Estimates for Linear Problems and Normally Distributed Measurements)

Let $\mathfrak{h} = h(c^\star) + \mathfrak{e}$, and correspondingly, let a specific realization of the random variables be given by $\eta = h(c^\star) + \epsilon$. Let further $g_i(c) = 0$ for $1 \leqslant i \leqslant n_g$, $n_g < n_c$, denote equality constraints on the parameters.

Assume that h and g are linear, i.e. $h(c) = \mathbf{B}_1 c + b_1$ and $g(c) = \mathbf{B}_2 c + b_2$, and assume that the measurement errors are normally distributed with expected value zero and covariance matrix $\mathbf{V}_\epsilon$, $\mathfrak{e} \sim \mathcal{N}(0, \mathbf{V}_\epsilon)$, with $\mathbf{V}_\epsilon$ being a regular matrix (which implies that a matrix $\mathbf{V}_\epsilon^{-\frac{1}{2}}$ exists such that $\mathbf{V}_\epsilon^{-\frac{1}{2}} \mathbf{V}_\epsilon^{-\frac{1}{2}} = \mathbf{V}_\epsilon^{-1}$). Further, let the regularity assumptions $\operatorname{rank}(\mathbf{B}_2) = n_g$ and $\operatorname{rank}\left(\begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}\right) = n_c$ be fulfilled.

Then the following holds:

1. *The maximum likelihood estimate $\hat{c}(\eta)$ is the solution of the following linear constrained least-*

squares problem:

$$\min_c \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 c + b_1 - \eta) \right\|_2^2, \tag{12.10a}$$

$$s.t. \quad \mathbf{B}_2 c + b_2 = 0. \tag{12.10b}$$

2. *The maximum likelihood estimate $\hat{c}(\eta)$ can be expressed as*

$$\hat{c}(\eta) = -\mathbf{J}^+ F(0), \tag{12.11}$$

where $\mathbf{J}^+$ is the generalized inverse of the (constant) Jacobian matrix $\mathbf{J} \equiv \mathbf{J}(c)$.

3. *The maximum likelihood estimate $c(\mathfrak{h})$ is normally distributed; more precisely it holds that*

$$\hat{c}(\mathfrak{h}) \sim \mathcal{N}(c^\star, \mathbf{V}_c), \quad with \quad \mathbf{V}_c := \mathbf{J}^+ \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{J}^+)^T. \tag{12.12}$$

Proof (see also Bock [39], page 133f)
Equation (12.10) is only the special form of problem (12.6), and thus assertion 1 follows directly from Theorem 10.3. Further, under the given assumptions, also Theorem 11.4 can be applied, which yields that $\hat{c}(\eta) = -\mathbf{J}^+ F(0)$ and thus assertion 2. Hence, it holds for the random counterpart $\hat{c}(\mathfrak{h})$ of $\hat{c}(\eta)$ that

$$\hat{c}(\mathfrak{h}) = -\mathbf{J}^+ \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} (b_1 - \mathfrak{h}) \\ b_2 \end{pmatrix}, \tag{12.13}$$

i.e. $\hat{c}(\mathfrak{h})$ is a linear transformation of $\mathfrak{h}$. From this it follows immediately that $\hat{c}(\mathfrak{h})$ is normally distributed. The expected value is given by $\mathbb{E}(\hat{c}(\mathfrak{h})) = \hat{c}(\eta^\star) = c^\star$, and the covariance follows from

$$\begin{aligned} \mathbb{V}(\hat{c}(\mathfrak{h})) =& \mathbb{E}((\hat{c}(\mathfrak{h}) - c^\star)(\hat{c}(\mathfrak{h}) - c^\star)^T) \\ =& \mathbf{J}^+ \mathbb{E}\left(\begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} (\eta^\star - \mathfrak{h}) \\ 0 \end{pmatrix} \begin{pmatrix} (\eta^\star - \mathfrak{h})^T \mathbf{V}_\epsilon^{-\frac{1}{2}} & 0 \end{pmatrix} \right) (\mathbf{J}^+)^T \end{aligned} \tag{12.14}$$

and from $\mathbb{E}((\eta^\star - \mathfrak{h})(\eta^\star - \mathfrak{h})^T) = \mathbf{V}_\epsilon$. ■

Remark 12.2 (Covariance of Parameter Estimates for Jacobian in Reduced Form)
Let, in addition to the assumptions of Theorem 12.1, the matrix $\mathbf{J}$ be given in reduced form (Definition 11.7), i.e.

$$\mathbf{J} = \begin{pmatrix} \mathbf{A} & \mathbf{S} \\ \mathbf{L} & \mathbf{0} \end{pmatrix}, \tag{12.15}$$

where $\mathbf{A} \in \mathbb{R}^{n_h \times n_g}$, $\mathbf{S} \in \mathbb{R}^{n_h \times (n_c - n_g)}$ is a diagonal matrix with diagonal elements $s_1 \geqslant s_2 \geqslant \cdots \geqslant s_{n_c - n_g}$, and $\mathbf{L} \in \mathbb{R}^{n_g \times n_g}$ is a regular lower triangular matrix.

Then the covariance matrix takes the special form

$$\mathbf{V}_c = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^\dagger (\mathbf{S}^\dagger)^T \end{pmatrix}. \tag{12.16}$$

Accordingly, if the entries of $\mathbf{S}$ are denoted by $s_1, \ldots, s_{n_c - n_g}$, then the variance of the parameter estimates $\hat{c}_1(\mathfrak{h}), \ldots, \hat{c}_{n_g}(\mathfrak{h})$ is 0 (because these parameters are determined by the equality constraints), and the variance of the parameter estimates $\hat{c}_{n_g+1}(\mathfrak{h}), \ldots, \hat{c}_{n_c}(\mathfrak{h})$ is given by $s_1^{-2}, \ldots, s_{n_c - n_g}^{-2}$. It is recalled that this interpretation of s_i^{-2} has already been used as a motivation for using equation (11.27) in the practical rank decision of the Generalized Gauss-Newton method (see discussion in Subsection 11.2.4).

Nonlinear Constrained Least-Squares Problems

For the case that h and g are general nonlinear (but smooth) functions, the maximum likelihood estimate $c(\mathfrak{h})$ is not any longer a normally distributed random variable. However, if the Jacobian

matrices $\mathbf{J}_1(\hat{c}(\eta))$ and $\mathbf{J}_2(\hat{c}(\eta))$ are, for specific measurements η, such that $\operatorname{rank}(\mathbf{J}_2(\hat{c}(\eta))) = n_g$ and $\operatorname{rank}(\mathbf{J}(\hat{c}(\eta))) = n_c$, then the expression

$$\tilde{\mathbf{V}}_c := \mathbf{J}^+(\hat{c}(\eta)) \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{J}^+)^T(\hat{c}(\eta)) \tag{12.17}$$

can still be used as an approximation of the covariance of the parameter estimates.

A recent result is that even the assumption of a correct model, i.e. the existence of $c^\star$ such that $h(c^\star) = \eta^\star$, can be dropped (see Hoffmann [149]). More precisely, the use of the above expression as an approximation of the covariance of parameter estimates is justified whenever the product of "incorrectness" and "nonlinearity" of the model – in suitable measures – is sufficiently small.

12.2.2. Distribution of Least-Squares Sum

The next issue that is considered is the distribution of the least-squares sum $\left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (h(\hat{c}(\mathfrak{h})) - \mathfrak{h}) \right\|_2^2$. As in Subsection 12.2.1, a rigorous result can be obtained if the functions h and g are linear, in which case the maximum likelihood estimate is given as solution of the linear constrained least-squares problems (12.10).

Theorem 12.3 (Distribution of Least-Squares Sum for Linear Problems and Normally Distributed Measurements)

Consider the linear constrained least-squares problem (12.10), where η represents a specific realization of the random number $\mathfrak{h} = \eta^\star + \mathfrak{e}$, with $\mathfrak{e} \sim \mathcal{N}(0, \mathbf{V}_\epsilon)$. Further, let the regularity conditions $\operatorname{rank}(\mathbf{B}_2) = n_g$ and $\operatorname{rank}\left(\begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} \right) = n_c$ be fulfilled.

Then it holds that the least-squares sum, evaluated at the solution $\hat{c}(\mathfrak{h})$, is χ^2-distributed with $n_h - (n_c - n_g)$ degrees of freedom, in short

$$\left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 \hat{c}(\mathfrak{h}) + b_1 - \mathfrak{h}) \right\|_2^2 \sim \chi^2_{n_h - (n_c - n_g)}. \tag{12.18}$$

Proof
The regularity assumption $\operatorname{rank}(\mathbf{B}_2) = n_g$ ensures that the (constant) Jacobian

$$\mathbf{J} \equiv \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} \tag{12.19}$$

can be brought into reduced form (Definition 11.7) by applying orthogonal transformations, see Sections 11.2 and 13.3. Therefore, without loss of generality, it is sufficient to prove the theorem for a Jacobian in the reduced form

$$\mathbf{J} = \begin{pmatrix} \mathbf{A} & \mathbf{S} \\ \mathbf{L} & \mathbf{0} \end{pmatrix}, \tag{12.20}$$

where $\mathbf{S}$ is a diagonal matrix and $\mathbf{L}$ is a regular lower triangular matrix.

The linear constrained least-squares problem then reads

$$\min_c \left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} c + \mathbf{V}_\epsilon^{-\frac{1}{2}} (b_1 - \eta) \right\|_2^2, \tag{12.21a}$$

$$\text{s.t.} \quad \begin{pmatrix} \mathbf{L} & \mathbf{0} \end{pmatrix} c + b_2 = 0. \tag{12.21b}$$

In view of equation (11.24), the solution $\hat{c}(\eta)$ can be written as

$$\hat{c}(\eta) = - \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \mathbf{S}^\dagger & -\mathbf{S}^\dagger \mathbf{A} \mathbf{L}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} (b_1 - \eta) \\ b_2 \end{pmatrix}, \tag{12.22}$$

where $\mathbf{S}^\dagger \in \mathbb{R}^{(n_c - n_g) \times n_h}$ is the Moore-Penrose pseudoinverse of $\mathbf{S}$. For the least-squares sum in the

solution, it then follows that

$$\left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \hat{c}(\mathfrak{h}) + \mathbf{V}_\epsilon^{-\frac{1}{2}}(b_1 - \mathfrak{h}) \right\|_2^2 = \left\| - \begin{pmatrix} \mathbf{S}\mathbf{S}^\dagger & \mathbf{A}\mathbf{L}^{-1} - \mathbf{S}\mathbf{S}^\dagger\mathbf{A}\mathbf{L}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}}(b_1 - \mathfrak{h}) \\ b_2 \end{pmatrix} + \mathbf{V}_\epsilon^{-\frac{1}{2}}(b_1 - \mathfrak{h}) \right\|_2^2$$
$$= \left\| (\mathbf{1} - \mathbf{S}\mathbf{S}^\dagger) \left(\mathbf{V}_\epsilon^{-\frac{1}{2}}(b_1 - \mathfrak{h}) - \mathbf{A}\mathbf{L}^{-1} b_2 \right) \right\|_2^2. \tag{12.23}$$

It further holds that $\mathfrak{h} = \eta^\star + \mathfrak{e} = \mathbf{B}_1 c^\star + b_1 + \mathfrak{e}$ and thus

$$\mathbf{V}_\epsilon^{-\frac{1}{2}}(b_1 - \mathfrak{h}) = \mathbf{V}_\epsilon^{-\frac{1}{2}}(-\mathbf{B}_1 c^\star - \mathfrak{e}). \tag{12.24}$$

Furthermore, $c^\star = \hat{c}(\eta^\star)$ and $\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathbf{B}_1 = \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix}$ yield

$$-\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathbf{B}_1 c^\star = \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \mathbf{S}^\dagger & -\mathbf{S}^\dagger\mathbf{A}\mathbf{L}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}}(-\mathbf{B}_1 c^\star) \\ b_2 \end{pmatrix}$$
$$= -\mathbf{S}\mathbf{S}^\dagger \mathbf{V}_\epsilon^{-\frac{1}{2}}\mathbf{B}_1 c^\star + \mathbf{A}\mathbf{L}^{-1} b_2 - \mathbf{S}\mathbf{S}^\dagger\mathbf{A}\mathbf{L}^{-1} b_2, \tag{12.25}$$

and hence

$$(\mathbf{1} - \mathbf{S}\mathbf{S}^\dagger)\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathbf{B}_1 c^\star = -(\mathbf{1} - \mathbf{S}\mathbf{S}^\dagger)\mathbf{A}\mathbf{L}^{-1} b_2. \tag{12.26}$$

Eventually, inserting this relation and equation (12.24) into equation (12.23) yields

$$\left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \hat{c}(\mathfrak{h}) + \mathbf{V}_\epsilon^{-\frac{1}{2}}(b_1 - \mathfrak{h}) \right\|_2^2 = \left\| (\mathbf{1} - \mathbf{S}\mathbf{S}^\dagger)\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathfrak{e} \right\|_2^2. \tag{12.27}$$

The matrix $\mathbf{1} - \mathbf{S}\mathbf{S}^\dagger$ is a diagonal matrix, and the diagonal contains $n_c - n_g$ entries that are equal to 0 and $n_h - (n_c - n_g)$ entries that are equal to 1. Furthermore, $\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathfrak{e} \sim \mathcal{N}(0, \mathbf{1})$. The least-squares sum $\left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 \hat{c}(\mathfrak{h}) + b_1 - \mathfrak{h}) \right\|_2^2$ hence is a sum of squares of $n_h - (n_c - n_g)$ standard normally distributed random variables, which by definition is a χ^2 distribution with $n_h - (n_c - n_g)$ degrees of freedom. ■

If the functions $h(c)$ and (or) $g(c)$ are nonlinear, the least-squares sum $\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (h(\hat{c}(\mathfrak{h})) - \mathfrak{h}) \|_2^2$ is only approximately $\chi^2_{n_h - (n_c - n_g)}$-distributed.

12.2.3. Confidence Regions

Having obtained parameter estimates $\hat{c}(\eta)$ for specific measurement data η, it is desirable to make statistical inference on the values of the correct parameters, i.e. on $c^\star$. For illustration purposes, it is appropriate to consider first the case of a linear unconstrained least-squares problem for a scalar parameter, i.e. $c \in \mathbb{R}$.

Linear Unconstrained Least-Squares Problems, Scalar Parameter

In the special case of linear unconstrained least-squares problems, it follows from Theorem 12.1 and Remark 12.2 that the distribution of the estimate is $\hat{c}(\mathfrak{h}) \sim \mathcal{N}(c^\star, s_1^{-2})$ (in the notation of the reduced form). It is well-known from the properties of the normal distribution that the probability for $\hat{c}(\mathfrak{h})$ to be in the interval $[c^\star - s_1^{-1}, c^\star + s_1^{-1}]$ is given by

$$P(\hat{c}(\mathfrak{h}) \in [c^\star - s_1^{-1}, c^\star + s_1^{-1}]) \approx 0.6827. \tag{12.28}$$

Vice versa, it is also possible to write

$$P(c^\star \in [\hat{c}(\mathfrak{h}) - s_1^{-1}, \hat{c}(\mathfrak{h}) + s_1^{-1}]) \approx 0.6827. \tag{12.29}$$

Having in mind that the measurements η are only one specific realization of the random variable $\mathfrak{h}$, it is possible to use the sloppy formulation that "the interval $[\hat{c}(\eta) - s_1^{-1}, \hat{c}(\eta) + s_1^{-1}]$ will enclose the correct parameter $c^\star$ with a probability that is approximately 0.6827".

It is immediately clear that the interval $[\hat{c}(\eta) - \nu s_1^{-1}, \hat{c}(\eta) + \nu s_1^{-1}]$ with $\nu > 1$ (with $\nu < 1$) has a higher (lower) probability to enclose the correct parameter $c^\star$. More precisely, by a specific choice of ν, any desired probability α, $0 < \alpha < 1$, can be obtained.

One way to define the exact size of the interval that corresponds to a given probability α is the approach via so-called *δ-indifference regions*. This approach is discussed in the following in the context of linear constrained least-squares problems and several parameters.

Linear Constrained Least-Squares Problems, Several Parameters

Consider the *δ-indifference region* around the correct parameters $c^\star$, i.e. that region in which the constraints are fulfilled and the least-squares sum is less than or equal to δ for the correct measurements $\eta^\star$:

$$\Upsilon_{c^\star}(\delta) := \left\{ c^\star + \Delta c \mid \mathbf{B}_2(c^\star + \Delta c) + b_2 = 0 \quad \wedge \quad \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1(c^\star + \Delta c) + b_1 - \eta^\star) \right\|_2^2 \leqslant \delta \right\}. \tag{12.30}$$

The question is how to choose δ such that this deterministic region contains the random maximum likelihood estimate $\hat{c}(\mathfrak{h})$ with a given probability α. The answer is given by the following lemma.

Lemma 12.4 (Confidence Region for Estimates)

Consider a linear constrained least-squares problem

$$\min_c \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 c + b_1 - \eta) \right\|_2^2, \tag{12.31a}$$

$$s.t. \quad \mathbf{B}_2 c + b_2 = 0, \tag{12.31b}$$

and assume that $\eta = \eta^\star + \mathfrak{e}$, $\mathfrak{e} \sim \mathcal{N}(0, \mathbf{V}_\epsilon)$. Further, let the regularity conditions $\operatorname{rank}(\mathbf{B}_2) = n_g$, $\operatorname{rank}\left(\begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} \right) = n_c$ *be fulfilled.*

Then it holds that the random maximum likelihood estimate $\hat{c}(\mathfrak{h})$ is contained in the region $\Upsilon_{c^\star}(q(\chi^2_{n_c - n_g}, \alpha))$ with probability α, i.e. formally:

$$P\left[\hat{c}(\mathfrak{h}) \in \Upsilon_{c^\star}(q(\chi^2_{n_c - n_g}, \alpha)) \right] = \alpha. \tag{12.32}$$

Herein, $q(\chi^2_{n_c - n_g}, \alpha)$ represents the quantile of the $\chi^2_{n_c - n_g}$ distribution to probability α.

Proof

Without loss of generality, it can be assumed that the Jacobian is given in reduced form (see discussion in the proof of Theorem 12.3). It then holds that $\mathbf{B}_2 = \begin{pmatrix} \mathbf{L} & \mathbf{0} \end{pmatrix}$ and $\mathbf{V}_\epsilon^{-\frac{1}{2}} \mathbf{B}_1 = \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix}$, with $\mathbf{L}$ being a regular lower triangular matrix and $\mathbf{S}$ being a diagonal matrix. Furthermore, it holds that $\mathbf{B}_1 c^\star + b_1 = \eta^\star$, and thus

$$\Upsilon_{c^\star}(\delta) = \left\{ c^\star + \Delta c \mid \begin{pmatrix} \mathbf{L} & \mathbf{0} \end{pmatrix} (c^\star + \Delta c) + b_2 = 0 \quad \wedge \quad \left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \Delta c \right\|_2^2 \leqslant \delta \right\}. \tag{12.33}$$

The maximum likelihood estimate $\hat{c}(\mathfrak{h})$ is, according to Remark 12.2, distributed as $\hat{c}(\mathfrak{h}) \sim \mathcal{N}(c^\star, \mathbf{V}_c)$ with

$$\mathbf{V}_c = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^\dagger (\mathbf{S}^\dagger)^T \end{pmatrix}. \tag{12.34}$$

Consequently, $\Delta c(\mathfrak{h}) := \hat{c}(\mathfrak{h}) - c^\star \sim \mathcal{N}(0, \mathbf{V}_c)$. In view of equation (12.33), the distribution of $\begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \Delta c(\mathfrak{h})$ is investigated:

$$\begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \Delta c(\mathfrak{h}) \sim \mathcal{N}\left(0, \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \mathbf{V}_c \begin{pmatrix} \mathbf{A}^T \\ \mathbf{S}^T \end{pmatrix} \right) = \mathcal{N}\left(0, \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}\mathbf{S}^\dagger (\mathbf{S}^\dagger)^T \mathbf{S}^T \end{pmatrix} \right). \tag{12.35}$$

The matrix $\mathbf{S}\mathbf{S}^\dagger (\mathbf{S}^\dagger)^T \mathbf{S}^T$ is of dimension $n_h \times n_h$, and it is diagonal with the first $n_c - n_g$ entries being 1 and the remaining entries being 0. From this it follows that $\left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \Delta c(\mathfrak{h}) \right\|_2^2 \sim \chi^2_{n_c - n_g}$.

Therefore, if the δ-indifference region $\Upsilon_{c^\star}(\delta)$ should contain the random maximum likelihood estimate $\hat{c}(\mathfrak{h})$ with probability α, the *indifference parameter* δ has to be chosen as the quantile $q(\chi^2_{n_c-n_g}, \alpha)$ of the $\chi^2_{n_c-n_g}$ distribution to probability α. ■

For statistical inference about the correct parameters $c^\star$, the following δ-indifference region is considered:

$$\Upsilon_{\hat{c}(\mathfrak{h})}(\delta) := \Big\{ \hat{c}(\mathfrak{h}) + \Delta c \mid \mathbf{B}_2(\hat{c}(\mathfrak{h}) + \Delta c) + b_2 = 0 \quad \wedge \\ \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1(\hat{c}(\mathfrak{h}) + \Delta c) + b_1 - \mathfrak{h}) \right\|_2^2 - \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 \hat{c}(\mathfrak{h}) + b_1 - \mathfrak{h}) \right\|_2^2 \leqslant \delta \Big\}. \tag{12.36}$$

For specific measurements (i.e., replace $\mathfrak{h}$ by η) this region comprises all those points c for which the constraints are fulfilled and the objective function $\left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 c + b_1 - \eta) \right\|_2^2$ differs from the value in the solution $\hat{c}(\eta)$ by less than δ.

The question is how to choose δ such that the random region $\Upsilon_{\hat{c}(\mathfrak{h})}(\delta)$ contains the deterministic point $c^\star$ with a given probability α. Once again, the answer is given by the quantile $q(\chi^2_{n_c-n_g}, \alpha)$ of the $\chi^2_{n_c-n_g}$ distribution to probability α, as stated by the following lemma.

Lemma 12.5 (Confidence Region for Correct Paramaters)

Consider a linear constrained least-squares problem

$$\min_c \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 c + b_1 - \eta) \right\|_2^2, \tag{12.37a}$$

$$s.t. \quad \mathbf{B}_2 c + b_2 = 0, \tag{12.37b}$$

and assume that the assumptions of Lemma 12.4 are fulfilled.

Then it holds that the correct parameters $c^\star$ are contained in the region $\Upsilon_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g}, \alpha))$ with probability α, i.e. formally:

$$P\left[c^\star \in \Upsilon_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g}, \alpha)) \right] = \alpha. \tag{12.38}$$

Proof
It is assumed – without loss of generality – that the Jacobian $\mathbf{J} = \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}$ is given in reduced form, i.e. $\mathbf{V}_\epsilon^{-\frac{1}{2}} \mathbf{B}_1 = \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix}$ and $\mathbf{B}_2 = \begin{pmatrix} \mathbf{L} & \mathbf{0} \end{pmatrix}$.

Consider the distribution of

$$\left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1(\hat{c}(\mathfrak{h}) + \Delta c(\mathfrak{h})) + b_1 - \mathfrak{h}) \right\|_2^2 - \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 \hat{c}(\mathfrak{h}) + b_1 - \mathfrak{h}) \right\|_2^2$$

for $\Delta c(\mathfrak{h}) := c^\star - \hat{c}(\mathfrak{h})$. From $\mathfrak{h} = \eta^\star + \mathfrak{e}$ and $\mathbf{B}_1 c^\star + b_1 = \eta^\star$ it follows for the first term that

$$\left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1(\hat{c}(\mathfrak{h}) + \Delta c(\mathfrak{h})) + b_1 - \mathfrak{h}) \right\|_2^2 = \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathfrak{e} \right\|_2^2. \tag{12.39}$$

Further, for the second term it holds that

$$\begin{aligned} \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 \hat{c}(\mathfrak{h}) + b_1 - \mathfrak{h}) \right\|_2^2 &= \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} \left(-\mathbf{B}_1 \Delta c(\mathfrak{h}) - \mathfrak{e} \right) \right\|_2^2 \\ &= \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathfrak{e} \right\|_2^2 + \left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} \Delta c(\mathfrak{h}) \right\|_2^2 + 2 \Delta c(\mathfrak{h})^T \begin{pmatrix} \mathbf{A}^T \\ \mathbf{S}^T \end{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathfrak{e}. \end{aligned} \tag{12.40}$$

With

$$\Delta c(\mathfrak{h}) = - \begin{pmatrix} \mathbf{0} \\ \mathbf{S}^\dagger \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathfrak{e} \end{pmatrix}, \tag{12.41}$$

and with $(\mathbf{S}^\dagger)^T\mathbf{S}^T = \mathbf{S}\mathbf{S}^\dagger$, it follows that

$$\left\|\mathbf{V}_\epsilon^{-\frac{1}{2}}(\mathbf{B}_1\hat{c}(\mathfrak{h}) + b_1 - \mathfrak{h})\right\|_2^2 = \left\|\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathfrak{e}\right\|_2^2 - \left\|\mathbf{S}\mathbf{S}^\dagger\mathbf{V}_\epsilon^{-\frac{1}{2}}\mathfrak{e}\right\|_2^2. \tag{12.42}$$

Since $\mathbf{S}\mathbf{S}^\dagger$ is an $n_h \times n_h$ diagonal matrix with $n_c - n_g$ entries that are equal to 1 and $n_h - (n_c - n_g)$ entries that are equal to 0, it follows that

$$\left\|\mathbf{V}_\epsilon^{-\frac{1}{2}}(\mathbf{B}_1(\hat{c}(\mathfrak{h}) + \Delta c(\mathfrak{h})) + b_1 - \mathfrak{h})\right\|_2^2 - \left\|\mathbf{V}_\epsilon^{-\frac{1}{2}}(\mathbf{B}_1\hat{c}(\mathfrak{h}) + b_1 - \mathfrak{h})\right\|_2^2 \sim \chi^2_{n_c-n_g}. \tag{12.43}$$

Thus, the proof is completed. ■

It is remarked that the Lemmas 12.4 and 12.5 are consistent with the equations (12.28) and (12.29), because for an unconstrained least-squares problem and a scalar parameter c, one has $q(\chi_1^2, 0.6827) = 1$ and

$$\begin{aligned}&\Upsilon_{c^\star}(1) = \{c^\star + \Delta c \mid |s_1\Delta c|_2^2 \leqslant 1\}\\ &\Rightarrow |\Delta c| = s_1^{-1} \Rightarrow \Upsilon_{c^\star}(1) = [c^\star - s_1^{-1}, c^\star + s_1^{-1}]\end{aligned} \tag{12.44}$$

(and analgously for $\Upsilon_{\hat{c}(\mathfrak{h})}(1)$).

Nonlinear Constrained Least-Squares Problems

Eventually, a general least-squares problem of the form

$$\min_c \ \|F_1(c)\|_2^2 \tag{12.45a}$$

$$\text{s.t. } F_2(c) = 0, \tag{12.45b}$$

with $F_1(c) = \mathbf{V}_\epsilon^{-\frac{1}{2}}(h(c) - \mathfrak{h})$ and $F_2(c) = g(c)$ is considered.

Consider, in analogy to equation (12.36), the following definition:

$$\begin{aligned}\Upsilon^N_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g}, \alpha)) := \Big\{&\hat{c}(\mathfrak{h}) + \Delta c \mid F_2(\hat{c}(\mathfrak{h}) + \Delta c) = 0 \quad \wedge\\ &\|F_1(\hat{c}(\mathfrak{h}) + \Delta c)\|_2^2 - \|F_1(\hat{c}(\mathfrak{h}))\|_2^2 \leqslant q(\chi^2_{n_c-n_g}, \alpha)\Big\}.\end{aligned} \tag{12.46}$$

For specific measurements η, this indifference region has the shape of a level set of $\|F_1(c)\|_2^2$, restricted on the subset of $\mathbb{R}^{n_c}$ for which the equality constraints are fulfilled. Since the set $\Upsilon^N_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g}, \alpha))$ is defined with the nonlinear functions F_1 and F_2, it is in the following called a *nonlinear indifference region*. This terminology is also characterized by using the superscript N.

Further, by using linear approximations of the functions F_1 and F_2 (around $\hat{c}(\mathfrak{h})$), the following *linearized indifference region* is defined (superscript L):

$$\begin{aligned}\Upsilon^L_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g}, \alpha)) := \Big\{&\hat{c}(\mathfrak{h}) + \Delta c \mid F_2(\hat{c}(\mathfrak{h})) + \mathbf{J}_2(\hat{c}(\mathfrak{h}))\Delta c = 0 \quad \wedge\\ &\|F_1(\hat{c}(\mathfrak{h})) + \mathbf{J}_1(\hat{c}(\mathfrak{h}))\Delta c\|_2^2 - \|F_1(\hat{c}(\mathfrak{h}))\|_2^2 \leqslant q(\chi^2_{n_c-n_g}, \alpha)\Big\}.\end{aligned} \tag{12.47}$$

For specific measurements η, this linearized indifference region has the elliptic shape of a level set of $\|F_1(\hat{c}(\eta)) + \mathbf{J}_1(\hat{c}(\eta))\Delta c\|_2^2$, constrained to the hyperplane of dimension $n_c - n_g$ that is defined by the linearized equality constraints.

The set $\Upsilon^L_{\hat{c}(\eta)}(q(\chi^2_{n_c-n_g}, \alpha))$ can also be represented in an alternative way that makes use of the generalized inverse.

Lemma 12.6 (Alternative Representation of Linear Indifference Region)

Let $\hat{c}(\eta)$ be the solution of the nonlinear constrained least-squares problem (12.45) for specific measurement data η. Assume that the Jacobian fulfills, in the solution, the regularity conditions $\operatorname{rank}(\mathbf{J}_2(\hat{c}(\eta))) = n_g$ *and* $\operatorname{rank}(\mathbf{J}(\hat{c}(\eta))) = n_c$.

Then it holds that

$$\Upsilon^L_{\hat{c}(\eta)}(q(\chi^2_{n_c-n_g},\alpha)) = \left\{\hat{c}(\eta)+\Delta c \mid \Delta c = -\mathbf{J}^+(\hat{c}(\eta))\begin{pmatrix}\Delta y\\ 0\end{pmatrix},\quad \|\Delta y\|_2^2 \leqslant q(\chi^2_{n_c-n_g},\alpha)\right\}. \quad (12.48)$$

Proof
See Bock [39], p. 136f, or Bock, Kostina, and Kostyukova [40]. ■

Due to the nonlinearity of F_1 and F_2, the distribution of the maximum likelihood estimate $\hat{c}(\mathfrak{h})$ is generally unknown. Consequently, also the distributions of the expressions that occur in the definitions of $\Upsilon^N_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g},\alpha))$ and $\Upsilon^L_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g},\alpha))$ are unknown, and the quantile $q(\chi^2_{n_c-n_g},\alpha)$ can only be regarded as an approximation of the "correct" indifference parameter δ (for the probability α). Therefore, both indifference regions contain the correct parameters only approximately with probability α:

$$P(c^\star \in \Upsilon^N_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g},\alpha))) \approx \alpha \quad (12.49a)$$

$$P(c^\star \in \Upsilon^L_{\hat{c}(\mathfrak{h})}(q(\chi^2_{n_c-n_g},\alpha))) \approx \alpha. \quad (12.49b)$$

The less nonlinear the functions F_1 and F_2 are, the better these approximations become.

12.2.4. Confidence Intervals

In the previous subsection, indifference regions were derived that enclose the correct parameters $c^\star$ either exactly with probability α (in the case of linear functions F_1 and F_2) or approximately with probability α (in the case of nonlinear functions F_1 and F_2). The next lemma shows that the indifference region $\Upsilon^L_{\hat{c}(\eta)}(q(\chi^2_{n_c-n_g},\alpha))$ as defined by equation (12.47) is "exactly" contained in a box.

Lemma 12.7 ("Exact" Bound for Linearized Indifference Region)

Let the assumptions of Lemma 12.6 be fulfilled. Then it holds that

$$\Upsilon^L_{\hat{c}(\eta)}(q(\chi^2_{n_c-n_g},\alpha)) \subset \Omega := \{\hat{c}(\eta)+\Delta c \mid |\Delta c_i| \leqslant \theta_i \quad \textit{for} \quad 1 \leqslant i \leqslant n_c\} \quad (12.50)$$

with $\theta_i := q(\chi^2_{n_c-n_g},\alpha)\cdot(\tilde{\mathbf{V}}_c)_{i,i}$, *where*

$$\tilde{\mathbf{V}}_c := \mathbf{J}^+(\hat{c}(\eta))\begin{pmatrix}\mathbf{1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}(\mathbf{J}^+)^T(\hat{c}(\eta)). \quad (12.51)$$

Moreover, Ω *contains* $\Upsilon^L_{\hat{c}(\eta)}(q(\chi^2_{n_c-n_g},\alpha))$ *exactly in the sense that*

$$\max_{c\in\Upsilon^L_{\hat{c}(\eta)}(q(\chi^2_{n_c-n_g},\alpha))} |c_i - \hat{c}_i(\eta)| = \theta_i. \quad (12.52)$$

Proof
See Bock [39], p. 137, or Bock, Kostina, and Kostyukova [40]. ■

Please note that, under the assumptions of Lemma 12.4, the indifference region $\Upsilon^L_{\hat{c}(\eta)}(q(\chi^2_{n_c-n_g},\alpha))$ encloses the correct parameters $c^\star$ exactly with probability α. In this case, $[\hat{c}_i-\theta_i,\hat{c}_i+\theta_i]$ is thus identified as confidence interval for the individual parameter c_i. If the functions F_1 and/or F_2 are nonlinear, $[\hat{c}_i-\theta_i,\hat{c}_i+\theta_i]$ can be used as an approximation of the confidence interval.

12.3. Analysis of "Rank-Deficient Solutions"

In Section 12.2, results were presented on the statistical distribution of maximum likelihood parameter estimates $\hat{c}(\mathfrak{h})$, on the statistical distribution of the least-squares sum, on confidence regions and on confidence intervals. All rigorous results thereby relied on the following assumptions:

- The functions $h(c)$ and $g(c)$ in the constrained least-squares problem (12.6) are linear.

- The measurements η in problem (12.6) are realizations of the random variables $\mathfrak{h} = \eta^\star + \mathfrak{e}$, and $\mathfrak{e} \sim \mathcal{N}(0, \mathbf{V}_\epsilon)$.
- The (constant) Jacobians $\mathbf{J}_1 \equiv \mathbf{J}_1(c)$ and $\mathbf{J}_2 \equiv \mathbf{J}_2(c)$ are such that the rank conditions $\operatorname{rank}(\mathbf{J}_2) = n_g$ and $\operatorname{rank}(\mathbf{J}) = n_c$ are fulfilled.

For nonlinear functions h and g, the results for the linear case are still approximately valid provided that the (non-constant) Jacobians $\mathbf{J}_1(c)$ and $\mathbf{J}_2(c)$ are such that the rank conditions are fulfilled in the solution, i.e. for $c = \hat{c}(\eta)$.

Unfortunately, it is quite frequent in practice that the rank condition $\operatorname{rank}(\mathbf{J}(\hat{c}(\eta))) = n_c$ is violated or that the matrix is (very) ill-conditioned. This situation may occur, e.g., from overparameterization of the dynamic process. The analysis and interpretation of solutions in this case is the subject of this section.

Linear Constrained Least-Squares Problems

It is instructive to regard, at first, the case of a linear constrained least-squares problem of the form

$$\min_c \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} (\mathbf{B}_1 c + b_1 - \eta) \right\|_2^2 \tag{12.53a}$$

$$\text{s.t.} \quad \mathbf{B}_2 c + b_2 = 0, \tag{12.53b}$$

for which the rank condition

$$\operatorname{rank}(\mathbf{J}) = n_c, \quad \mathbf{J} := \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix} \tag{12.54}$$

is violated.

As in Section 12.2, it is assumed that the measurements are normally distributed ($\mathfrak{h} = \eta^\star + \mathfrak{e}$, $\mathfrak{e} \sim \mathcal{N}(0, \mathbf{V}_\epsilon)$, $\eta^\star = \mathbf{B}_1 c^\star + b_1$, $\mathbf{B}_2 c^\star = b_2$) and that the rank condition $\operatorname{rank}(\mathbf{B}_2) = n_g$ is fulfilled. Further, for numerical purposes, let the matrix $\mathbf{B}_2$ be well-conditioned.

Since $\mathbf{B}_2$ has full rank (and is well-conditioned), it is still possible to transform the problem into the reduced form (Definition 11.8) such that the above problem becomes

$$\min_c \left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} c + \mathbf{V}_\epsilon^{-\frac{1}{2}} (b_1 - \eta) \right\|_2^2, \tag{12.55a}$$

$$\text{s.t.} \quad \begin{pmatrix} \mathbf{L} & \mathbf{0} \end{pmatrix} c + b_2 = 0. \tag{12.55b}$$

Violation of the rank condition (12.54) is equivalent to the case that one or several entries in the diagonal matrix $\mathbf{S}$ are zero. Hence, the solution is not unique. However, there exists a unique solution of minimum norm $\|c\|_2$, and this solution can be expressed in terms of the rank-deficient generalized inverse $\mathbf{J}_{[r]}^+$ that was introduced in Section 11.2:

$$\hat{c}(\eta) = -\mathbf{J}_{[r]}^+ \begin{pmatrix} \mathbf{V}_\epsilon^{-\frac{1}{2}} (b_1 - \eta) \\ b_2 \end{pmatrix}, \quad \mathbf{J}_{[r]}^+ = \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \mathbf{S}^\dagger & -\mathbf{S}^\dagger \mathbf{A} \mathbf{L}^{-1} \end{pmatrix}. \tag{12.56}$$

Thereby, r denotes the rank of the matrix $\mathbf{J}$.

A result for the practical interpretation of this "minimum-norm" solution is as follows (see also Bock [39], page 154).

Theorem 12.8 (Characterization of Rank-Deficient Solutions of Linear Problems)

The solution $\hat{c}(\eta)$ of problem (12.55) given by equation (12.56) is equivalent to the unique solution of the regularized problem

$$\min_c \left\| \begin{pmatrix} \mathbf{A} & \mathbf{S} \end{pmatrix} c + \mathbf{V}_\epsilon^{-\frac{1}{2}} (b_1 - \eta) \right\|_2^2, \tag{12.57a}$$

$$\textit{s.t.} \quad \begin{pmatrix} \mathbf{L} & \mathbf{0} \end{pmatrix} c + b_2 = 0, \tag{12.57b}$$

$$(\mathbf{1} - \mathbf{P}_r) c = 0. \tag{12.57c}$$

where $\mathbf{P}_r$ *is given by* $\mathbf{P}_r = \begin{pmatrix} \mathbf{1}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}$, *and* $\mathbf{1}_r$ *is the identity matrix of dimension* $r \times r$.

Proof
Let $\hat{c}(\eta) = (z_1^T, z_2^T, z_3^T)$, with $z_1 \in \mathbb{R}^{n_g}$, $z_2 \in \mathbb{R}^{r-n_g}$ and $z_3 \in \mathbb{R}^{n_c-r}$. It is obvious that z_1 and z_2 are, for both problems, given by $z_1 = -\mathbf{L}^{-1}b_2$ and $z_2 = -\bar{\mathbf{S}}^{-1}(\mathbf{V}_\epsilon^{-\frac{1}{2}}(b_1 - \eta) - \mathbf{A}\mathbf{L}^{-1}b_2)$, respectively, where $\bar{\mathbf{S}}$ denotes the upper left block of $\mathbf{S}$ of dimension $(r-n_g)\times(r-n_g)$ that contains the non-zero entries of $\mathbf{S}$.

The vector z_3 is arbitrary in problem (12.55) because it is multiplied by the zero entries of $\mathbf{S}$. In equation (12.56), these components of $\hat{c}(\eta)$ are set to zero because of the zero elements in $\mathbf{S}^\dagger$. For the solution of problem (12.57), these components are zero as well due to the additional equality constraints. ■

By interpreting a *rank-deficient solution* as the unique solution of the regularized problem (12.57), it is immediately possible to apply the analysis of Section 12.2. That is, if the undetermined parameters of problem (12.55) are considered as fixed, the statistical distribution of the parameter estimates can be expressed as (cf. Theorem 12.1)

$$\hat{c}(\mathfrak{h}) \sim \mathcal{N}(c^\star, \mathbf{V}_c), \quad \mathbf{V}_c = \mathbf{J}_{[r]}^+ \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{J}_{[r]}^+)^T, \tag{12.58}$$

i.e. $\mathbf{V}_c$ represents the covariance of the those parameters that are estimated (i.e. locally identifyable).

Further, confidence intervals can be given for the subset of estimated parameters (cf. Lemma 12.7)

$$\Upsilon_{\hat{c}(\eta)}(q(\chi^2_{r-n_g}, \alpha)) \subset \Omega := \{\hat{c}(\eta) + \Delta c \mid |\Delta c_i| < \theta_i \quad \text{for} \quad 1 \leqslant i \leqslant n_c\} \tag{12.59}$$

with $\theta_i := q(\chi^2_{r-n_g}, \alpha) \cdot (\mathbf{V}_c)_{i,i}$.

For ill-conditioned, yet non-singular, problems for which some elements s_i are "almost" zero, two approaches for the analysis exist. The first is to regard the solution as singular, and thus to compute the statistical distribution of the estimated parameters by means of equation (12.58). Alternatively, it is also possible to apply the analysis of Section 12.2 directly. This will yield huge entries in the covariance matrix and thus huge confidence intervals for the estimates of those parameters that correspond to the near-singular values s_i.

Nonlinear Constrained Least-Squares Problems

For nonlinear problems of the form

$$\min_c \ \|F_1(c)\|_2^2 \tag{12.60a}$$

$$\text{s.t. } F_2(c) = 0, \tag{12.60b}$$

the Generalized Gauss-Newton method computes the iterates by $c^{k+1} = c^k + \Delta c^k$, where Δc^k is the solution of the linear constrained least-squares problem

$$\min_{\Delta c} \ \|F_1(c^k) + \mathbf{J}_1(c^k)\Delta c\|_2^2 \tag{12.61a}$$

$$\text{s.t. } F_2(c^k) + \mathbf{J}_2(c^k)\Delta c = 0. \tag{12.61b}$$

The modification of the Generalized Gauss-Newton method for ill-conditioned problems as described in Section 11.2 thereby uses, if necessary, a rank-deficient generalized inverse $\mathbf{J}_{[r]}^+(c^k)$ of $\mathbf{J}(c^k) = \begin{pmatrix} \mathbf{J}_1(c^k) \\ \mathbf{J}_2(c^k) \end{pmatrix}$ to determine the increment Δc^k. Accordingly, the method may end up in a point $\hat{c}(\eta)$ where $\Delta c = 0$ is a rank-deficient solution of the linear constrained least-squares problem. Then – as a generalization of Theorem 12.8 – the following theorem holds.

Theorem 12.9 (Rank-Deficient Solutions in the Context of Nonlinear Problems)

Let $\hat{c}(\eta)$ *be a point such that the locally linearized constrained least-squares problem has a rank-*

deficient solution $\Delta c = 0$. *Then it holds that* $\hat{c}(\eta)$ *is a KKT point of the problem*

$$\min_{c} \ \|F_1(c)\|_2^2 \tag{12.62a}$$

$$s.t. \ F_2(c) = 0 \tag{12.62b}$$

$$(\mathbf{1} - \mathbf{P}_r)(c - \hat{c}(\eta)) = 0 \tag{12.62c}$$

with $\mathbf{P}_r = \mathbf{J}_{[r]}^{+}(\hat{c})\mathbf{J}(\hat{c})$.

Proof
See Bock [39], page 153f. ■

13. Parameter Estimation in the Context of IHDDEs

> *Parameter estimation for DDEs can often be achieved successfully by minimizing the objective function $\Phi(p)$, but in doing so computationally one needs to be aware that $\Phi(p)$ and its derivatives may suffer jumps.*
>
> Baker and Paul, in the conclusion of their paper "Pitfalls in Parameter Estimation for Delay Differential Equations" [13].

In Chapter 10 it was shown that constrained maximum likelihood parameter estimation is, for measurements η that are normally distributed with known covariance matrix $\mathbf{V}_\epsilon$, equivalent to solving nonlinear constrained least-squares problems of the form (10.15), i.e.

$$\min_c \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} \left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t}, c) \right) \right\|_2^2 \tag{13.1a}$$

$$\text{subject to} \quad g(\{y(t_j;c)\}_{j=1}^{n_t}, c) = 0. \tag{13.1b}$$

Thereby, h is the $\mathbb{R}^{n_h}$-dimensional vector of measurement functions and g is the n_g-dimensional vector of equality constraint functions. The parameters are denoted by $c \in \mathbb{R}^{n_c}$.

The state $y(t;c)$ is defined as solution of a dynamic model, see Subsection 10.1. So far, no particular form of the dynamic model has been specified. In this chapter, the case is considered that the dynamic model is a system of differential equations. This task requires to solve infinite-dimensional optimization problems. In particular, for the comparably simple case of ordinary differential equations (ODEs), the following optimization problem arises:

$$\min_{y(t;c),c} \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} \left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t}, c) \right) \right\|_2^2 \tag{13.2a}$$

$$\text{subject to} \quad g(\{y(t_j;c)\}_{j=1}^{n_t}, c) = 0 \tag{13.2b}$$

$$\dot{y}(t;c) = f(t, y(t;c), c) \quad \forall \quad t \in \mathcal{T}(c) := [t^{ini}(c), t^{fin}(c)]. \tag{13.2c}$$

It is recalled that f is called the right-hand-side function of the ODE, and that $t^{ini}(c)$ and $t^{fin}(c)$ are the initial time and the final time. The interval $\mathcal{T}(c)$ has to be chosen in such a way that $t_j \in \mathcal{T}(c)$ for all $1 \leqslant j \leqslant n_t$.

The problem (13.2) is "of infinite dimension" because the state vector $y(t;c)$ occurs as an (infinite-dimensional) optimization variable, which has to fulfill the (infinite-dimensional) condition (13.2c), i.e. the differential equation system. In order to apply, in the numerical practice, an optimization method such as the Generalized Gauss-Newton method (Chapter 11), a finite-dimensional parameterization of the problem (13.2) is needed first.

In the context of this thesis, the goal is of course to go beyond problems of the form (13.2). More precisely, in this chapter, parameter estimation problems are addressed in which the state $y(t;c)$ fulfills a system of impulsive hybrid discrete-continuous delay differential equations (IHDDEs) rather than a system of ODEs. Further, a finite-dimensional parameterization for such problems is presented.

Literature Survey

In order to give the context for the methods presented in this chapter, a literature survey on parameter estimation methods is given first. The emphasis is thereby on the following two aspects:

(a) strategies for finite-dimensional parameterization of infinite-dimensional parameter estimation problems, and

(b) numerical optimization methods for solving the resulting finite-dimensional optimization problems.

Issue (b) is addressed first. The underlying idea of many methods that have been developed for this purpose is to locally apply Newton's method to the KKT conditions (see Definition 10.16) of the optimization problem. However, away from the solution the exact Hessian matrix (cf. equation (11.9)) may not be positive indefinite, and/or the exact computation of the Hessian matrix may be too expensive. This has lead to the development of many approximation strategies for the Hessian.

The (Generalized) Gauss-Newton method (as defined in Chapter 11), e.g., is specific to problems with least-squares objective function. It uses an approximation of the Hessian that disregards second order derivatives of the functions h and g, which significantly reduces the computational costs compared to an "exact" Newton method. In addition, it has been discussed in Subsection 11.1.2 that the use of this approximation ensures that the (Generalized) Gauss-Newton method converges only to "statistically stable" solutions. This makes the (Generalized) Gauss-Newton method particularly well-suited for the solution of (constrained) least-squares parameter estimation problems.

A popular alternative method for the solution of unconstrained least-squares problems is the *Levenberg-Marquardt method* (see Levenberg [174], Marquardt [185]). This method can be viewed as a hybrid between the Gauss-Newton method and a rudimentary steepest descent method (which chooses the increments into the direction of the negative gradient of the objective function). Formally, the Levenberg-Marquardt method is obtained by adding a regularizing term to the Hessian, which also acts as a limiting factor on the norm of the increment. A modification of the Levenberg-Marquardt method for constrained least-squares problems is described, e.g., in Holt and Fletcher [150].

In the context of more general optimization problems – i.e. those of "non-least-squares type" – other strategies have been proposed for the approximation of the Hessian. Of frequent use are those strategies that compute, from a given (approximate) Hessian in one iteration, a new Hessian approximation by so-called "update strategies". Popular examples of such update strategies are called *Davidon-Fletcher-Powell* (see Davidon [74] and Fletcher and Powell [109]), and *Broyden-Fletcher-Goldfarb-Shanno* (see Broyden [50], Fletcher [107], Goldfarb [117], and Shanno [234]).

Optimization methods further differ with respect to the handling of equality constraints (and possibly inequality constraints). One approach is to design optimization algorithms that are capable of handling constraints – this has been the case for the Generalized Gauss-Newton method (Chapter 11). Another prominent method, suitable for problems with general ("non-least-squares") objective functions, and with both equality and inequality constraints, is the *sequential quadratic programming method* (also called "Wilson method", in reference to Wilson [257]).

An alternative approach for the handling of constraints is to reformulate the original constrained optimization problem as an unconstrained problem. This can be achieved by adding terms to the objective function that penalize violations in the constraints. This leads to so-called *penalty methods* and *augmented Lagrangian methods*, all of which require an underlying method for solving unconstrained optimization problems.

Detailed presentation of all aforementioned methods can be found, e.g., in the following textbooks: Fletcher [108], Geiger and Kanzow [114], Bonnans et al. [47], Nocedal and Wright [195], Biegler [30], and Ulbrich and Ulbrich [251].

As a second step in this literature review, methods for finite-dimensional parameterization are discussed next (i.e., item (a) above). An obvious and straightforward way for the finite-dimensional parameterization of problem (13.2) is the so-called *initial value problem approach* (also called *single shooting*). This approach relies on using the initial value for the finite-dimensional parameterization, which leads to the following problem:

$$\min_{c} \| \mathbf{V}_\epsilon^{-\frac{1}{2}} \left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t}, c) \right) \|_2^2 \tag{13.3a}$$

$$\text{subject to} \quad g(\{y(t_j;c)\}_{j=1}^{n_t}, c) = 0 \tag{13.3b}$$

$$y(t;c) \quad \text{is the solution of the following IVP on } \mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$$

$$\dot{y}(t;c) = f(t, y(t;c), c) \tag{13.3c}$$

$$y(t^{ini}(c);c) = y^{ini}(c). \tag{13.3d}$$

For the formulation of the finite-dimensional problem in this form it might be necessary to augment the original parameter vector in order to be able to parameterize y^{ini} in terms of c.

Alternative finite-dimensional parameterizations of problem (13.2) can be found by the so-called *boundary value problem approach*. Two realizations of the boundary value problem approach are popular in the context of optimization problems that are constrained by ODEs or by differential-algebraic equations.

One realization of the boundary value problem approach is the *multiple shooting method*, which goes back to Bock [36, 38, 39]. In this approach, several so-called *multiple shooting nodes* are introduced in order to split up the interval $\mathcal{T}(c)$ into several subintervals. The state vectors at the multiple shooting nodes are introduced as additional optimization variables, and they are used as initial values for the IVPs on the subintervals. Further, additional equality constraints (often called *matching conditions*) are added to the optimization problem in order to ensure that the trajectory is continuous in the solution.

Another boundary value problem approach is the *collocation method*, see e.g. Bock [38] and Tjoa and Biegler [247]. This method relies on choosing a mesh such that the interval $\mathcal{T}(c)$ is split up into a number of subintervals that is typically much larger than in the case of multiple shooting. On each subinterval, a polynomial representation of the state is assumed, and the coefficients of the polynomials are determined by the condition to fulfill the differential equation at the *collocation nodes*. As a result, the parameterization by the collocation method leads to a high dimensional nonlinear constrained least-squares problem that does not require explicit solutions of IVPs in each iteration of the optimization method.

The parameterization of optimization problems – not necessarily with least-squares objective function as in problem (13.2) – by boundary value problem approaches has many advantages. In particular, multiple shooting and collocation approaches are well-suited for highly unstable and chaotic systems for which it is numerically impossible to solve the IVP on the whole time interval $\mathcal{T}(c)$ (see e.g. Bock [39], pages 24 and 226f, and Bock, Kostina, and Schlöder [42]). Furthermore, boundary value problem approaches allow to exploit knowledge about the solution, e.g. the available measurement data, in the initial guess of the optimization variables. This often drastically improves the convergence behavior, see e.g. Lenz et al. [172].

After having reduced the infinite-dimensional parameter estimation problem to finite dimension, a method for solving finite-dimensional nonlinear constrained least-squares problems can be applied. Many combinations of finite-dimensional parameterizations and optimization methods are suggested in the literature, and only those works are cited in the following that have dealt with problems with time delays, in particular with delay differential equations (DDEs).

In one of the first approaches by Burns and Hirsch [52], the DDE is discretized on the whole considered time interval by an explicit Euler scheme or a Runge-Kutta scheme, which yields a delay difference equation. The solution of the resulting finite-dimensional least-squares problem is either done by a rudimentary steepest-descent method or by a quasi-Newton method that employs the Davidon-Fletcher-Powell update strategy.

Other early works have proposed a rather indirect approach: Burns and Cliff [16], Banks and Daniel Lamm [17], and Murphy [191] transform the DDE into an abstract operator equation. They then find finite-dimensional approximations of the abstract operator equation, which are shown to be equivalent to a system of ODEs. Murphy [191] then follows the single shooting approach and solves the resulting finite-dimensional optimization problem with a Levenberg-Marquardt method.

Bocharov and Romanyukha [33] take a more straightforward approach and use the single shooting parameterization in order to obtain a finite-dimensional least-squares problem. This problem is subsequently minimized by combining the derivative-free Nelder-Mead algorithm (see Nelder and Mead [192], and a survey on derivative-free optimization methods in Rios and Sahinidis [215]) and quasi-Newton methods that use, e.g., the Davidon-Fletcher-Powell update strategy.

Baker and Paul [13] raise attention to the issue that the objective function for parameter estimation problems in the context of DDEs is, in general, non-differentiable with respect to the parameters. Nevertheless, Baker, Bocharov, and Paul [10], Baker et al. [11], and Baker et al. [9] combine the single shooting approach with derivative-based optimization methods like sequential quadratic programming or Levenberg-Marquardt and apply them to parameter estimation problems constrained by DDEs and "DDEs of neutral type". Despite the fact that the objective functions may be non-smooth, they used this approach successfully for solving parameter estimation problems in several real-world applications.

The single shooting parameterization has further been used in Hartung and Turi [138], Hartung [134], and Hartung [136]. All three papers put emphasis on the fact that the theory is developed for infinite-dimensional parameters. However, all numerical computations use, of course, finite-dimensional representations of such infinite-dimensional parameters. The finite-dimensional problems are, in Hartung and Turi [138], solved with a quasi-Newton method using a least-squares-specific update strategy for the Hessian that is due to Dennis, Gay, and Welsch [75]. In Hartung [134] and in Hartung [136], a Gauss-Newton method is employed, which, however, is therein called "quasi-linearization method" (see Xuyen and Svrcek [263], who discuss the equivalence of these approaches).

Several application-oriented works have also used the single shooting approach for DDE-constrained parameter estimation problems, see Lehn, Tibken, and Hofer [170], Reinecke [212], and Olufsen and Ottesen [197]. In Lehn, Tibken, and Hofer [170], the optimization is performed by a combination of the Nelder-Mead method and a sequential quadratic programming method. Reinecke [212] uses a Gauss-Newton method, and Olufsen and Ottesen [197] rely on a Levenberg-Marquardt method.

The above-listed references suggest that the single shooting parameterization combined with a straightforward application of a derivative-based optimization method are – despite a possible non-smooth behavior of the objective function (and of the constraint functions) – the most often used techniques for DDE-constrained parameter estimation problems. The remainder of this literature survey mentions comparably recent publications, in which different approaches have been explored.

There is, for example, the work by ZivariPiran [271]. Therein, the "usual" single shooting parameterization is used, but the realization of the optimization is non-standard in two respects. At first, it is noted that the derivative-based optimization method presented therein avoids the use of finite differences for computing the sensitivities $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$. Instead, a "first differentiate, then discretize" approach is used that also accounts for possible jumps in $\mathbf{W}(t;c)$ (see Chapter 7). Second, this work addresses the issue that the non-smoothness of y and $\mathbf{W}$ may cause a non-smoothness of the objective function (or of the constraint functions). In this context, it is suggested to embed a sequential quadratic programming method into a higher level algorithm that ensures smoothness of the optimization problems by means of additional constraints. ZivariPiran [271] reports encouraging numerical results for this strategy for two problems with either one or two constant delays. It is remarked, however, that the proposed strategy is implicitly based on the assumption that the total number and the order of the discontinuities in $y(t;c)$ remains unchanged during the iterations of the sequential quadratic programming method. Hence, it is non-trivial to generalize this approach to the case that the delays are state-dependent.

The works by Horbelt [151] and Horbelt, Timmer, and Voss [152] are distinctive in the respect that they present a multiple shooting approach for DDE-constrained parameter estimation problems. Herein, the IVPs on the subintervals are defined by using splines as initial functions, and these splines are requested to coincide with the IVP solution on the preceding interval on a small number of discrete time points ("ad hoc" modification of matching conditions for problems with time delays). The method seems to have performed satisfactorily for the considered applications. However, the method is somewhat heuristic in the sense that it lacks an analysis of the errors that are introduced to the specific matching conditions. Further, the method has been applied to scalar DDEs with single delays, and the generalization to state-dependent delays is non-trivial.

The author is not aware of any works that have used the collocation approach as a finite-dimensional parameterization of DDE-constrained parameter estimation problems. However, it is referred to Schumann-Bischoff, Luther, and Parlitz [227] and to Mehrkanoon, Mehrkanoon, and Suykens [187]. The approaches used therein are at least distantly related to collocation in that they avoid the use of an IVP solver. It should be noted that they make some rather restrictive assumptions on the problem (linearity of the right-hand-side function in the past states, small time delays, dense measurements) and are limited in their capabilities (e.g. only the initial function or the delay can be estimated).

For completeness, it should be mentioned that the collocation approach has been used for computing steady states or periodic solutions in DDEs by solving boundary value problems; in particular, this is the case in the software package DDE-BIFTOOL by Engelborghs, Luzyanina, and Roose [93], see also Luzyanina, Engelborghs, and Roose [181]. Further, for an example of a collocation discretization used for optimal control problems constrained by DDEs with constant or time-dependent delays, see Günterberg [125].

Novel Results Presented in This Chapter

At first, an infinite-dimensional parameter estimation problem is formulated for the case that the state y is defined by an impulsive hybrid discrete-continuous delay differential equation (IHDDE) model. Of course, this comprises parameter estimation problems in the context of all subclasses of differential equations that were presented in Section 1.2. It is remarked that, to the knowledge of the author, also the simpler class of hybrid discrete-continuous delay differential equations (HDDEs) has not yet been treated in the context of parameter estimation.

As a second contribution of this chapter, a finite-dimensional parameterization of the IHDDE-constrained parameter estimation problem by means of the single shooting approach is presented.

It is further elaborated on the issue that IHDDE-constrained parameter estimation problems must generally be considered as non-smooth optimization problems. A justification is given why the application of derivative-based optimization methods may nevertheless be successful.

Eventually, this chapter presents the numerical methods that are implemented in *PARAMeter Estimation in Differential Equations* (ParamEDE) – a newly developed code for practical parameter estimation in IHDDEs. This code is the first that realizes a damped Generalized Gauss-Newton method based on the restrictive monotonicity test for parameter estimation in DDEs. ParamEDE uses Colsol-DDE as underlying IVP solver and exploits its capabilities for computing the derivatives of IVP solutions with respect to parameters.

Organization of This Chapter

The chapter is divided into three sections. Section 13.1 introduces the infinite-dimensional IHDDE-constrained least-squares problem and presents a finite-dimensional parameterization by means of the single shooting approach. Section 13.2 discusses why IHDDE-constrained least-squares problems have to be regarded as non-smooth, and elaborates on the consequences of this non-smoothness for the properties of local solutions and for the applicability of derivative-based optimization methods. Eventually, Section 13.3 presents a detailed algorithm for practical parameter estimation in IHDDEs that is based on a damped Generalized Gauss-Newton method with restrictive monotonicity test and a regularization strategy. It further discusses the practical realization of this algorithm in the software package ParamEDE.

13.1. Problem Formulation: Parameter Estimation in IHDDEs

13.1.1. Infinite-Dimensional Problem

Consider the case that a real-world process is modeled by an IHDDE as in Definition 1.1. Then the following infinite-dimensional parameter estimation problem is formulated as a generalization of problem (13.2):

$$\min_{y(t;c),c} \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} \left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t}, c) \right) \right\|_2^2 \tag{13.4a}$$

$$g(\{y(t_j;c)\}_{j=1}^{n_t}, c) = 0 \tag{13.4b}$$

$$\dot{y}(t;c) = f(t, y(t;c), c, \{y(t-\tau_i(t, y(t;c), c); c)\}_{i=1}^{n_\tau}, \zeta(t)) \qquad \text{for } t \in \mathcal{D}_1^t(\mathcal{T}(c)) \tag{13.4c}$$

$$\begin{aligned} y(t;c) =& y^+(t;c) \\ =& y^-(t;c) + \omega(t, y^-(t;c), c, \{y^\bullet(t-\tau_i(t, y^-(t;c), c); c)\}_{i=1}^{n_\tau}, \zeta(t)) \text{ for } t \in \mathcal{D}_0^t(\mathcal{T}(c)). \end{aligned} \tag{13.4d}$$

Herein, $\mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)]$ is the considered interval, with $t^{ini}(c)$ denoting the initial time and $t^{fin}(c)$ denoting the final time. For the problem formulation given above, the interval has to be such that $t_j \in \mathcal{T}(c)$ for all $1 \leqslant j \leqslant n_t$.

For completeness, also the meaning of all other symbols in equation (13.4) is recalled from Section 1.1: $\mathcal{D}_1^t(\mathcal{T}(c))$ is the set that contains those times for which all switching function signs $\zeta(t)$ are non-zero, and $\mathcal{D}_0^t(\mathcal{T}(c)) = \mathcal{T}(c) \backslash \mathcal{D}_1^t(\mathcal{T}(c))$ is a set containing those times for which at least one switching function sign $\zeta(t)$ is zero. Further, f is the right-hand-side function, τ_i are the delay functions, $\zeta(t) = (\zeta_1(t), \ldots, \zeta_{n_\sigma}(t))^T$ are the signs of the switching functions σ_i, and ω is the impulse function. Eventually, $t^{ini}(c)$ and $t^{fin}(c)$ are the initial time and the final time of the considered time interval, respectively.

The IHDDE-constrained parameter estimation problem (13.4) is, like problem (13.2), of infinite dimension: The function $y(t;c)$ occurs as infinite-dimensional optimization variable, which is subject to the infinite-dimensional constraint (13.4c). In order to make the problem computationally treatable, a finite-dimensional parameterization is needed.

13.1.2. Finite-Dimensional Problem

A finite-dimensional parameterization of the optimization problem (13.4) is given by

$$\min_c \left\| \mathbf{V}_\epsilon^{-\frac{1}{2}} \left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t}, c) \right) \right\|_2^2 \tag{13.5a}$$

$$g(\{y(t_j;c)\}_{j=1}^{n_t}, c) = 0 \tag{13.5b}$$

$$\begin{aligned} y(t;c) \quad & \text{is solution of the following IHDDE-IVP on } \mathcal{T}(c) = [t^{ini}(c), t^{fin}(c)] : \\ \dot{y}(t;c) =& f(t, y(t;c), c, \{y(t-\tau_i(t, y(t;c), c); c)\}_{i=1}^{n_\tau}, \zeta(t)) \qquad \text{for } t \in \mathcal{D}_1^t(\mathcal{T}(c)) \\ y(t;c) =& y^+(t;c) \\ =& y^-(t;c) + \omega(t, y^-(t;c), c, \{y^\bullet(t-\tau_i(t, y^-(t;c), c); c)\}_{i=1}^{n_\tau}, \zeta(t)) \text{ for } t \in \mathcal{D}_0^t(\mathcal{T}(c)) \end{aligned} \tag{13.5c, 13.5d}$$

$$y(t^{ini}(c);c) = y^{ini}(c) \tag{13.5e}$$

$$y(t;c) = \phi(t,c) \quad \text{for} \quad t < t^{ini}(c). \tag{13.5f}$$

As in Chapter 1, y^{ini} is called the initial value and ϕ is called the initial function.

The difference to problem (13.4) is that $y(t;c)$ is not longer an optimization variable, but that it is, for given parameter values c, determined as solution of an IHDDE-IVP. It is noted that finite-dimensional parameterization implies, for IHDDEs, that also the initial function is parameterized in terms of c.

In taking the step from problem (13.4) to problem (13.5), the parameter vector c may have to be augmented in order to parameterize y^{ini} and ϕ in terms of c.

Problem (13.5) can be regarded as a *single shooting approach* (or *initial value problem approach*) applied to problem (13.4) in the context of IHDDEs.

13.2. Non-Smooth Parameter Estimation Problems

As in Chapter 10, let the following abbreviating notations be defined for the objective function and for the constraint functions in problem (13.5):

$$F_1(c) := \mathbf{V}_\epsilon^{-\frac{1}{2}} \left(\eta - h(\{y(t_j;c)\}_{j=1}^{n_t}, c) \right) \tag{13.6a}$$

$$F_2(c) := g(\{y(t_j;c)\}_{j=1}^{n_t}, c). \tag{13.6b}$$

Further, let $n_{F_1} := n_h$ and $n_{F_2} := n_g$ be the dimensions of $F_1(c)$ and $F_2(c)$.

Since IHDDE-IVP solutions may have discontinuities of order 0, the functions $F_1(c)$ and $F_2(c)$ are generally discontinuous functions of c even if the functions h and g are smooth. More precisely, discontinuities occur for those points $c \in \mathbb{R}^{n_c}$ in parameter space for which a discontinuity of order 0 is located at one of the measurement times t_j.

Consider next the first derivative of the functions $F_1(c)$ and $F_2(c)$. In the abbreviating notation defined above, the Jacobians

$$\mathbf{J}_i(c) = \left. \frac{\partial F_i(c')}{\partial c'} \right|_{c'=c} \tag{13.7}$$

can formally be expressed as follows:

$$\mathbf{J}_1(c) = -\mathbf{V}_\epsilon^{-\frac{1}{2}} \left(\sum_{j=1}^{n_t} \frac{\partial h}{\partial y_j} \frac{\partial y(t_j;c)}{\partial c} + \frac{\partial h}{\partial c} \right) \tag{13.8a}$$

$$\mathbf{J}_2(c) = \sum_{j=1}^{n_t} \frac{\partial g}{\partial y_j} \frac{\partial y(t_j;c)}{\partial c} + \frac{\partial h}{\partial c}. \tag{13.8b}$$

The partial derivatives of h and g are thereby evaluated at $(\{y(t_j;c)\}_{j=1}^{n_t}, c)$.

Equations (13.8) show that the Jacobians $\mathbf{J}_1(c)$ and $\mathbf{J}_2(c)$ involve the derivative of the IHDDE-IVP solution with respect to the parameters c. In Chapter 7, sufficient conditions were given under which there exists a piecewise continuously differentiable function $\mathbf{W}(t;c)$ such that $\mathbf{W}(t;c) = \partial y(t;c)/\partial c$. However, even in the case that these sufficient differentiability assumptions are fulfilled, the Jacobians $\mathbf{J}_1(c)$ and $\mathbf{J}_2(c)$ are non-differentiable at a point $c \in \mathbb{R}^{n_c}$ in parameter space if one of the jumps in $\mathbf{W}(t;c)$ occurs at one of the measurement times t_j.

Parameter estimation problems constrained by IHDDEs therefore need to be considered as non-smooth optimization problems. The various consequences of this non-smoothness are discussed in the following.

13.2.1. Necessary and Sufficient Optimality Conditions

Necessary and sufficient optimality conditions were given in Section 10.4 under the assumption that $F_1(c)$ and $F_2(c)$ are (twice) continuously differentiable with respect to the parameters. If the state $y(t;c)$ is given by the solution of an IHDDE-IVP, this assumption might not be fulfilled.

Consider the case that h and g are smooth functions of their arguments, and let $\hat{c}$ be a local solution (Definition 10.8) such that the sufficient conditions for differentiability of the IVP solution (see Chapter 7) are fulfilled. Moreover, assume that the discontinuities in y and $\mathbf{W}$ do not occur at the measurement times t_j, $1 \leqslant j \leqslant n_t$. Then it holds that $F_1(c)$ and $F_2(c)$ are locally differentiable in a neighborhood of the local solution $\hat{c}$. Accordingly, under these assumptions, the necessary condition of first order given in Theorem 10.15 remains valid.

Second order differentiability has not been discussed in this thesis so far. Assume, however, that also the second derivative $\partial^2 y(t;c)/\partial c^2$ is a piecewise smooth function. If its time points of discontinuities are, for a local solution $\hat{c}$, safely away from the measurements times t_j, $1 \leqslant j \leqslant n_t$, then $F_1(c)$ and $F_2(c)$ are locally twice continuously differentiable and the Theorems 10.15 and 10.17 hold.

Of course, the question remains whether a set of sufficient conditions can be found, which ensure that $\partial^2 y(t;c)/\partial c^2$ is a piecewise smooth function. In general, this question can be approached by the techniques that were employed in Chapter 7 for showing first order differentiability, i.e. applying the method of steps and exploiting regularity conditions on the switching functions. However, the concrete formulation of suitable regularity conditions and the derivation of the "second-order variational IVP" and the corresponding jump expressions for the second order derivatives become very technical and are thus not given in this work.

13.2.2. Derivative-Based Optimization

As mentioned before, the functions $F_1(c)$ and $F_2(c)$ (as defined in equations (13.6a) and (13.6b)) are generally discontinuous or non-differentiable if $y(t;c)$ represents an IHDDE-IVP solution. Clearly, this fact calls for a justification of using derivative-based optimization methods.

The justification is as follows. If h and g are smooth, and if the discontinuities of order 0 in y and $\mathbf{W}$ do, at a local minimum $\hat{c}$, not occur at the measurement times t_j, $1 \leqslant j \leqslant n_t$ then the functions $F_1(c)$ and $F_2(c)$ are differentiable in a neighborhood of $\hat{c}$. Accordingly, the differentiability assumption of the local contraction theorem (Theorem 11.6) is fulfilled within this neighborhood. This directly yields a local convergence result for the Generalized Gauss-Newton method under the condition that the initial guess is sufficiently close to the solution.

If no sufficiently good initial guess is available, the Generalized Gauss-Newton method may fail to converge. More over, also a damped Generalized Gauss-Newton method – like the one presented in Section 11.3 – may fail to converge. For example, it can get stuck at a boundary of the domain on which F_1 and F_2 are differentiable. Numerical investigations have shown, however, that damped Generalized Gauss-Newton methods are quite successful for practical non-smooth least-squares problems (see Lenz [171]), even though standard convergence theory does not apply. Further numerical investigations on the practical performance of damped Generalized Gauss-Newton methods applied to a non-smooth least-squares parameter estimation problem are given in Chapter 16.

13.2.3. Analysis of Solutions

Rigorous statistical results for the distribution of parameter estimates as a function of the (random) measurement data, for confidence regions, and for confidence intervals were obtained in Chapter 12

under the assumption that $F_1(c)$ and $F_2(c)$ are linear.

In the vast majority of practical least-squares problems the functions F_1 and F_2 are nonlinear. This is, in particular, true for parameter estimation problems where the state $y(t;c)$ is the solution of an IVP in differential equations. In this case, only approximations of confidence regions are available, see equations (12.46), (12.47), and (12.49). Accordingly, also the intervals $[\hat{c}_i - \theta_i, \hat{c}_i + \theta_i]$ with θ_i as defined in Lemma 12.7 are only approximations of confidence intervals. The quality of these approximations thereby depends on the "amount of nonlinearity" of the functions $F_1(c)$ and $F_2(c)$.

In the case that $y(t;c)$ is an IHDDE-IVP solution, the functions $F_1(c)$ and $F_2(c)$ may be discontinuous or non-differentiable for some parameter values. Nevertheless, if F_1 and F_2 are locally differentiable at a solution $\hat{c}$, the linear indifference region $\Upsilon^L_{\hat{c}}(q(\chi^2_{n_c-n_g}, \alpha))$ is still defined (equation (12.47)) and the intervals $[\hat{c}_i - \theta_i, \hat{c}_i + \theta_i]$ can formally be computed (Lemma 12.7). However, both the indifference region and the intervals have to be interpreted with caution regarding a statistical inference about the correct parameters $c^\star$.

13.3. Practical Parameter Estimation in IHDDEs

13.3.1. A Practical Algorithm

In this section, an algorithm is suggested for estimating parameters in model functions of IHDDEs. It should be noted that the strategy for stepsize selection and for the computation of rank-deficient generalized inverses in this algorithm resembles the implementation in the parameter estimation software PARFIT as presented in Bock [36, 38, 39], Bock, Kostina, Schlöder [41], see also Lenz [171].

For convenience, it is briefly recalled that the Generalized Gauss-Newton solves nonlinear constrained least-squares problems of the form

$$\min_c \|F_1(c)\|_2^2 \tag{13.9a}$$

$$\text{s.t. } F_2(c) = 0 \tag{13.9b}$$

by starting from an initial guess c^0 and iterating from c^k to c^{k+1} by setting $c^{k+1} = c^k + \Delta c^k$ (see Chapter 11). Thereby, Δc^k is the solution of the linear constrained least-squares problem

$$\min_{\Delta c} \|F_1(c^k) + \mathbf{J}_1(c^k)\Delta c\|_2^2 \tag{13.10a}$$

$$\text{s.t. } F_2(c^k) + \mathbf{J}_2(c^k)\Delta c = 0. \tag{13.10b}$$

By setting $F(c^k) := \begin{pmatrix} F_1(c^k) \\ F_2(c^k) \end{pmatrix}$ and $\mathbf{J}(c^k) := \begin{pmatrix} \mathbf{J}_1(c^k) \\ \mathbf{J}_2(c^k) \end{pmatrix}$, the increment Δc^k can be expressed as

$$\Delta c^k = -\mathbf{J}^+(c^k)F(c^k), \tag{13.11}$$

where $\mathbf{J}^+(c^k)$ is the generalized inverse of $\mathbf{J}(c^k)$.

The algorithm given below makes use of the restrictive monotonicity test as globalization strategy, i.e. the iterates are given by $c^{k+1} = c^k + \alpha^k \Delta c^k$, $\alpha^k \in (0, 1]$, see Section 11.3 for details. The algorithm further uses the modification for ill-conditioned and singular problems described in Section 11.2. This means that the increment is determined by

$$\Delta c^k = -\mathbf{J}^+_{[r]}(c^k)F(c^k), \tag{13.12}$$

with $\mathbf{J}^+_{[r]}(c^k)$ being a rank-deficient generalized inverse with rank $r \leqslant n_c$. It is further recalled that $\tilde{r} = r - n_{F_2}$, where n_{F_2} is the number of equality constraints.

As in Section 11.2, it is assumed that the rank condition $\operatorname{rank}(\mathbf{J}_2(c^k)) = n_{F_2}$ is fulfilled for all iterates.

Algorithm 13.1 (A Generalized Gauss-Newton Method for Parameter Estimation in IHDDEs)

Start with $j = 0$, $k = 0$, and with an initial guess c^0 for the unknown parameters. Let further $\alpha_{min} \in (0, 1)$ be a minimum stepsize for the Generalized Gauss-Newton method, and let $\alpha_{acc} \in$

$(\alpha_{min}, 1]$. *Moreover, let* γ_{max} *be an upper bound of the acceptable condition (in a given norm) of* $\mathbf{J}_1(c)$ *on the kernel of* $\mathbf{J}_2(c)$. *Finally, let* ϵ_{term} *be a termination criterion and let* η, η_1, *and* η_2 *be given such that* $0 < \eta_1 < \eta < \eta_2 < 2$. *Initialize* $\alpha_{bnd} = 1$.

1. *Solve the IHDDE-IVP in order to obtain* $y(t; c^k)$ *and compute the sensitivities* $\mathbf{W}(t; c^k) = \partial y(t; c^k)/\partial c$.
   - *If an error occurs in the (numerical) solution of the IHDDE-IVP or during the computation of the sensitivities, proceed with step 2.*
   - *If the (numerical) integration was successful, proceed with step 3.*
2. *Proceed according to the iteration number* k *as follows:*
   - *If* $k = 0$ *(i.e. first iteration), stop and exit with an error message.*
   - *If* $k > 0$, *Propose a new stepsize* $\alpha^{k-1,j+1}$ *or a new increment* Δc^{k-1} *according to the following rule:*
     - *If* $\alpha^{k-1,j} > \alpha_{min}$, *set* $\alpha_{bnd} = \alpha^{k-1,j}$ *and*
     $$\alpha^{k-1,j+1} = \max\left(\frac{\alpha^{k-1,j}}{2}, \alpha_{min}\right).$$
     - *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} = 0$, *and* $\tilde{r} > 1$, *then set* $\alpha^{k-1,j+1} = \alpha_{min}$, $\tilde{r} = \tilde{r} - 1$, *and compute a new increment with a rank-deficient generalized inverse, i.e.*
     $$\Delta c^{k-1} = -\mathbf{J}^+_{[r]}(c^{k-1})F(c^{k-1}),$$
     where $r = \tilde{r}$. *Reset* $\alpha_{bnd} = 1$.
     - *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} = 0$, *and* $\tilde{r} = 1$, *then stop and exit with an error message.*
     - *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} > 0$ *and* $\tilde{r} > 0$, *then set* $\alpha^{k-1,j+1} = \alpha_{min}$, $\tilde{r} = \tilde{r} - 1$, *and compute a new increment with a rank-deficient generalized inverse, i.e.*
     $$\Delta c^{k-1} = -\mathbf{J}^+_{[r]}(c^{k-1})F(c^{k-1}),$$
     where $r = n_{F_2} + \tilde{r}$. *Reset* $\alpha_{bnd} = 1$.
     - *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} > 0$, *and* $\tilde{r} = 0$, *then stop and exit with an error message.*

   Set $j = j + 1$, $c^k = c^{k-1} + \alpha^{k-1,j}\Delta c^{k-1}$, *and go back to step 1.*
3. *Evaluate the functions* $F_1(c^k)$ *and* $F_2(c^k)$ *(as defined by equations (13.6a), (13.6b)). Then proceed according to the iteration number* k *as follows:*
   - *If* $k = 0$ *(first iteration), then proceed with step 7.*
   - *If* $k > 0$, *proceed with step 4.*
4. *Compute an auxiliary increment* $\overline{\Delta c^k} = -\mathbf{J}_{[r]}(c^{k-1})F(c^k)$, *set* $\delta c = \overline{\Delta c^k} - (1 - \alpha^{k-1,j})\Delta c^{k-1}$, *and compute* $\hat{w} = 2 \cdot \|\delta c\|_2/(\alpha^{k-1,j} \cdot \|\Delta c^k\|_2)^2$. *Proceed as follows (restrictive monotonicity test):*
   - *If* $\hat{w}\alpha^{k-1,j}\|\Delta c^k\|_2 > \eta_2$, *proceed with step 5.*
   - *If* $\hat{w}\alpha^{k-1,j}\|\Delta c^k\|_2 < \eta_1$, *proceed with step 6.*
   - *If* $\eta_1 \leqslant \hat{w}\alpha^{k-1,j}\|\Delta c^k\|_2 \leqslant \eta_2$, *proceed with step 7.*
5. *Propose a new stepsize* $\alpha^{k-1,j+1}$ *or a new increment* Δc^{k-1} *according to the following rule:*
   - *If* $\alpha^{k-1,j} > \alpha_{min}$, *set* $\alpha_{bnd} = \alpha^{k-1,j}$ *and*
   $$\alpha^{k-1,j+1} = \max\left(\frac{\eta}{\hat{w}\|\Delta c^k\|_2}, \alpha_{min}\right).$$

- *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} = 0$, *and* $\tilde{r} > 1$, *then set* $\alpha^{k-1,j+1} = \alpha_{min}$, $\tilde{r} = \tilde{r} - 1$, *and compute a new increment with a rank-deficient generalized inverse, i.e.*

$$\Delta c^{k-1} = -\mathbf{J}^{+}_{[r]}(c^{k-1})F(c^{k-1}),$$

  where $r = \tilde{r}$. *Reset* $\alpha_{bnd} = 1$.
- *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} = 0$, *and* $\tilde{r} = 1$, *then stop and exit with an error message.*
- *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} > 0$ *and* $\tilde{r} > 0$, *then set* $\alpha^{k-1,j+1} = \alpha_{min}$, $\tilde{r} = \tilde{r} - 1$, *and compute a new increment with a rank-deficient generalized inverse, i.e.*

$$\Delta c^{k-1} = -\mathbf{J}^{+}_{[r]}(c^{k-1})F(c^{k-1}),$$

  where $r = n_{F_2} + \tilde{r}$. *Reset* $\alpha_{bnd} = 1$.
- *If* $\alpha^{k-1,j} = \alpha_{min}$, $n_{F_2} > 0$, *and* $\tilde{r} = 0$, *then stop and exit with an error message.*

Set $j = j + 1$, $c^k = c^{k-1} + \alpha^{k-1,j}\Delta c^{k-1}$, *and go back to step 1.*

6. *Proceed as follows depending on the value of the employed stepsize* $\alpha^{k-1,j}$*:*
   - *If* $\alpha^{k-1,j} \geqslant \min(\alpha_{acc}, 0.5 \cdot \alpha_{bnd})$, *proceed with step 7.*
   - *If* $\alpha^{k-1,j} < \min(\alpha_{acc}, 0.5 \cdot \alpha_{bnd})$, *then propose a new stepsize*

$$\alpha^{k-1,j+1} = \min\left(\frac{\eta}{\hat{w}\|\Delta c^k\|_2}, 0.5 \cdot (\alpha_{bnd} + \alpha^{k-1,j})\right).$$

     Then, set $j = j + 1$, $c^k = c^{k-1} + \alpha^{k-1,j}\Delta c^{k-1}$, *and go back to step 1.*
7. *Accept* c^k *as new iterate. Compute the Jacobians* $\mathbf{J}_1(c^k)$ *and* $\mathbf{J}_2(c^k)$ *by means of equations (13.8).*
8. *Choose the largest possible rank* r, $r = n_{F_2} + \tilde{r}$, $0 \leqslant \tilde{r} \leqslant n_c - n_{F_2}$, *such that the condition of* $\mathbf{J}_1(c^k)$ *on the kernel of* $\mathbf{J}_2(c^k)$ *is, in the chosen norm, at most* γ_{max}.
9. *Compute the new increment*

$$\Delta c^k = -\mathbf{J}^{+}_{[r]}(c^k)F(c^k).$$

10. *Perform termination check:*
    - *If* $\|\Delta c^k\|_2 \leqslant \epsilon_{term}$, *then exit with message "convergence achieved", define the solution* $\hat{c} := c^k$ *and provide an analysis of the obtained solution.*
    - *Otherwise, proceed with step 11.*
11. *Propose a stepsize for the next step:*
    - *If* $k = 0$ *(first iteration), then set* $\alpha^{k,0} = 1$.
    - *If* $k > 0$, *then set* $\hat{w} = 2 \cdot \|\delta c\|_2/(\alpha^{k-1,j} \cdot \|\Delta c^k\|_2)^2$, *and*

$$\alpha^{k,0} = \max\left(\min\left(\frac{\eta}{\hat{w} \cdot \|\Delta c^k\|_2}, 1\right), \alpha_{min}\right).$$

12. *Set* $j = 0$, $c^{k+1} = c^k + \alpha^{k,j}\Delta c^k$, $k = k + 1$. *Reset* $\alpha_{bnd} = 1$. *Then go back to step 1.*

This algorithm is an augmented version of Algorithm 11.13, which allows the use of rank-deficient generalized inverses for the computation of increments in the case that the restrictive monotonicity test fails for a user-given minimum stepsize α_{min}. Furthermore, Algorithm 13.1 employs a lower bound η_1 for the restrictive monotonicity test, and provides an error handling for the case that the IHDDE-IVP solution or its derivative cannot be computed in parts of the parameter space.

Some strategies used in Algorithm 13.1 are now explained in detail.

- If the integration or the sensitivity computation fails in step 1, then the algorithm attempts, for $k > 0$, different stepsizes or different increments that are computed with rank-deficient generalized inverses. Preference is given to a reduction of the stepsize, and rank reductions are only attempted once the minimum stepsize is reached.

- A similar strategy is employed if the upper bound of the restrictive monotonicity test is violated, see step 5. The only difference to the decision tree in step 2 is the choice of $\alpha^{k-1,j+1}$ in the case that $\alpha^{k-1,j} > 1$. Here, a new stepsize is proposed, which is based on the "optimal" value η. Contrariwise, the stepsize is simply multiplied by a factor 1/2 in step 2.

- For unconstrained problems, rank reductions are performed in the steps 2 and 5 until the rank of $\mathbf{J}_1(c^k)$ has reached the value 1. For constrained problems, the rank of $\mathbf{J}_1(c^k)$ on the kernel of $\mathbf{J}_2(c^k)$ may even be reduced to 0; in this case, the iteration only aims at fulfilling the linearized equality constraints.

- The quantity α_{bnd} records, for a given increment (corresponding to a certain rank r of the generalized inverse), the smallest attempted stepsize that has led to a failure of the integration or to a violation of the restrictive monotonicity test. Before the start of an iteration, and also if a new increment is computed with a rank-deficient generalized inverse, the value is reset to 1.

- The value of α_{bnd} is used in step 6, which is called if the lower bound of the restrictive monotonicity test is violated. Here, it is proposed to accept the iterate if the stepsize is at least half as large as α_{bnd} – i.e. the stepsize for which the upper bound of the restrictive monotonicity test had been violated or for which integration had failed.

- The proposed stepsize is further accepted in step 6 if it is greater than or equal to α_{acc}. The algorithmic parameter α_{acc} can thus be used to "override" the restrictive monotonicity test with respect to the lower bound η_1.

- A good default value for η is 1, which corresponds to the "optimal" stepsize according to the discussion in Section 11.3. Reasonable values for η_1 and η_2 are 0.5 and 1.5, respectively.

13.3.2. Realization in ParamEDE

Algorithm 13.1 is realized in a new software package called ParamEDE (PARAMeter Estimation in Differential Equations). In the following, specific topics concerning the implementation are discussed in detail.

Solution of IHDDE-IVPs and Computation of Sensitivities

In ParamEDE, the IHDDE-IVP solution $y(t; c^k)$ that is needed in step 1 of Algorithm 13.1 are computed by calling Colsol-DDE (see Chapter 6). The sensitivities $\mathbf{W}(t; c^k)$ are also computed by Colsol-DDE by using the principle of Internal Numerical Differentiation (see Chapters 8 and 9).

It is recalled at this point that Colsol-DDE performs a number of numerical checks that aim at detecting a possible non-differentiable behavior of the IVP solution with respect to the parameters c, see Subsection 9.1.11. If one of the numerical checks is not passed, the integration is stopped, and Algorithm 13.1 attempts to find a different stepsize or determines a rank-deficient increment, see step 2.

Computation of Derivatives of the Functions h and g

The vector of measurement functions h and the vector of constraint functions g need to be differentiated with respect to their arguments in order to compute the Jacobian $\mathbf{J}_1$ and $\mathbf{J}_2$. For this purpose, an Automatic Differentiation tool can be used. In particular, ParamEDE is designed to be used in conjunction with Tapenade, see Hascoët and Pascual [140, 141]. ParamEDE can therefore be used in such a way that it is derivative-free for the user.

Multiple Experiments

In practical applications, it is typical that a sequence of similar experiments has been carried out. ParamEDE supports the user in specifying such "multi-experiment parameter estimation problems" by providing the opportunity to specify "global" parameters, which occur in several or in all experiments, and "local" parameters, which occur in only one experiment. For the solution of the IHDDE-IVP in each of the experiments, Colsol-DDE is called with a parameter vector that comprises both the global parameters and the correct set of local parameters.

Computation of Increments

In each iteration, the increments are determined by $\Delta c^k = -\mathbf{J}^+_{[r]}(c^k)F(c^k)$, where $n_{F_2} \leqslant r \leqslant n_c$ if $n_{F_2} > 0$ and $1 \leqslant r \leqslant n_c$ if $n_{F_2} = 0$. If $r < n_c$, then $\mathbf{J}^+_{[r]}(c^k)$ represents a rank-deficient generalized inverse.

For the practical computation of the increment Δc^k, define

$$\mathbf{J}(c^k) := \begin{pmatrix} \mathbf{J}_1(c^k) \\ \mathbf{J}_2(c^k) \end{pmatrix}. \tag{13.13}$$

By a householder decomposition of $\mathbf{J}_2^T(c^k)$, the following representation can be obtained:

$$\mathbf{J}(c^k) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{L} & \mathbf{0} \end{pmatrix} \mathbf{Q}, \tag{13.14}$$

where $\mathbf{Q}$ is an orthogonal matrix. If the rank condition $\operatorname{rank}(\mathbf{J}_2(c^k)) = n_{F_2}$ is fulfilled (which is assumed by ParamEDE), then $\mathbf{L}$ is a regular $n_{F_2} \times n_{F_2}$ lower triangular matrix. Further, $\mathbf{A}$ and $\mathbf{B}$ represent the first n_{F_2} columns and the last $n_c - n_{F_2}$ columns of $\mathbf{J}_1(c^k)\mathbf{Q}^T$, respectively. By computing a singular value decomposition of $\mathbf{B}$ such that $\mathbf{B} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, it follows that

$$\mathbf{J}(c^k) = \begin{pmatrix} \mathbf{U} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{A}} & \mathbf{S} \\ \mathbf{L} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^T \end{pmatrix} \mathbf{Q}. \tag{13.15}$$

In this equation, $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and $\mathbf{S}$ is a diagonal matrix containing the singular values s_i (for $1 \leqslant i \leqslant n_c - n_{F_2}$) of $\mathbf{J}_1(c^k)$ on the kernel of $\mathbf{J}_2(c^k)$. Further, $\tilde{\mathbf{A}} = \mathbf{U}^T\mathbf{A}$, and $\mathbf{1}$ and $\mathbf{0}$ represent identity matrices and zero matrices of appropriate dimension, respectively.

At this point, it is observed that any matrix $\mathbf{J}(c^k)$ can be brought into the reduced form (Definition 11.7) by using orthogonal transformations, provided that $\mathbf{J}_2(c^k)$ is regular. It is further noted that $s_{n_c - n_{F_2}} > 0$ is a sufficient condition for positive definiteness of $\mathbf{J}_1^T(c^k)\mathbf{J}_1(c^k)$ on the kernel of $\mathbf{J}_2(c^k)$, and hence, an increment computed with a "full-rank" generalized inverse (i.e. $\Delta c^k = -\mathbf{J}^+_{[n_c]}(c^k)F(c^k)$) is a strict local minimum of the linear constrained least-squares problem, compare Section 11.1.

Consider now the case that a generalized inverse $\mathbf{J}^+_{[r]}(c^k)$ of $\mathbf{J}(c^k)$ should be computed, where $r \leqslant n_c$ denotes the possibly reduced rank. For this purpose, define $\tilde{r} := r - n_{F_2}$, and let $\tilde{\mathbf{S}}$ be a diagonal matrix whose entries are given by $\tilde{s}_i = s_i$ for $1 \leqslant i \leqslant \tilde{r}$ and by $\tilde{s}_i = 0$ for $\tilde{r}+1 \leqslant i \leqslant n_c - n_{F_2}$ (cf. Subsection 11.2.3). The (possibly rank-deficient) generalized inverse $\mathbf{J}^+_{[r]}(c^k)$ of $\mathbf{J}(c^k)$ is then given by

$$\mathbf{J}^+_{[r]}(c^k) = \mathbf{Q}^T \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{A}} & \tilde{\mathbf{S}} \\ \mathbf{L} & \mathbf{0} \end{pmatrix}^+ \begin{pmatrix} \mathbf{U}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \tag{13.16}$$

(verify $\mathbf{J}^+_{[r]}(c^k)\mathbf{J}(c^k)\mathbf{J}^+_{[r]}(c^k) = \mathbf{J}^+_{[r]}(c^k)$). Further, by recalling equation (11.25), it follows that

$$\mathbf{J}^+_{[r]}(c^k) = \mathbf{Q}^T \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \tilde{\mathbf{S}}^\dagger & -\tilde{\mathbf{S}}^\dagger\mathbf{A}\mathbf{L}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{U}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}, \tag{13.17}$$

where $\tilde{\mathbf{S}}^\dagger$ is the Moore-Penrose pseudoinverse of $\tilde{\mathbf{S}}$.

ParamEDE computes the above-described matrix decompositions and obtains a possibly "rank-

deficient increment" by

$$\Delta c^k = -\mathbf{Q}^T \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \tilde{\mathbf{S}}^\dagger & -\tilde{\mathbf{S}}^\dagger \mathbf{A} \mathbf{L}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{U}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \begin{pmatrix} F_1(c^k) \\ F_2(c^k) \end{pmatrix}. \tag{13.18}$$

The matrices $\mathbf{Q}$, $\mathbf{A}$, $\mathbf{L}$, $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}$ are stored until a step is eventually accepted (step 7 in Algorithm 13.1). Thus, for the computation of the auxiliary increment $\overline{\Delta c^k}$ in step 4, a costly recomputation of the matrix decomposition is avoided. The matrix decompositions are also reused in case that the rank-reduction strategy is activated (step 2 or step 5).

Practical Rank Decision

Algorithm 13.1 requires to bound the condition of the matrix $\mathbf{J}_1(c^k)$ on the kernel of $\mathbf{J}_2(c^k)$. ParamEDE bounds the condition in the spectral norm by setting $\tilde{r}$ such that $s_i \geqslant s_1/\gamma_{max}$ for $1 \leqslant i \leqslant \tilde{r}$, where γ_{max} is a user-given input parameter. Furthermore, ParamEDE allows to specify a lower bound on the singular values, because the singular values correspond to the standard error of the corresponding parameter combination, see Subsection 11.2.4 and Section 12.2.

Scaling of Parameters

In practical parameter estimation problems, the unknown parameters may have very different orders of magnitude. ParamEDE therefore uses an internal scaling of the parameters. The employed initialization and update strategies for the scaling factors thereby resembles the heuristics used in Colsol-DDE (see Subsection 6.5.8). The computed scaling factors are then used in the solution of the linear constrained least-squares problems and for the computation of increment norms. The latter is relevant, in particular, in the termination criterion and for the computation of $\hat{w}$.

Estimation of κ

One of the conditions in the local contraction theorem (Theorem 11.6) is that there exists $\kappa < 1$ such that $\|\mathbf{J}^+(\tilde{z})R(y)\| \leqslant \kappa\|\tilde{z}-y\|$ for all y, $\tilde{z}$ in the considered domain, with $R(y) := F(y)+\mathbf{J}(y)\Delta y$. It was further discussed in Subsection 11.1.2 that $\kappa < 1$ can be interpreted as a condition on the quality of the measurement data, and that the existence of $\kappa < 1$ plays a key role regarding the statistical stability of the obtained solution.

Because of the significant role that κ plays in the convergence theory and the analysis of solutions, it is desirable to estimate it numerically. ParamEDE uses, when iterating from c^k to $c^{k+1} = c^k + \alpha^k \Delta c^k$, the following estimate:

$$\hat{\kappa} = \frac{\|\mathbf{J}^+_{[r]}(c^{k+1})R(c^k)\|_2}{\|c^{k+1} - c^k\|_2} = \frac{\|\mathbf{J}^+_{[r]}(c^{k+1})\left(F(c^k) + \mathbf{J}(c^k)\Delta c^k\right)\|_2}{\alpha^k \|\Delta c^k\|_2}. \tag{13.19}$$

Thereby, the rank r in the nominator is the same as the one that has been used for computing Δc^k. The Euclidean norms in both the nominator and in the denominator are computed with the same scaling factors.

Alternative estimation formulae for κ and their use in specific regularization strategies are discussed in Becker [22] and Hass [142].

Analysis of Solutions

In Section 12.2 it has been discussed that the matrix

$$\tilde{\mathbf{V}}_c := \mathbf{J}^+(\hat{c}) \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} (\mathbf{J}^+)^T(\hat{c}) \tag{13.20}$$

can be used as an approximation of the covariance of the estimates as a function of the (random) measurement data (cf. equation (12.17)).

Given a decomposition of $\mathbf{J}(c^k)$ as in equation (13.15), ParamEDE computes for $s_{n_c - n_{F_2}} > 0$

the matrix $\tilde{\mathbf{V}}_c$ by

$$\tilde{\mathbf{V}}_c = \mathbf{Q}^T \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^\dagger (\mathbf{S}^\dagger)^T \end{pmatrix} \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^T \end{pmatrix} \mathbf{Q}. \tag{13.21}$$

This matrix can subsequently be used to compute approximations of confidence intervals for all parameters, see Lemma 12.7.

It is further observed that equation (13.18) can be rewritten as

$$\underbrace{\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}^T \end{pmatrix} \mathbf{Q}}_{=:\mathbf{\Omega}} \Delta c^k = - \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{S}^\dagger \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{L}^{-1} \\ \mathbf{1} & -\mathbf{A}\mathbf{L}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{U}^T & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix} \begin{pmatrix} F_1(c^k) \\ F_2(c^k) \end{pmatrix} \tag{13.22}$$

Therefore, in a solution $\hat{c}$, the first n_{F_2} rows of $\mathbf{\Omega}$ give the parameter combinations that are locally determined by the equality constraints. The lower $n_c - n_{F_2}$ rows of $\mathbf{\Omega}$ combinations are locally determined by the least-squares conditions. Further, the $n_{F_2}+1$-st row corresponds to the "best determined parameter combination", and the n_c-th row corresponds to the "least determined parameter combination".

Because of the relevance of the matrix $\mathbf{\Omega}$ for the analysis of the obtained solution it is provided to the user of ParamEDE.

Let further $e_i \in \mathbb{R}^{n_c}$ be the unit vector into the direction of the i-th coordinate, and compute

$$z = \mathbf{\Omega} e_i. \tag{13.23}$$

Then

$$\rho_1 = \sum_{i=1}^{n_{F_2}} (z_i)^2, \quad \rho_2 = \sum_{i=n_{F_2}+1}^{r} (z_i)^2, \quad \rho_3 = \sum_{i=r+1}^{n_c} (z_i)^2 \tag{13.24}$$

are the projections of z on the subspaces spanned by those parameter combinations that are locally determined by the equality constraints, by the "large" singular values and by the "small" singular values, respectively. Thereby, the meaning of "large" and "small" is defined according to the rank decision in the last iteration. ParamEDE provides the values of ρ_1, ρ_2, and ρ_3 for each individual parameter c_i, $1 \leqslant i \leqslant n_c$, in order to assist the user in the analysis of the parameter estimates.

Part V.

Numerical Investigations

14. Solution of IHDDE-IVPs

> *A good test problem should have an analytic solution. Finding interesting and nontrivial functional differential equations whose solution is known is sometimes a difficult task.*
>
> Neves [193], commenting on the difficulty to find suitable test problems for initial value problem solvers.

This chapter presents results for the numerical solution of initial value problems (IVPs) in differential equations with time delays, with discontinuities in the right-hand-side function, and with impulses.

Numerical Results Presented in This Chapter

The presented results are related to four topics.

The first issue addressed in this chapter is the current lack of established test problems for the challenging class of differential equations considered in this thesis. Finding interesting and non-trivial test problems with known analytic solutions is indeed difficult (see quote by Neves above). Therefore accurate numerical solutions are provided as reference values for several IVPs with unknown analytic solution.

The second issue is related to the modified standard approach, which has been introduced in Subsection 5.2.2. If a current trial step of the integration method is such that the deviating arguments cross discontinuities in the past, then the modified standard approach employs smooth extrapolations beyond past discontinuities for computing the past states. In this chapter, a numerical example is given that demonstrates the benefit of using the modified standard approach rather than the standard approach in an algorithm for locating propagated discontinuities.

Furthermore, this chapter includes a simulation study that assesses the influence of some of the parameters in the model for the voting behavior of the viewers of the TV singing competition "Unser Star für Baku" (see Section 3.3). It is demonstrated that the use of a time delay in the differential equation model is crucial in order to make the simulation results qualitatively consistent with the observations in the TV show.

Eventually, the performance of the newly developed solver Colsol-DDE (see Chapter 6) is investigated. In particular, convergence of the results obtained with Colsol-DDE is demonstrated in the limit of small tolerances, and the capabilities and limitations of the implemented methods for solving stiff problems are studied.

Organization of This Chapter

In Section 14.1, accurate numerical reference solutions are given and the performance of the methods implemented in Colsol-DDE is investigated. Section 14.2 shows how localization of discontinuities works with the modified standard approach and points out the advantages over the use the standard approach. The simulation study for the model of the voting behavior of the viewers of the TV singing competition "Unser Star für Baku" is presented in Section 14.3.

Notation

In this chapter, the notation $y(t)$ (instead of $y(t;c)$) is used for the state because the dependency of the IVP solution on parameters in the model functions is not addressed here.

14.1. Accurate Reference Solutions and Performance of Colsol-DDE

This section gives accurate numerical reference solutions for IVPs in delay differential equations (DDEs), impulsive delay differential equations (IDDEs), and impulsive hybrid discrete-continuous delay differential equations (IHDDEs). The reference solutions are obtained by using Colsol-DDE and are validated by using other IVP solvers.

The convergence behavior of the methods implemented in Colsol-DDE is analysed in the limit of small relative tolerances. This analysis makes use of the given reference solutions. Furthermore, the two-stage Radau IIA method and the three-stage Lobatto IIIA method are applied to a stiff DDE-IVP. The chosen stepsizes are analysed and put into relation with the stability properties of the methods (see Section 6.7).

In general, the accuracy of numerically computed IVP solutions depends on the relative tolerance σ_{tol}^{rel}, on the absolute tolerance σ_{tol}^{abs} and on the "zero criterion" γ_{crit} that is used in the strategy for locating zeros of state-dependent switching functions, and also for locating zeros of the propagation switching functions that correspond to state-dependent delays (see Subsection 6.6.3 and Section 6.9 for details). Unless otherwise noted, very small values are used for both σ_{tol}^{abs} and γ_{crit} for all numerical computations presented in this chapter. In particular, this is the case for the computation of reference results and for the computations done for convergence analyses.

14.1.1. DDE with State-Dependent Delay

Problem Definition

The following differential equation is considered as an introductory example:

$$\dot{y}(t) = y(y(t)). \tag{14.1}$$

The right-hand-side function of this differential equation can also be expressed as $y(t-\tau(t,y(t)))$, with the state-dependent delay defined by

$$\tau(t,y(t)) = t - y(t). \tag{14.2}$$

The employed initial conditions are

$$y(2) = 1 \tag{14.3a}$$

$$y(t) = \frac{1}{2} \quad \text{for} \quad t < 2. \tag{14.3b}$$

This means that the initial time is $t^{ini} = 2$, the initial value is $y^{ini} = 1$, and the initial function is $\phi(t) \equiv 1/2$. Note that the initial function does not link continuously to the initial value, i.e. $\phi(t^{ini}) \neq y^{ini}$. The final time is set to $t^{fin} = 5.5$. Thus the IVP (14.1), (14.3) is considered on the interval $\mathcal{T} = [2, 5.5]$.

Categorization and References

The IVP (14.1), (14.3) is a DDE-IVP with one state-dependent delay, and it is due to Paul [201], see also Paul [202].

Analytic Solution

Among all IVPs considered in this chapter, the DDE-IVP (14.1), (14.3) is the only one for which an analytic solution is known. The analytic solution is given in Paul [201, 202]:

$$y(t) = \begin{cases} \frac{1}{2}t & 2 \leqslant t \leqslant 4 \\ 2\exp(\frac{1}{2}t - 2) & 4 \leqslant t \leqslant \ln(4\exp(4)) \\ -2\ln\big(\exp(-2)\,[1 + \ln(4\exp(4)) + t]\big) & \ln(4\exp(4)) \leqslant t \leqslant \frac{1}{2} + \ln(4\exp(4)) \end{cases}. \tag{14.4}$$

It is noted that $\ln(4\exp(4)) \approx 5.386294361119891$, and thus $\frac{1}{2} + \ln(4\exp(4)) = 5.886294361119891$. The final time $t^{fin} = 5.5$ has been chosen such that it lies within the interval for which an

analytic solution is known. A plot of the solution within the considered interval $[2, 5.5]$ is given in Figure 14.1.

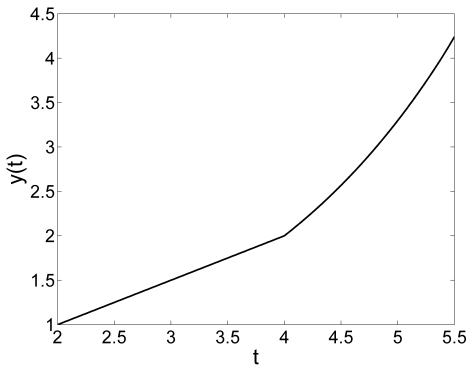


Figure 14.1.: Solution of the DDE-IVP (14.1), (14.3). The discontinuity of order 1 at t=4 is clearly visible.

Convergence Behavior

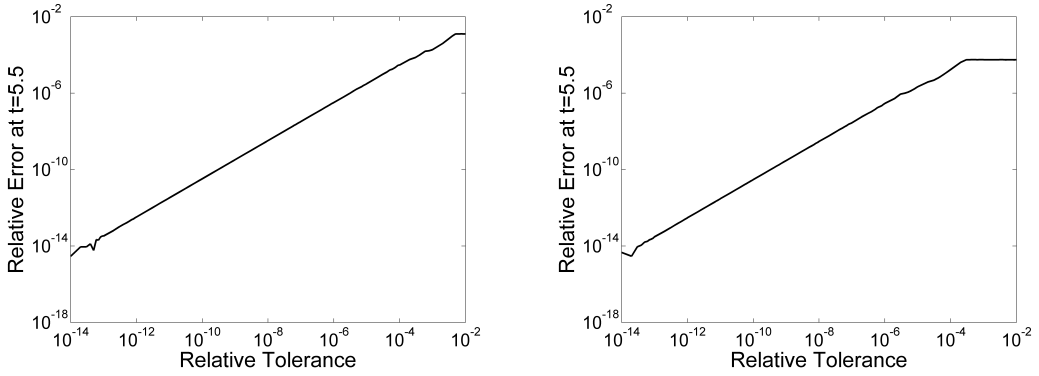
The convergence behavior of the numerical methods implemented in Colsol-DDE is studied. For this purpose, the relative tolerance is varied over several orders of magnitude as follows: $\sigma_{tol}^{rel} = 10^{-2}$, and $\sigma_{tol}^{rel} = n \cdot 10^{-m}$, with $n \in \{1, 2, \ldots, 9\}$ and $m \in \{3, \ldots, 14\}$.

For all above-given values of the relative tolerance, the DDE-IVP (14.1), (14.3) is solved. This gives, for each choice of σ_{tol}^{rel}, a numerical result $\eta(5.5)$. Then the relative error of the numerical solution at the final time $t^{fin} = 5.5$ is computed as follows:

$$\epsilon_{rel} = \frac{|\eta(5.5) - y(5.5)|}{|y(5.5)|}. \tag{14.5}$$

The obtained relative errors as a function of the relative tolerance are shown in Figure 14.2. Figure 14.2a (left) displays the results for the two-stage Radau IIA method implemented in Colsol-DDE, and Figure 14.2b (right) displays the results for the three-stage Lobatto IIIA method implemented in Colsol-DDE.

For both methods, a very good proportionality of the relative error at the final time to the relative tolerance is obtained for a wide range of relative tolerances. For small tolerances, small violations from the error-tolerance proportionality are observed only when the relative error is approximately 10^{-14} and thus close to the reachable machine precision $\epsilon_{mach} \approx 10^{-16}$.



(a) (b)

Figure 14.2.: Convergence of the results obtainted with Colsol-DDE to the exact solution of the DDE-IVP (14.1), (14.3): (a) relative errors ϵ_{rel}^{nom} obtained with the two-stage Radau IIA method, (b) relative errors ϵ_{rel}^{nom} obtained with the three-stage Lobatto IIIA method.

For crude tolerances, it is observed that the relative error levels off at a certain value. For the Lobatto method, this happens for $\sigma_{tol}^{rel} \gtrsim 10^{-4}$, and for the Radau method, this happens for $\sigma_{tol}^{rel} \gtrsim 5 \cdot 10^{-3}$. In both cases, the reason is that the variation of the tolerance in this regime does not affect the choice of the mesh. In fact, only 5 integration steps are taken for these crude tolerance values, and the choice of the stepsizes is only determined by the initial stepsize (which has here been chosen as $h_1 = (t^{fin} - t^{ini})/10 = 0.35$) and by the requirement to include the time points of the propagated discontinuities (approximately) into the mesh.

14.1.2. IDDE with State-Dependent Delay

Problem Definition

Consider the following modification of the DDE (14.1), which includes impulses:

$$\dot{y}(t) = y(y(t)) \qquad \text{for} \quad \zeta(t) = (\pm 1, \pm 1, \pm 1)^T \tag{14.6a}$$

$$y(t) = y^-(t) + \omega(\zeta(t)) \quad \text{else} \tag{14.6b}$$

As before, the right-hand-side function of the differential equation can be written as $y(t-\tau(t,y(t)))$ with the state-dependent delay

$$\tau(t, y(t)) = t - y(t). \tag{14.7}$$

Further, $\zeta(t) = (\zeta_1(t), \zeta_2(t), \zeta_3(t))^T$, and $\zeta_i(t)$ for $1 \leqslant i \leqslant 3$ are the signs of three switching functions:

$$\zeta_i(t) = \operatorname{sign}(\sigma_i(t, y^-(t), y^\bullet(t - \tau(t, y(t))))), \quad \text{for} \quad i = 1, 2, 3. \tag{14.8}$$

Here, three simple time-dependent switching functions are considered:

$$\sigma_i(t, y^-(t), y^\bullet(t - \tau(t, y(t)))) \equiv \sigma_i(t) \quad \text{for} \quad 1 \leqslant i \leqslant 3, \tag{14.9a}$$

$$\sigma_1(t) = t - 2.5, \quad \sigma_2(t) = t - 3.4, \quad \sigma_3(t) = t - 3.88. \tag{14.9b}$$

At the zeros of all three switching functions, impulses are applied. The following impulse function is used, which is independent of the time and of the states:

$$\omega(t, y(t), y(t - \tau(t, y(t)))) \equiv \omega(\zeta(t)) \tag{14.10a}$$

$$\omega(\zeta(t)) = \begin{cases} 0.7 & \text{if} \quad \zeta(t) = (0, \pm 1, \pm 1)^T \\ -1 & \text{if} \quad \zeta(t) = (\pm 1, 0, \pm 1)^T \\ -1.5 & \text{if} \quad \zeta(t) = (\pm 1, \pm 1, 0)^T \end{cases}. \tag{14.10b}$$

The initial conditions that are used here are the same as in Subsection 14.1.1:

$$y(2) = 1 \tag{14.11a}$$

$$y(t) = \frac{1}{2} \quad \text{for} \quad t < 2. \tag{14.11b}$$

As before, the model functions are $t^{ini} = 2$, $y^{ini} = 1$, $\phi(t) \equiv 1/2$, and $t^{fin} = 5.5$. This implies that the IVP (14.6), (14.10), (14.11) is considered on the interval $\mathcal{T} = [2, 5.5]$.

Categorization

The IVP (14.6), (14.10), (14.11) is an IDDE-IVP with one constant delay and with three simple time-dependent switching functions.

Numerical Reference Solution

A numerically computed reference solution $\eta^{ref}(5.5)$ is given by

$$y(5.5) \approx \eta^{ref}(5.5) = 4.303485743099. \tag{14.12}$$

This reference result has been obtained by using both Colsol-DDE and the Matlab solver dde23 (see Shampine and Thompson [233]). Using a sequence of stringent tolerances, the solvers yielded the same result in the leading 13 digits given above.

A plot of the numerical reference solution is given in Figure 14.3. The three impulses at the time points $t = 2.5$, at $t = 3.4$, and at $t = 3.88$ are clearly visible. Furthermore, there are many discontinuities of order 1 and higher. A complete list of all discontinuities is given in Table 14.1. In total, there are 4 discontinuities of order 0 (including the discontinuity at t^{ini}), 7 discontinuities of order 1, 8 discontinuities of order 2 and of order 3, 5 discontinuities of order 4 and 1 discontinuity of order 5.

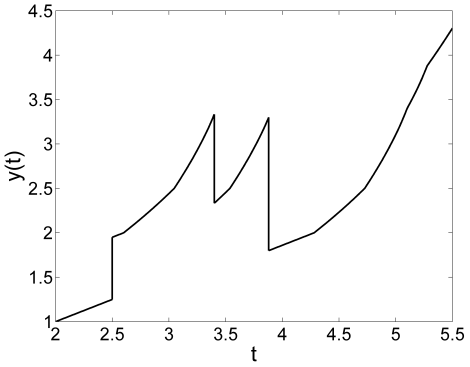


Figure 14.3.: Solution of the IDDE-IVP (14.6), (14.10), (14.11). There are three impulses (i.e. discontinuities of order 0) at $t = 2.5$, at $t = 3.4$, and at $t = 3.88$. In addition, there are many discontinuities of order 1 and higher, see Table 14.1.

Convergence Behavior

The convergence of the numerical results obtained with Colsol-DDE to the reference result $\eta^{ref}(5.5)$ is investigated. As in Subsection 14.1.1, the relative tolerance is varied over several orders of magnitude: $\sigma_{tol}^{rel} = 10^{-2}$, and $\sigma_{tol}^{rel} = n \cdot 10^{-m}$, with $n \in \{1, 2, \ldots, 9\}$ and $m \in \{3, \ldots, 14\}$.

For each value of the relative tolerance, the IDDE-IVP (14.6), (14.10), (14.11) is solved and the relative error in the numerical result $\eta(5.5)$ is determined by

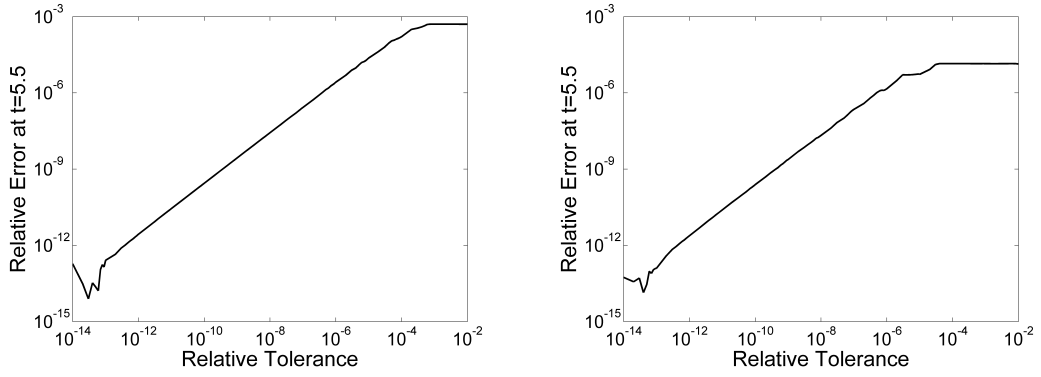
$$\epsilon_{rel} = \frac{|\eta(5.5) - \eta^{ref}(5.5)|}{|\eta^{ref}(5.5)|}. \tag{14.13}$$

Figure 14.4 displays the relative error as a function of the relative tolerance for both the two-stage Radau IIA method (left part, Figure 14.4a) and the three-stage Lobatto IIIA method (right part, Figure 14.4b) implemented in Colsol-DDE. A very good proportionality of the obtained numerical error at the final time to the relative tolerance is observed. For small tolerances, error-tolerance proportionality is observed up to the accuracy of the reference result.

For crude tolerances, the relative error levels off. Compared to the results presented in Figure 14.2, this "levelling off" occurs for smaller values of σ_{tol}^{rel}, and the obtained relative error is smaller. The reason for this phenomenon is that Colsol-DDE tracks discontinuities, i.e. the code includes all those discontinuities into the mesh whose order is less than or equal to the order of the discrete error-estimating method (see Section 6.3 and Section 6.6 for details). The code thus takes at least 33 integration steps independent of the chosen tolerance, which leads to a more accurate approximation of the solution even for very crude tolerances (in comparison to the results presented in Subsection 14.1.1).

Time Point of Discontinuity	Order	Discontinuity Type	Parent Discontinuity
2.000000000000	0	Initial	
2.500000000000	0	Root	
2.600000000000	1	Propagated	2.000000000000
3.046287102628	1	Propagated	2.500000000000
3.096922718597	2	Propagated	2.600000000000
3.296922718597	2	Propagated	3.046287102628
3.316784011696	3	Propagated	3.096922718597
3.387984216475	3	Propagated	3.296922718597
3.394450807608	4	Propagated	3.316784011696
3.400000000000	0	Root	
3.537332584912	1	Propagated	2.500000000000
3.587968200881	2	Propagated	2.600000000000
3.787968200881	2	Propagated	3.046287102628
3.807829493979	3	Propagated	3.096922718597
3.879029698759	3	Propagated	3.296922718597
3.880000000000	0	Root	
4.280235740772	1	Propagated	2.000000000000
4.726522843400	1	Propagated	2.500000000000
4.777158459369	2	Propagated	2.600000000000
4.977158459369	2	Propagated	3.046287102628
4.997019752467	3	Propagated	3.096922718597
5.068219957247	3	Propagated	3.296922718597
5.074686548380	4	Propagated	3.316784011696
5.096983677979	4	Propagated	3.387984216475
5.098939211900	5	Propagated	3.394450807608
5.100607935648	1	Propagated	3.400000000000
5.157470913611	2	Propagated	3.537332584912
5.177332206710	3	Propagated	3.587968200881
5.248532411489	3	Propagated	3.787968200881
5.254999002622	4	Propagated	3.807829493979
5.277296132222	4	Propagated	3.879029698759
5.277590305387	1	Propagated	3.880000000000
5.488442308087	2	Propagated	4.280235740772

Table 14.1.: List of all numerically determined discontinuities in the solution of the IDDE-IVP (14.6), (14.10), (14.11). The first column gives the time point of the determined discontinuity, the second one gives the order of the discontinuity. The third column provides the information whether the discontinuity is an initial discontinuity, a root discontinuity, or a propagated discontinuity. Eventually, for all propagated discontinuities, the last column lists the corresponding parent discontinuity.



(a) (b)

Figure 14.4.: Convergence of the results obtainted with Colsol-DDE to the reference solution of the IDDE-IVP (14.6), (14.10), (14.11): (a) relative errors ϵ_{rel}^{nom} obtained with the two-stage Radau IIA method, and (b) relative errors ϵ_{rel}^{nom} obtained with the three-stage Lobatto IIIA method.

14.1.3. Epidemiology: An IHHDE Model with State-Dependent Switching Functions

Problem Definition

In Section 3.1, an extension of the SEIRS epidemiological model by Cooke and van den Driessche [68] has been introduced. The extensions allow for an invasion of a healthy population by an infected population, for an improved medical treatment of the infected population after a new drug becomes available and for a vaccination of susceptibles after a vaccine is developed.

First, all differential equations from Section 3.1 are collected:

$$\dot{y}_1(t) = \begin{cases} b\tilde{Y}(t) - \lambda\frac{y_1(t)y_3(t)}{Y(t)} + \gamma y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) & \text{for } \zeta(t) = (\pm1,\pm1,-1,-1)^T \\ b\tilde{Y}(t) - \lambda\frac{y_1(t)y_3(t)}{Y(t)} + \tilde{\gamma} y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) & \text{for } \zeta(t) = (\pm1,\pm1,+1,-1)^T \\ b\tilde{Y}(t) - \lambda\frac{y_1(t)y_3(t)}{Y(t)} + \gamma y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) - \rho y_1(t) & \text{for } \zeta(t) = (\pm1,\pm1,-1,+1)^T \\ b\tilde{Y}(t) - \lambda\frac{y_1(t)y_3(t)}{Y(t)} + \tilde{\gamma} y_3(t-\tau_1)\exp(-d\tau_1) & \\ -dy_1(t) - \rho y_1(t) & \text{for } \zeta(t) = (\pm1,\pm1,+1,+1)^T \end{cases} \tag{14.14a}$$

$$\dot{y}_2(t) = \lambda\frac{y_1(t)y_3(t)}{Y(t)} - \lambda\frac{y_1(t-\tau_2)y_3(t-\tau_2)}{Y(t-\tau_2)}\exp(-d\tau_2) - dy_2(t) \quad \text{for } \zeta(t) = (\pm1,\pm1,\pm1,\pm1)^T \tag{14.14b}$$

$$\dot{y}_3(t) = \begin{cases} \lambda\frac{y_1(t-\tau_2)y_3(t-\tau_2)}{Y(t-\tau_2)}\exp(-d\tau_2) - (\epsilon+\gamma+d)y_3(t) & \text{for } \zeta(t) = (\pm1,-1,\pm1,\pm1)^T \\ \lambda\frac{y_1(t-\tau_2)y_3(t-\tau_2)}{Y(t-\tau_2)}\exp(-d\tau_2) - (\tilde{\epsilon}+\tilde{\gamma}+d)y_3(t) & \text{for } \zeta(t) = (\pm1,+1,\pm1,\pm1)^T \end{cases} \tag{14.14c}$$

$$\dot{y}_4(t) = \begin{cases} \gamma y_3(t) - \gamma y_3(t-\tau_1)\exp(-d\tau_1) - dy_4(t) & \text{for } \zeta(t) = (\pm1,-1,\pm1,\pm1)^T \\ \tilde{\gamma} y_3(t) - \gamma y_3(t-\tau_1)\exp(-d\tau_1) - dy_4(t) & \text{for } \zeta(t) = (\pm1,+1,-1,\pm1)^T \\ \tilde{\gamma} y_3(t) - \tilde{\gamma} y_3(t-\tau_1)\exp(-d\tau_1) - dy_4(t) & \text{for } \zeta(t) = (\pm1,+1,+1,\pm1)^T \end{cases} \tag{14.14d}$$

$$\dot{y}_5(t) = \begin{cases} 0 & \text{for } \zeta(t) = (\pm1,\pm1,\pm1,-1)^T \\ \rho y_1(t) & \text{for } \zeta(t) = (\pm1,\pm1,\pm1,+1)^T \end{cases} \tag{14.14e}$$

$$\dot{y}_6(t) = \begin{cases} (\epsilon+d)y_3(t) & \text{for } \zeta(t) = (\pm1,-1,\pm1,\pm1)^T \\ (\tilde{\epsilon}+d)y_3(t) & \text{for } \zeta(t) = (\pm1,+1,\pm1,\pm1)^T \end{cases}. \tag{14.14f}$$

$$y(t) = y^-(t) + \omega(\zeta(t)) \quad \text{for } \zeta(t) \neq (\pm1,\pm1,\pm1,\pm1)^T \tag{14.14g}$$

It is recalled that $Y(t) = y_1(t)+y_2(t)+y_3(t)+y_4(t)+y_5(t)$, and that $\tilde{Y}(t) = y_1(t)+y_2(t)+y_4(t)+y_5(t)$. There are five constant delays, i.e. the delays are given by

$$\tau_i(t, y(t)) \equiv \tau_i, \quad 1 \leqslant i \leqslant 5. \tag{14.15}$$

As explained in Section 3.1, it holds that $\tau_4 = \tau_1 + \tau_3$. Moreover, $\zeta(t) = (\zeta_1(t), \zeta_2(t), \zeta_3(t), \zeta_4(t))$, and $\zeta_i(t)$ for $1 \leqslant i \leqslant 4$ are the signs of switching functions σ_i:

$$\zeta_i(t) := \operatorname{sign}(\sigma_i(t, y^-(t), \{y^-(t-\tau_i)\}_{i=1}^5)) \quad \text{for} \quad 1 \leqslant i \leqslant 4. \tag{14.16}$$

One of the switching functions is simple time-dependent, whereas the other three are state-dependent:

$$\sigma_1(t, y(t), \{y(t-\tau_i)\}_{i=1}^5) \equiv \sigma_1(t) = t - s \tag{14.17a}$$

$$\sigma_2(t, y(t), \{y(t-\tau_i)\}_{i=1}^5) \equiv \sigma_2(y(t-\tau_3)) = y_6(t-\tau_3) - \varphi \tag{14.17b}$$

$$\sigma_3(t, y(t), \{y(t-\tau_i)\}_{i=1}^5) \equiv \sigma_3(y(t-\tau_4)) = y_6(t-\tau_4) - \varphi \tag{14.17c}$$

$$\sigma_4(t, y(t), \{y(t-\tau_i)\}_{i=1}^5) \equiv \sigma_4(y(t-\tau_5)) = y_6(t-\tau_5) - \varphi. \tag{14.17d}$$

The impulse function for this problem is defined by

$$\omega(t, y(t), \{y(t-\tau_i)\}_{i=1}^{5}, \zeta(t)) \equiv \omega(\zeta(t)) \tag{14.18a}$$

$$\omega(\zeta(t)) = \begin{cases} \begin{pmatrix} 0 & 0 & \nu & 0 & 0 & 0 \end{pmatrix}^T & \text{for} \quad \zeta(t) = (0, \pm 1, \pm 1, \pm 1)^T \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T & \text{else} \end{cases}, \tag{14.18b}$$

i.e. an impulse is only applied in the zero of the first switching function, and the impulse is independent of the time and independent of the states. In the zeros of the switching functions σ_2, σ_3, and σ_4 the state vector is continuous but the right-hand-side function f is discontinuous.

The differential equation system (14.14) and the impulse equation (14.18) are associated with the following initial condition:

$$y(t) = \begin{pmatrix} 100 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T \quad \text{for} \quad t \leqslant 0. \tag{14.19}$$

This means that $t^{ini} = 0$, $\phi(t) \equiv \phi(0) = y^{ini} = \begin{pmatrix} 100 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}^T$. Further, the final time is set to $t^{fin} = 350$, i.e. the considered time interval is $\mathcal{T} = [0, 350]$. The numerical values for the parameters in the differential equations (14.14) are given in Table 14.2.

Parameter	Description	Numerical Value
τ_1	time interval of immunization	42
τ_2	latency time	4
τ_3	time needed for drug development	30
τ_4	$= \tau_1 + \tau_3$	72
τ_5	time needed for vaccine development	140
b	birth rate	0.005
d	death rate independent of disease	0.004
λ	infection rate	0.2
ϵ	additional death rate due to disease (before drug development)	0.06
$\tilde{\epsilon}$	additional death rate due to disease (after drug development)	0.006
γ	recovery rate (before drug development)	0.04
$\tilde{\gamma}$	recovery rate (after drug development)	0.08
ρ	vaccination rate	0.03
s	arrival time of infected population	50
ν	size of infected population that arrives at s	2
φ	threshold number of deaths that triggers drug & vaccine developm.	5

Table 14.2.: Description and numerical values of parameters for simulation of the epidemiological model (14.14), (14.18).

Categorization

The IVP (14.14), (14.18), (14.19) is an IHDDE-IVP with five constant delays, with one simple time-dependent switching function and with four state-dependent switching functions.

Numerical Reference Solution

A numerical reference solution $\eta^{ref}(350)$ is given by

$$y(350) \approx \eta^{ref}(350) = \begin{pmatrix} 17.964515428508 \\ 0.044210519674 \\ 0.2564968120450 \\ 1.963180626737 \\ 70.99920119505 \\ 40.72925671756 \end{pmatrix}. \tag{14.20}$$

This reference result has been obtained by using both Colsol-DDE and the Matlab solver dde23 (Shampine and Thompson [233]) with a sequence of stringent tolerances. The results of the two solvers were identical in the leading $11 - 13$ digits given above.

A plot of the IVP solution is given in Figure 14.5. In the interval $[0, 50]$, the entire population is healthy, and thus $Y(t) = \tilde{Y}(t) = y_1(t)$, and the number of susceptibles $y_1(t)$ shows an exponential growth ($\dot{y}_1(t) = (b - d) \cdot y_1(t) = 0.001 y_1(t)$). Meanwhile, all other components of the state vector are identically zero, i.e. $y_i(t) \equiv 0$ for $2 \leqslant i \leqslant 6$.

At $t = 50$, i.e. at the zero of the simple time-dependent switching function σ_1, the infected population arrives. From this time on, the epidemics spreads within the population, which leads to a rapid decrease in $y_1(t)$ (number of susceptibles), and an increase of the size of the exposed, infected, and recovered population. The total number of deaths within the infected class reaches the threshold $\varphi = 5$ at $t \approx 78.2$. However, from that time on $\tau_3 = 30$ time units are needed until the new drug is available.

When the drug becomes available at $t \approx 108.2$, the right-hand-side functions of the differential equations for $y_3(t)$, $y_4(t)$, and $y_6(t)$ change discontinuously. In Figure 14.5, it is clearly visible that the size of the recovered population increases more rapidly, and that the number of deaths increases much slower than before. There is also a discontinuity in $\dot{y}_4(t)$ (rate of change of the number of infected individuals), but the "kink" is not clearly visible in the plot because $\tilde{\epsilon} + \tilde{\gamma} = 0.086 \approx \epsilon + \gamma = 0.1$. This means that the sum of revovery and death rate remains almost the same as before; in fact it decreases slightly, leading to a slightly more rapid increase of the number of infected individuals.

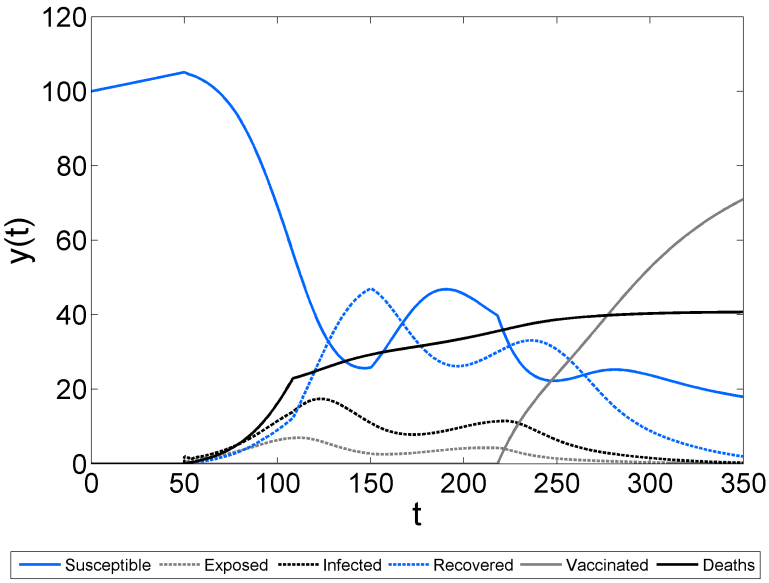


Figure 14.5.: Solution of the IHDDE-IVP (14.14), (14.18), (14.19), which simulates the spread of an epidemic. Population in each of the five classes: susceptible (blue, solid line), exposed (gray, dashed line), infected (black, dashed line), recovered (blue, dashed line), vaccinated (gray, solid line). In addition, the total number of deaths in the infected class is displayed (black, solid line).

After an additional 42 time units, at $t \approx 150.2$, the switching function σ_3 becomes zero. At this time, the fraction of the population that has recovered thanks to the drug at $t \approx 108.2$ becomes susceptible again. This leads to a significant increase in the number of susceptibles at $t \approx 150.2$ and to a corresponding significant decrease in the number of recovered.

Eventually, at $t \approx 218.2$, the vaccine becomes available. This leads to discontinuities in $\dot{y}_1(t)$ and $\dot{y}_5(t)$, such that the size of the susceptible population starts to decrease rapidly and the size of the vaccinated population starts to increase rapidly. Thanks to the vaccination of the population, the number of exposed and infected monotonically decreases for $t \gtrsim 218.2$, such that the epidemic is eventually defeated. Correspondingly, the number of deaths levels off when t approaches the final time $t^{fin} = 350$.

The total population $Y(t)$ is plotted as a function of time in Figure 14.6. Clearly visible is the

impulse at $t = 50$, the time point when the infected population arrives. From then on, the total population drops quickly, approximately from 107 to 90. At $t \approx 108.2$, the new drug is available, and the decrease in the total population is slowed down. However, the total population starts to increase only at approximately $t \approx 250$, i.e. some time after the vaccine has become available. At the final time $t^{fin} = 350$, the total population has not yet reached its initial value 100.

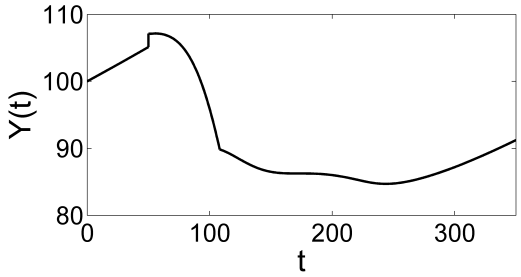


Figure 14.6.: Solution of the IHDDE-IVP (14.14), (14.18), (14.19): total population.

Convergence Behavior

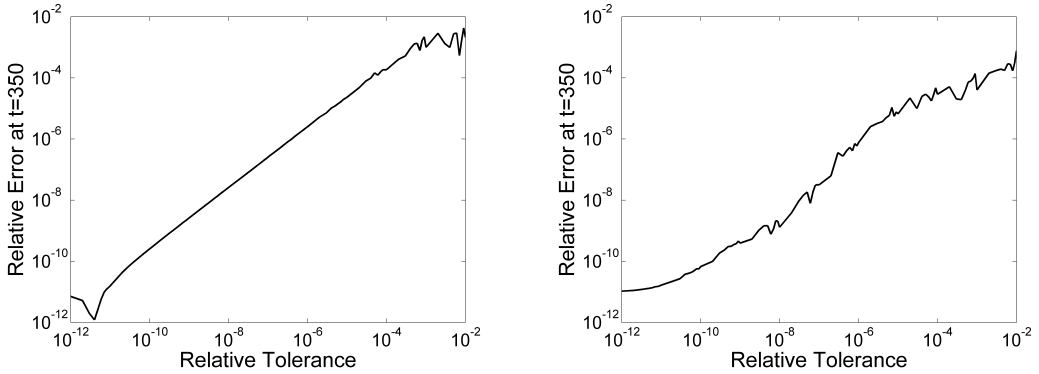
As a last step in the discussion of the epidemiological model, the convergence behavior of the numerical results obtained with Colsol-DDE for the IHDDE-IVP (14.14), (14.18), (14.19) is investigated. For this purpose, the relative tolerance is varied as follows: $\sigma_{tol}^{rel} = 10^{-2}$, and $\sigma_{tol}^{rel} = n \cdot 10^{-m}$, with $n \in \{1, 2, \ldots, 9\}$ and $m \in \{3, \ldots, 12\}$.

For each value of the relative tolerance, the IHDDE-IVP is solved, which yields a numerical approximation $\eta(350)$ of $y(350)$. The relative error in the solution is then computed as

$$\epsilon_{rel} = \max_{1 \leqslant i \leqslant 6} \left(\frac{|\eta_i(350) - \eta_i^{ref}(350)|}{|\eta_i^{ref}(350)|} \right), \tag{14.21}$$

where the maximum is taken over all 6 components of the state vector.

Figure 14.7 shows the relative error as a function of the relative tolerance. A good proportionality of the relative error to the relative tolerance is observed for both the two-stage Radau IIA and the three-stage Lobatto IIIA method, which indicates a good reliablity of the code. With both methods, the reference result can be reproduced up to a relative precision of 10^{-11}, which is the accuracy of the second component of the reference solution.



(a) (b)

Figure 14.7.: Convergence of the results obtainted with Colsol-DDE to the reference solution of the IHDDE-IVP (14.14), (14.18), (14.19): (a) relative errors ϵ_{rel}^{nom} obtained with the two-stage Radau IIA method, and (b) relative errors ϵ_{rel}^{nom} obtained with the three-stage Lobatto IIIA method.

The convergence behavior of the Lobatto IIIA method (Figure 14.7b) shows some irregularities in the sense that the relative error is not an almost perfectly linear function of σ_{tol}^{rel} as it is observed

for the Radau IIA method (Figure 14.7a). Please note that this does not contradict Theorem 5.23, because this theorem only guarantees that a bound on the global error (i.e. the maximum error on the whole considered time interval) is proportional to the relative tolerance.

14.1.4. Stiff Test Problem

Problem Definition

In this subsection, numerical results are presented for the following differential equation:

$$\dot{y}_1(t) = -k_1 y_1(t) + k_3 y_2(t-\tau) y_3(t) \tag{14.22a}$$

$$\dot{y}_2(t) = k_1 y_1(t) - k_3 y_2(t-\tau) y_3(t) - k_2 (y_2(t))^2 \tag{14.22b}$$

$$\dot{y}_3(t) = +k_2 (y_2(t))^2 \tag{14.22c}$$

The differential equation system contains one constant delay, i.e.

$$\tau(t, y(t)) \equiv \tau. \tag{14.23}$$

The differential equation system (14.22) is associated with the following initial condition:

$$y(t) = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T \quad \text{for} \quad t \leqslant 0, \tag{14.24}$$

which means that $t^{ini} = 0$, $\phi(t) \equiv \phi(0) = y^{ini} = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$. The final time is set to $t^{fin} = 10000$, such that the IVP is considered on the interval $\mathcal{T} = [0, 10000]$. The constants are set as follows: $k_1 = 0.04$, $k_2 = 3 \cdot 10^5$, $k_3 = 100$. The delay is given by $\tau = 0.01$.

Categorization and References

The IVP (14.22), (14.24) is a DDE-IVP with one constant delay. The ODE variant (i.e. set $\tau = 0$) is motivated as a model for a chemical reaction and has frequently been used as a standard stiff test problem for ODE-IVP solvers, see e.g. Robertson and Williams [217], Enright, Hull, and Lindberg [97], and Hairer and Wanner [127]. The variant with delay has been proposed by Guglielmi and Hairer [122] as a stiff DDE test problem.

Numerical Reference Solution

A numerical reference solution $\eta^{ref}(10000)$ is given by

$$y(10000) \approx \eta^{ref}(10000) = \begin{pmatrix} 2.08929059712 \cdot 10^{-3} \\ 8.3536874368 \cdot 10^{-7} \\ 0.9979098740341. \end{pmatrix} \tag{14.25}$$

This reference solution has been obtained by solving the stiff DDE-IVP (14.22), (14.24) with both Colsol-DDE and RADAR5 (see Guglielmi and Hairer [122, 123]). Using a sequence of stringent tolerances in both solvers, the obtained results were identical in the leading $11 - 13$ digits given above.

A plot of the trajectories $y_i(t)$, $1 \leqslant i \leqslant 3$, for $t \in [0, 10000]$, is given in Figure 14.8. It can be observed that the first chemical species is consumed (monotonically decreasing concentration $y_1(t)$), and that the third chemical species is produced (monotonically increasing concentration $y_3(t)$). Only very small amounts of the intermediate product, represented by $y_2(t)$, are available during the entire time horizon, because it quickly reacts to $y_1(t)$ and $y_3(t)$ ($k_2 \gg k_1$ and $k_3 \gg k_1$).

Application of Colsol-DDE to the Stiff DDE-IVP (14.22), (14.24)

The stiff DDE-IVP (14.22), (14.24) is suitable for investigating the capabilities and limitations of the collocation methods that are implemented in Colsol-DDE.

As seen in Figure 14.8, all three components of the state vector eventually approach a steady state. Therefore, if a numerical method is sufficiently stable in order to deal with the stiffness in the DDE-IVP, it is expected that it takes larger and larger stepsizes toward the final time (cf. Guglielmi and Hairer [122]).

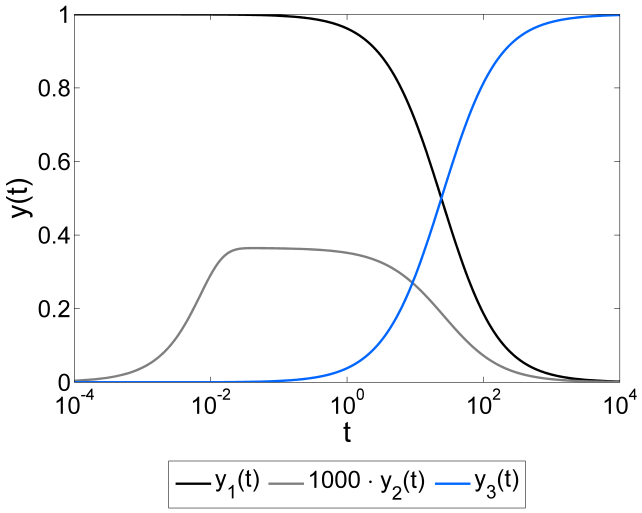
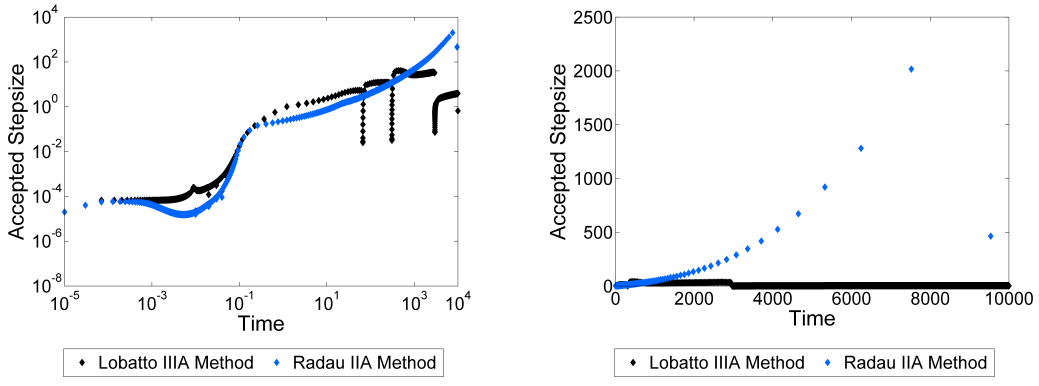


Figure 14.8.: Solution of the stiff DDE-IVP (14.22), (14.24). Note that the horizontal axis is logarithmic in order to better resolve the rapid transient behavior in the beginning of the considered interval. Note further that the state vector component $y_2(t)$ is amplified by a factor of 1000.

In order to test whether this is the case, the IVP is solved by using both the two-stage Radau IIA method and the three-stage Lobatto IIIA method implemented in Colsol-DDE. A moderate relative tolerance $\sigma_{tol}^{rel} = 10^{-3}$ is used, and the absolute tolerance is set to $\sigma_{tol}^{abs} = 10^{-8}$.

Figure 14.9 displays, at each mesh point t_k, the stepsize h_{k+1} that was accepted to proceed to the next mesh point $t_{k+1} = t_k + h_{k+1}$. The two-stage Radau IIA method takes larger and larger stepsizes toward the final time, with the largest stepsize being greater than 2000. At $t = k \cdot 10^{-2}$, $k \in \{1, 2, 3, 4\}$, some steps are taken with smaller stepsizes than the neighboring steps. The reason for this is that discontinuity points are included into the mesh. Furthermore, the last stepsize is short compared to the second-to-last one in order to stop at the final time $t^{fin} = 10000$.

For the three-stage Lobatto IIIA method, a less regular behavior of the sequence of accepted stepsizes is observed. In particular, the Lobatto IIIA method takes for $t \gtrsim 3000$ – where the steady state is almost reached – integration steps that are small compared to the integration steps in the Radau IIA method. This behavior can possibly be attributed to the fact that the three-stage Lobatto IIIA method has, compared to the two-stage Radau IIA method, less favorable stability properties, see Section 6.7.



(a) (b)

Figure 14.9.: Accepted stepsizes for solving the stiff DDE-IVP (14.22), (14.24), (a) in double logarithmic scaling and (b) in linear scaling. The stepsizes that are taken by the two-stage Radau IIA method are plotted in blue, and the stepsizes that are taken by the three-stage Lobatto IIIA method are plotted in black.

Despite the fact that the lack of stability appears to be a limiting factor for the performance of the three-stage Lobatto IIIA method, the method is still much faster than explicit methods. For example, for the above-given absolute and relative tolerances, Colsol-DDE needs 2840 accepted integration steps and $0.2s$ of computation time with the three-stage Lobatto IIIA method (1311 accepted integration steps and $0.1s$ for the two-stage Radau IIA method). In comparison, DDE_SOLVER by Thompson and Shampine [246] – as an example for a code that is based on explicit methods – needs 189946 (accepted) integration steps and $167s$ of computation time.[1]

14.2. Accuracy and Efficiency of the Modified Standard Approach for Locating Discontinuities

A key element of the modified standard approach is to use smooth extrapolations if a current trial integration step is such that the deviating argument crosses discontinuities in the past. The use of extrapolations has also been proposed in Guglielmi and Hairer [122], ZivariPiran [271], ZivariPiran and Enright [272], and Ernst [101]. However, to the knowledge of the author, the advantage of using extrapolations has not yet been demonstrated on a practical example. This is done in the following.

Problem Definition

Consider the DDE-IVP

$$\dot{y}(t) = y(y(t)) \tag{14.26a}$$
$$y(2) = 0.1 \tag{14.26b}$$
$$y(t) = 1 + \sin\left((t-1)\cdot\frac{\pi}{4} - \frac{\pi}{8}\right) \quad \text{for} \quad t < t^{ini} = 2 \tag{14.26c}$$

on the interval $\mathcal{T} = [2, 10]$. The differential equation is the same as in Subsection 14.1.1, but the initial value, the initial function, and the final time are different. In particular, the choice of a sine-like initial function ensures that locating the time point of the child discontinuity of the discontinuity at t^{ini} becomes harder compared to the constant initial function used in Subsection 14.1.1.

Numerical Reference Solution

A highly accurate DDE-IVP solution has been computed by applying Colsol-DDE with stringent tolerances. The obtained solution is plotted in Figure 14.10. There is only one propagated discontinuity within the considered time interval at $t \approx 6.453$, and this propagated discontinuity is of order 1 in y.

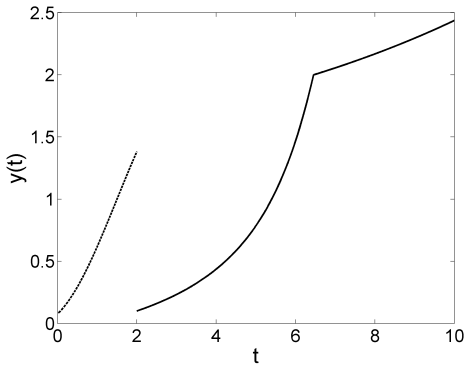


Figure 14.10.: Plot of the solution of the DDE-IVP (14.26) (solid line). The dashed line plotted for $t < 2$ shows the employed initial function.

[1] On an Intel i7 960 cpu with a frequency of 3.2GHz and 8MB cache, compilation with g95 compiler and identical choice of compiler flags.

14.2.1. Discontinuity Location during an Integration with Moderate Relative Tolerance

Consider the situation that the DDE-IVP should be solved with a moderate relative tolerance $\sigma_{tol}^{rel} = 10^{-3}$. Colsol-DDE (three-stage Lobatto IIIA method) reaches, after 6 successful integration steps, the mesh point $t_6 = 6.336915699252338$, and the numerical approximation of the IVP solution at that time is given by $y_6 = 1.845963562901120$.

With the approximation y_6 at the mesh point t_6, the "optimal" stepsize is given by $h_7^{opt} = 0.1162441676165301$. "Optimality" is thereby meant in the following sense: If the three-stage Lobatto IIIA method in Colsol-DDE is used with this stepsize, then the continuous representation $\eta_7^{opt}(t)$ on the interval $[t_6, t_7^{opt}]$, $t_7^{opt} = t_6 + h_7^{opt} = 6.453159866868869$, is such that $y_7^{opt} = \eta_7^{opt}(t_7^{opt}) = 2.000000000000001$. This means that the deviating argument, evaluated along the numerical solution, reaches the past discontinuity point $t = 2$ (the initial time) up to machine precision at the end of the integration step. The continuous representation $\eta_7^{opt}(t)$ is plotted as a red line in Figure 14.11 (for $t \in [6.452, t_7^{opt}]$), and the point (t_7^{opt}, y_7^{opt}) is plotted as a red dot.

For DDE-IVPs with state-dependent delays, such as the one considered here, it is in general impossible to know the optimal stepsize a priori. In the particular example (14.26), the stepsize selection strategy in Colsol-DDE (see Section 6.6) proposes the value $h_7^0 = 0.6515050019875654$ as a stepsize for the next step. This stepsize leads to a time point $t_7^0 = t_6 + h_7^0$ that is well to the right of t_7^{opt}. In the following, it is discussed how discontinuity location works with the modified standard approach realized in Colsol-DDE. Furthermore, the alternative of using the standard approach is considered and compared to the modified standard approach.

Modified Standard Approach

When taking the integration step with stepsize h_7^0, Colsol-DDE first solves the nonlinear equations that arise in the three-stage Lobatto IIIA method, in the uniform order correction, and in the implicit quadrature rule (see Sections 6.1-6.5). The Newton method used for solving the nonlinear equations converges after 5, 5, and 3 iterations, respectively. This yields a discrete approximation $y_7^0 = 2.881694350084759$, and further a continuous representation $\eta_7^0(t)$ for $t \in [t_6, t_7^0]$. The continuous representation $\eta_7^0(t)$ is plotted as black line in Figure 14.11. Please note that the black line is, on the interval $[6.452, t_7^{opt}]$, very close to the continuous representation $\eta_7^{opt}(t)$ (red line) that is obtained with the optimal stepsize.

The code proceeds with the computation of error estimates, which obey the employed tolerance criterion. Colsol-DDE then checks the signs of the propagation switching function

$$\sigma^\alpha(y(t)) := y(t) - 2 \tag{14.27}$$

at the beginning and end of the integration step, see Section 6.9. This leads to the detection of a sign change because $\sigma^\alpha(y_6) \approx -0.154 < 0$ and $\sigma^\alpha(y_7^0) \approx 0.882 > 0$. Subsequently, regula falsi is used to determine an approximation of the root of

$$\sigma^\alpha(\eta_7^0(t)) = \eta_7^0(t) - 2. \tag{14.28}$$

The obtained approximation of the root is given by $t_7^1 = 6.453183048573098$. The point $(t_7^1, \eta_7^0(t_7^1))$ is plotted in Figure 14.11 as a black diamond, which is almost overlayed by the red dot that indicates the "optimal" point (t_7^{opt}, y_7^{opt}). More precisely, it holds that $|t_7^1 - t_7^{opt}| \approx 2.318 \cdot 10^{-5}$.

The result of the regula falsi method suggests to use $h_7^1 = 0.1162673493207601$ as a new stepsize in order to include the time point of the propagated discontinuity (approximately) into the mesh. Colsol-DDE then computes a step with this stepsize, which yields $y_7^1 = 2.000032053223695$ such that $\sigma^\alpha(y_7^1) \approx 3.2 \cdot 10^{-5}$. Since only a moderate relative tolerance $\sigma_{tol}^{rel} = 10^{-3}$ is requested, it is reasonable to accept this step, i.e. to set $t_7 = t_7^1$, $y_7 = y_7^1$, and to proceed with the integration. Alternatively, one may decide to go through an additional cycle, i.e. calling regula falsi again for proposing a new stepsize and computing an integration step with the new stepsize. Additional numerical investigations have indicated, however, that the error that is introduced due to an inaccurate localization of the discontinuity point is of the same order of magnitude than the error that is made anyway during the integration with the selected moderate relative tolerance $\sigma_{tol}^{rel} = 10^{-3}$. Hence, only one root finding by regula falsi and one repeated integration step allow to include the discontinuity point into the mesh with sufficient accuracy.

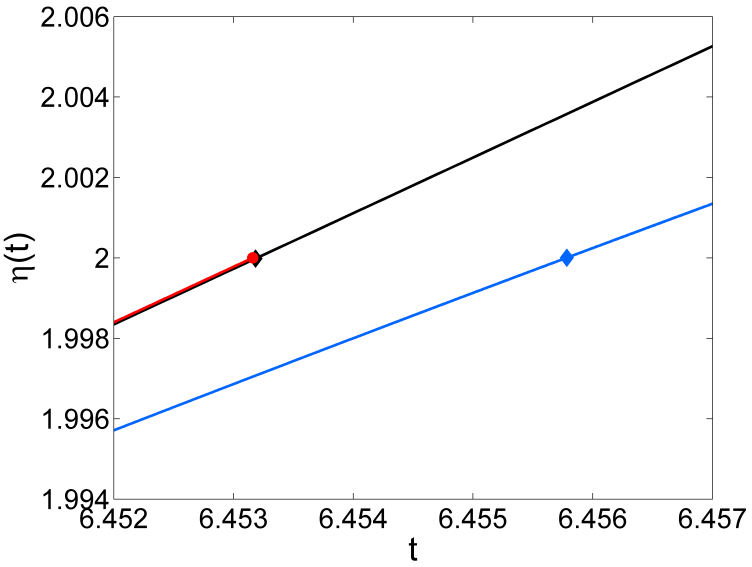


Figure 14.11.: Plot of continuous representations of the solution of the DDE-IVP (14.26) in the vicinity of t_7^{opt}. Red: Continuous representation $\eta_7^{opt}(t)$ for $t \in [6.452, t_7^{opt}]$. Black: Continuous representation $\eta_7^0(t)$ that is obtained by using the modified standard approach for computing past states (the red line and the black line partly overlay each other). Blue: Continuous representation $\tilde{\eta}_7^0(t)$ that is obtained by using the standard approach for computing past states.
The red dot indicates the zero of $\eta_7^{opt}(t) - 2$, and the black diamond and the blue diamond represent the zeros of $\eta_7^0(t)$ and $\tilde{\eta}_7^0(t)$, respectively. Clearly, the use of the modified standard approach allows for a much more accurate approximation of the propagated discontinuity as the standard approach.

Standard Approach

Consider now the case that the standard approach is used for approximating past states. This time, when taking the integration step with stepsize h_7^0, the Newton method fails to converge when attempting to solve the nonlinear equation for the uniform order correction (the method ends up in a two-cycle). As a remedy, smaller stepsizes can be tried out. For $\tilde{h}_7^0 = h_7^0/4 = 0.1628762504968901$, the Newton method successfully solves the nonlinear equations that arise in the collocation method, in the implicit uniform correction, and in the implicit quadrature rule. The obtained continuous representation $\tilde{\eta}_7^0(t)$ is displayed in blue in Figure 14.11. A root finding algorithm applied to the propagation switching function $\sigma^\alpha(\tilde{\eta}_7^0(t)) = \tilde{\eta}_7^0(t) - 2$ yields, approximately, the result $\tilde{t}_7^1 \approx 6.45578$.

The point $(\tilde{t}_7^1, \tilde{\eta}_7^0(\tilde{t}_7^1))$ is displayed as a blue diamond in Figure 14.11. Obviously, the error in the approximation of the discontinuity point by using the standard approach is much larger than the error that was obtained by using the modified standard approach. More precisely, it holds that $|\tilde{t}_7^1 - t_7^{opt}| \approx 2.620 \cdot \mathbf{10^{-3}} \gg |t_7^1 - t_7^{opt}| \approx 2.318 \cdot \mathbf{10^{-5}}$.

In order to determine the propagated discontinuity with an accuracy of $\approx 10^{-5}$, one or several additional trial steps will be needed, which requires additional computation time. In addition, computation time has already been spent in the attempt to take the larger stepsize h_7^0, for which the Newton method failed to converge.

These results demonstrate clearly that the modified standard approach is a more efficient strategy for locating propagated discontinuities than the standard approach.

14.3. Simulation Study: Voting Behavior of TV Viewers of "Unser Star für Baku"

In this section, simulation results are presented for the voting behavior of the viewers of the TV singing competition "Unser Star für Baku". The focus lies on a qualitative investigation of the role of certain parameters in the model introduced in Section 3.3, namely the laziness parameter λ, the delay τ, and the panic factor ρ. For all investigations in this section, it is assumed that there are three candidates, two of which may proceed to the next round while the last one is voted out.

Problem Definition

For convenience, the differential equation model introduced in Section 3.3 is recalled. The situation that is considered here is that two out of three candidates are selected to continue in the next round of the show while the last candidate is voted out. In this case, the state vector has four components, i.e. $y_i(t)$, $1 \leqslant i \leqslant 4$. The first three components represent the percentages of votes of the three candidates that are displayed in the livescore, while the fourth component represents the total number of votes. The differential equations are given by

$$\dot{y}_i(t) = 100 \cdot \frac{k_i \cdot g^{panic}(t,\zeta(t)) \cdot \beta_i(t, y(t-\tau), \zeta(t)) \cdot y_4(t) - \dot{y}_4(t) y_i(t) y_4(t)/100}{(y_4(t))^2} \quad \text{for} \quad 1 \leqslant i \leqslant 3 \tag{14.29a}$$

$$\dot{y}_4(t) = \sum_{i=1}^{3} \left(k_i \cdot g^{panic}(t,\zeta(t)) \cdot \beta_i(t, y(t-\tau), \zeta(t)) \right) \tag{14.29b}$$

There is one constant delay in this differential equation system:

$$\tau(t, y(t)) \equiv \tau \tag{14.30}$$

The symbol $\zeta(t)$ represents the signs of the switching functions: $\zeta(t) = (\zeta_1(t), \zeta_2(t), \zeta_3(t), \zeta_4(t))^T$, with

$$\zeta_i(t) = \operatorname{sign}(\sigma_i(t, y^-(t), y^-(t-\tau))) \quad \text{for} \quad i = 1,2,3,4. \tag{14.31}$$

One of the four switching functions is simple time-dependent and the other three are state-dependent:

$$\sigma_1(t, y(t), y(t-\tau)) \equiv \sigma_1(t) = t - t^{fin} - \delta \tag{14.32a}$$

$$\sigma_2(t, y(t), y(t-\tau)) \equiv \sigma_2(y(t-\tau)) = y_1(t-\tau) - y_2(t-\tau) \tag{14.32b}$$

$$\sigma_3(t, y(t), y(t-\tau)) \equiv \sigma_3(y(t-\tau)) = y_1(t-\tau) - y_3(t-\tau) \tag{14.32c}$$

$$\sigma_4(t, y(t), y(t-\tau)) \equiv \sigma_4(y(t-\tau)) = y_2(t-\tau) - y_3(t-\tau). \tag{14.32d}$$

The switching function signs $\zeta(t)$ are arguments to both the function g^{panic} and to the functions β_i (with $1 \leqslant i \leqslant 3$). The function g^{panic} is continuous and piecewise linear:

$$g^{panic}(t, \zeta(t)) = \begin{cases} 1 & \text{if} \quad \zeta(t) = (-1, \pm 1, \pm 1, \pm 1) \\ 1 + \frac{t - (t^{fin} - \delta)}{\delta} \rho & \text{if} \quad \zeta(t) = (+1, \pm 1, \pm 1, \pm 1) \end{cases}. \tag{14.33}$$

Only the sign of the first switching function is relevant for the definition of g^{panic}. The signs of the other three switching functions determine the result of the functions β_i:

$$\begin{aligned} &\beta_1(t, y(t-\tau), \zeta(t)) \\ &= \begin{cases} \frac{y_1(t-\tau)}{\frac{1}{2}[y_1(t-\tau) + y_2(t-\tau)]} & \text{for} \quad \zeta(t) = (\pm 1, -1, -1, -1) \\ \frac{y_1(t-\tau)}{\frac{1}{2}[y_1(t-\tau) + y_3(t-\tau)]} & \text{for} \quad \zeta(t) = (\pm 1, -1, -1, +1) \\ \exp\left(-\lambda (y_1(t-\tau) - \frac{1}{2}[y_1(t-\tau) + y_2(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, +1, -1, -1) \\ \exp\left(-\lambda (y_1(t-\tau) - \frac{1}{2}[y_1(t-\tau) + y_3(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, -1, +1, +1) \\ \exp\left(-\lambda (y_1(t-\tau) - \frac{1}{2}[y_2(t-\tau) + y_3(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, +1, +1, \pm 1) \end{cases}, \end{aligned} \tag{14.34}$$

and

$$\beta_2(t, y(t-\tau), \zeta(t)) = \begin{cases} \frac{y_2(t-\tau)}{\frac{1}{2}[y_2(t-\tau)+y_1(t-\tau)]} & \text{for} \quad \zeta(t) = (\pm 1, +1, -1, -1) \\ \frac{y_2(t-\tau)}{\frac{1}{2}[y_2(t-\tau)+y_3(t-\tau)]} & \text{for} \quad \zeta(t) = (\pm 1, +1, +1, -1) \\ \exp\left(-\lambda(y_2(t-\tau) - \frac{1}{2}[y_2(t-\tau) + y_1(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, -1, -1, -1) \\ \exp\left(-\lambda(y_2(t-\tau) - \frac{1}{2}[y_2(t-\tau) + y_3(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, +1, +1, +1) \\ \exp\left(-\lambda(y_2(t-\tau) - \frac{1}{2}[y_1(t-\tau) + y_3(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, -1, \pm 1, +1) \end{cases}, \tag{14.35}$$

and

$$\beta_3(t, y(t-\tau), \zeta(t)) = \begin{cases} \frac{y_3(t-\tau)}{\frac{1}{2}[y_3(t-\tau)+y_1(t-\tau)]} & \text{for} \quad \zeta(t) = (\pm 1, -1, +1, +1) \\ \frac{y_3(t-\tau)}{\frac{1}{2}[y_3(t-\tau)+y_2(t-\tau)]} & \text{for} \quad \zeta(t) = (\pm 1, +1, +1, +1) \\ \exp\left(-\lambda(y_3(t-\tau) - \frac{1}{2}[y_3(t-\tau) + y_1(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, -1, -1, +1) \\ \exp\left(-\lambda(y_3(t-\tau) - \frac{1}{2}[y_3(t-\tau) + y_2(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, +1, +1, -1) \\ \exp\left(-\lambda(y_3(t-\tau) - \frac{1}{2}[y_1(t-\tau) + y_2(t-\tau)])\right) & \text{for} \quad \zeta(t) = (\pm 1, \pm 1, -1, -1) \end{cases}. \tag{14.36}$$

It is remarked that the right-hand-side function of the DDE (14.29) is continuous in the zeros of the switching functions.

The differential equation (14.29) is associated with the following initial condition:

$$y(t) = \begin{pmatrix} 36 & 33 & 31 & 100000 \end{pmatrix}^T \quad \text{for} \quad t \leqslant 0. \tag{14.37}$$

This means that $t^{ini} = 0$ and that $\phi(t) \equiv \phi(0) = y^{ini} = (36, 33, 31, 100000)^T$. The final time is set to $t^{fin} = 600$, i.e. the considered time interval is $\mathcal{T} = [0, 600]$. The chosen initial conditions are interpreted as follows: At the beginning of the considered time interval, 100000 votes have been received. Candidate 1 has received 36% of the votes, candidate 2 has received 33% of the votes, and candidate 3 has received 31% of the votes.

The parameters in the differential equation system (14.29), and in the functions g^{panic} and β_i have the following meaning: k_i represents the size of the fan-base of candidate i and can be interpreted as a measure for the "true" singing performance of candidate i. Further, δ stands for the duration of the time interval at the end of the voting time in which viewers vote more frequently ("panic"). The parameter ρ is the amplification factor due to the panic at the end of the voting time, and λ characterizes the laziness of TV viewers if a candidate is currently on one of the two winning ranks. The delay τ represents the time that passes between the emission of the current livescore results from the TV studio until the votes of the TV viewers – as a reaction to the displayed results – are received and counted.

The following values for k_i and δ are used in all simulations:

$$k_1 = 40, \quad k_2 = 25, \quad k_3 = 18, \quad \delta = 120. \tag{14.38}$$

This means, in particular, candidate 1 has a bigger fan-base than candidate 2, who in turn has a bigger fan-base than candidate 3. The values of the other parameters τ, λ, and ρ are varied in order to study their influence.

Categorization

The IVP (14.29), (14.37) is an HDDE-IVP with one constant delay, one simple time-dependent switching function, and three state-dependent switching functions.

Integrator Settings

All numerical results presented in this section have been obtained with Colsol-DDE (three-stage Lobatto IIIA method), and with the following settings for the relative and absolute tolerance: $\sigma_{tol}^{rel} = 10^{-4}$, $\sigma_{tol}^{abs} = 10^{-8}$.

14.3.1. Influence of the Time Delay

In order to investigate the influence of the time delay, the IVP is solved for four different values of τ: 0, 15, 30, and 60. The laziness parameter is thereby fixed to $\lambda = 10$, and the panic factor is set to $\rho = 0$.

The results are displayed in Figure 14.12. If the delay is set to $\tau = 0$, the differential equation is an HODE rather than an HDDE. In this case, the percentages of votes for the three candidates approach a steady state. The differences in the percentages of votes at the end of the time interval are very small: $y_1(600) \approx 33.39$, $y_2(600) \approx 33.34$, $y_3(600) = 33.27$. However, the ranking order of the three candidates is the same on the entire time interval, i.e. the best candidate always stays on rank 1, and the worst candidate always stays on rank 3. The livescore yields a fair result, but it is also boring.

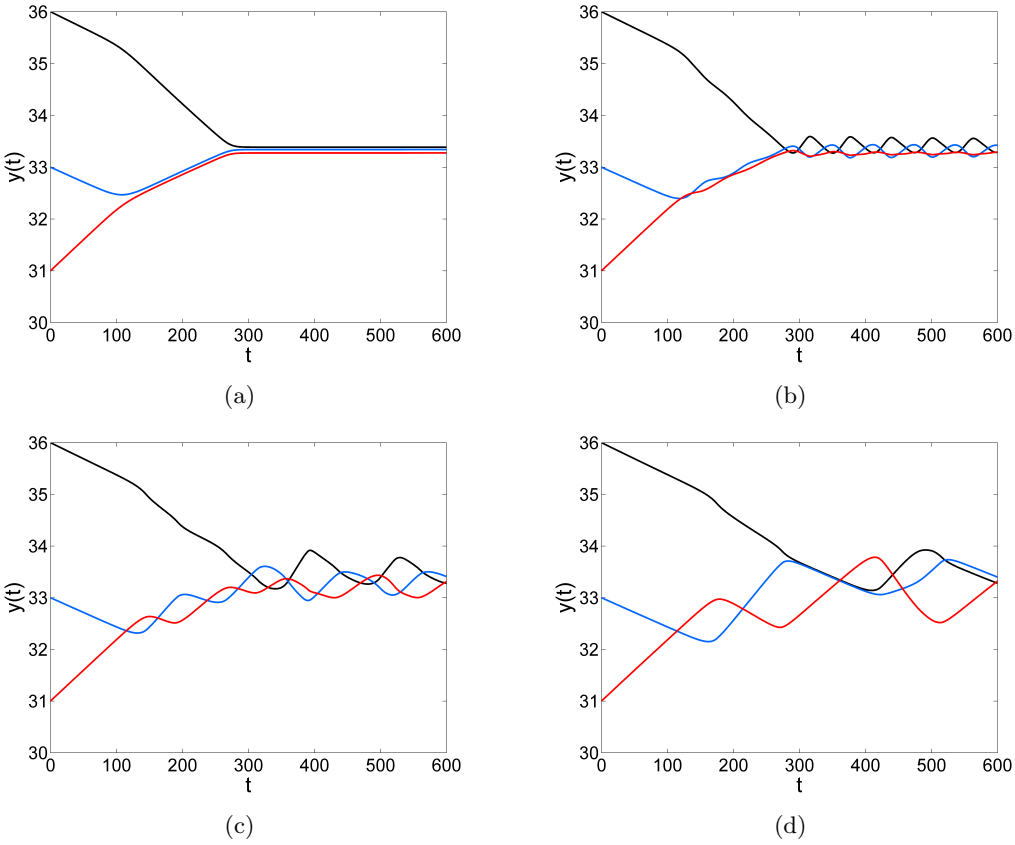


Figure 14.12.: Solution of the DDE-IVP (14.29), (14.37), which simulates the percentages of votes displayed in the livescore for various values of the delay τ: (a) $\tau = 0$, (b) $\tau = 15$, (c) $\tau = 30$, (d) $\tau = 60$. The black, blue, and red lines show the simulated percentages of votes for candidate 1, 2, and 3, respectively.

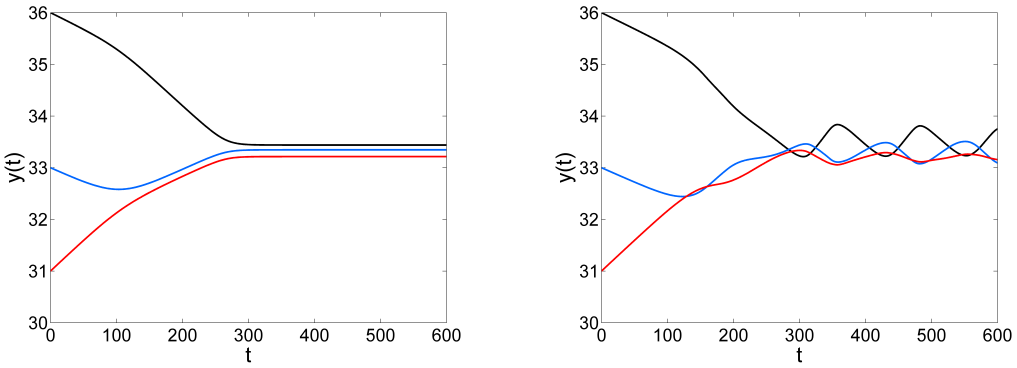
The situation changes if a time delay is introduced. For all three tested values of the delay, the ranking order of the candidates changes over time, which makes the results qualitatively consistent with the observations in the TV show. The changes in the ranking order are a combined effect of the laziness and the time delay: The viewers only vote for a candidate on rank 1 or 2 if he or she is in danger to drop to rank 3. However, by the time when the viewers react it might already be too late. This leads to oscillations in the percentages of votes for the three candidates, and the amplitudes of the oscilations become larger vor increasing values of the time delay τ. Accordingly, a time delay in the voting procedure makes the livescore more interesting, but also unfair.

14.3.2. Influence of Laziness

In order to study the influence of the laziness parameter, IVP solutions for $\lambda = 5$ are computed for $\tau = 0$ and $\tau = 30$. The panic factor is set to $\rho = 0$.

For the non-delayed case ($\tau = 0$), the simulation results displayed in Figure 14.13a are qualitatively the same as those seen in Figure 14.12a, which was computed with the larger laziness parameter $\lambda = 10$. However, the differences in the percentages of votes for the different candidates are larger for the smaller laziness parameter: $y_1(600) \approx 33.44$, $y_2(600) \approx 33.35$, $y_3(600) = 33.22$.

In the variant with delay ($\tau = 30$), compare Figures 14.12c and 14.13b, a reduced value of the laziness decreases the chances of the "worst" candidate 3 to make it to the next round; instead, candidate 3 is now almost always on the last rank. Moreover, candidate 3 never reaches rank 1, which was temporarily the case for the higher laziness value $\lambda = 10$.

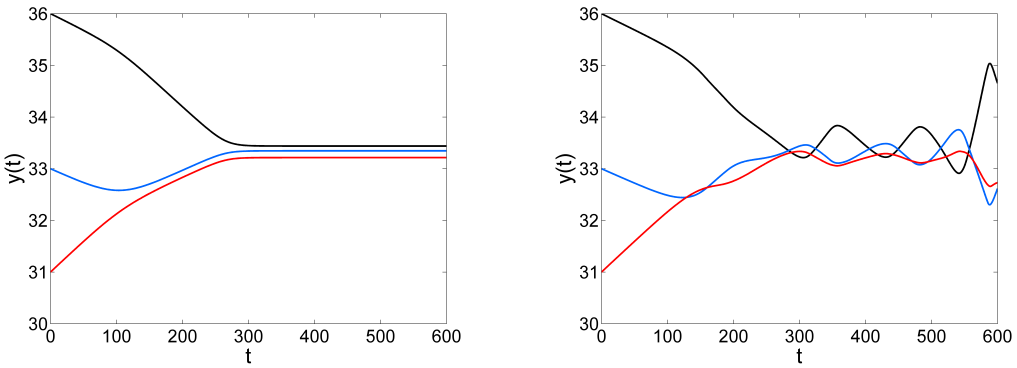


(a) (b)

Figure 14.13.: Solution of the DDE-IVP (14.29), (14.37), for a reduced value of the laziness parameter: $\lambda = 5$. The left figure, (a), shows the IVP solution for $\tau = 0$, and the right figure, (b), shows the IVP solution for $\tau = 30$. The black, blue, and red lines show the simulated percentages of votes for candidate 1, 2, and 3, respectively.

14.3.3. Influence of Panic

The last topic of this simulation study is to investigate the influence of the panic parameter ρ. Therefore, the HDDE-IVP is solved for $\rho = 3$ and for delay values $\tau = 0$ and $\tau = 30$. The laziness parameter is in both cases set to $\lambda = 5$.



(a) (b)

Figure 14.14.: Solution of the DDE-IVP (14.29), (14.37) with panic parameter $\rho = 3$. The left figure, (a), shows the IVP solution for $\tau = 0$, and the right figure, (b), shows the IVP solution for $\tau = 30$. The black, blue, and red lines show the simulated percentages of votes for candidate 1, 2, and 3, respectively.

The results are shown in Figure 14.14. For the case $\tau = 0$, the panic factor does not visibly influence the percentages of votes for the three candidates; in fact, the computed results differ only in the 10th digit. The reason for this behavior is that the panic affects the voting activities for all three candidates in a very similar way. There is, however, a significant increase in the number

of the total votes that are received, i.e. in the state vector component $y_4(t)$ (not displayed in the figures). With $\rho = 0$, the total number of votes at $t = 600$ is approximately 125000, whereas in the case $\rho = 3$ it is approximately 135000.

In the variant with delay ($\tau = 30$), the panic at the end of the voting time leads to a larger amplitude in the last oscillation, compare Figures 14.14b and 14.13b. As a consequence, candidate 1 on rank 1 has almost 2% more votes at the end of the voting time than candidate 3 on rank 2. In contrast to this, the gap is only 0.6% in the case without panic.

15. Sensitivity Analysis

> *There are a number of articles in which the above discussed methods are compared. The conclusion of each article is that the authors' own method is faster and maybe more accurate than the previously published methods.*
>
> Turányi [250], commenting on the literature on numerical comparisons for various methods for sensitivity computation.

This chapter presents numerical results for the computation of sensitivities, i.e. for the computation of the derivatives of initial value problem (IVP) solutions with respect to parameters. Challenging differential equations are considered that feature time delays, discontinuities in the right-hand-side function, and discontinuities in the state.

Numerical Results Presented in This Chapter

The results presented in this chapter are related to four topics.

The first issue is to provide accurate reference values for the sensitivities of IVP solutions in the context of differential equations with time delays. For a specific IVP with chaotic solution, the results of this chapter suggest that previously given reference sensitivities (ZivariPiran [271] and ZivariPiran and Enright [273]) are incorrect.

The second topic of this chapter is to investigate the performance of the newly developed Internal Numerical Differentiation method (see Section 8.2). More precisely, the accuracy and efficiency of the method are compared to two traditional methods for sensitivity computation: finite differences (so-called External Numerical Differentiation) and solution of the combined nominal and variational IVP. In both comparisons, it turns out that the newly developed Internal Numerical Differentiation approach is one or several orders of magnitude faster, while being at least as accurate as the alternative approaches.

This chapter further presents results for different realizations of Internal Numerical Differentiation. More precisely, equivalence of forward and adjoint Internal Numerical Differentiation for delay differential equations (see Sections 8.2 and 8.3) is shown, i.e. the two approaches yield the same result except for numerical round-off errors. For a problem with many parameters, it is demonstrated that the adjoint approach is more efficient for sensitivity computation than the forward approach (see Section 8.2). In addition, an IVP is presented for which it is crucial to use the Internal Numerical Differentiation approach in combination with an error-control strategy for sensitivity computation (see Subsection 8.2.8).

Eventually, the reliability of Colsol-DDE for computing sensitivities is assessed by considering the convergence behavior of the implemented methods in the limit of small relative tolerances.

Organization of This Chapter

The presentation of reference sensitivities and the investigation of the convergence behavior of the methods implemented in Colsol-DDE is the subject of Section 15.1. The comparison of the Internal Numerical Differentiation approaches to alternative methods for sensitivity computation is discussed in Section 15.2. Finally, Section 15.3 presents the results for the forward and the adjoint mode of Internal Numerical Differentiation and demonstrates the benefit of error-controlled sensitivity computation.

General Remarks

For the computation of all numerical results presented in this chapter, a very stringent absolute tolerance is used unless otherwise noted. Further, a restrictive "zero criterion" is employed in

the algorithm that locates zeros of switching functions and propagation switching functions (see Section 6.9).

Throughout this chapter, the use of Colsol-DDE always refers to the use of the three-stage Lobatto IIIA method in Colsol-DDE. Further, sensitivities computation with Internal Numerical Differentiation always refers to the use of the "direct Internal Numerical Differentiation" variant implemented in Colsol-DDE (see Subsection 9.1.6 for details). However, very similar results have also been obtained with the two-stage Radau IIA method and with iterative Internal Numerical Differentiation.

Whenever computation times are reported in this section, they have been obtained on an Intel i7 960 cpu with $3.2GHz$ frequency rate and $8GB$ cache. Furthermore, the same compiler and same compiler flags have been used in order to obtain fair comparisons, e.g. of different approaches for sensitivity computation.

Notation

Parameters in the model functions are denoted by c_i if the sensitivity of the solution with respect to this parameter is of interest. Consequently, the notation $y(t;c)$ is used for the IVP solutions. Furthermore, the i-th row of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{i,*}$, and the j-th column of a matrix $\mathbf{A}$ is denoted by $\mathbf{A}_{*,j}$.

15.1. Accurate Reference Sensitivities and Convergence Analysis

This section presents accurate sensitivities of solutions of challenging IVPs in delay differential equations (DDEs), hybrid discrete-continuous delay differential equations (HDDEs), and impulsive delay differential equations (IDDEs). These accurate sensitivities can be used as reference values. Furthermore, it is demonstrated that sensitivities computed by Colsol-DDE converge to the reference values in the limit of small tolerances.

15.1.1. Physiology: The Mackey-Glass DDE

Problem Definition

Consider the following differential equation:

$$\dot{y}(t;c) = \frac{c_4 y(t-\tau(c))}{c_5 + (y(t-\tau(c)))^{c_2}} - c_4 y(t). \tag{15.1}$$

The differential equation has one constant (but parameter-dependent) delay:

$$\tau(t, y(t;c), c) \equiv \tau(c) = c_3. \tag{15.2}$$

The following initial condition is employed:

$$y(t;c) = c_1 \quad \text{for} \quad t \leqslant 0. \tag{15.3}$$

This means that $t^{ini}(c) \equiv t^{ini} = 0$, $\phi(t,c) \equiv \phi(0,c) = y^{ini}(c) = c_1$. The final time is set to $t^{fin}(c) \equiv t^{fin} = 100$, such that the considered interval is $\mathcal{T} = [0, 100]$.

There are 5 parameters in the IVP: The parameter c_1 represents the constant value of the initial function, c_3 is the constant delay, and the parameters c_2, c_4, and c_5 are used in the right-hand-side function. The values of the parameters are chosen as follows:

$$c_1 = 0.5, \quad c_2 = 9.65, \quad c_3 = 2, \quad c_4 = 2, \quad c_5 = 1. \tag{15.4}$$

Categorization

The IVP (15.1), (15.3) is a DDE-IVP with one constant delay.

References and Background

Mackey and Glass [182] have proposed the DDE (15.1) as a heuristic model for the regulation of hematopoesis and for respiratory behavior. In the former case, the state $y(t)$ represents the density of mature circulating cells and the time delay $\tau(c) = c_3$ stands for the time that passes between the initiation of blood cell production in the bone marrow and the release of mature cells into the blood.

It is known that the DDE-IVP (15.1), (15.3) exhibits, for the parameter values given above, a chaotic behavior (see Glass and Mackey [116]). For this reason, huge values can be expected in the Wronskian matrix $\mathbf{W}(t;c)$. This is indeed the case, as shown in the following.

Numerical Reference Solution and Numerical Reference Sensitivities

A reference solution $\eta^{ref}(100)$ is given by

$$y(100;c) \approx \eta^{ref}(100) = 0.840255042, \tag{15.5}$$

and reference sensitivities $\mathbf{E}^{ref}(100)$ are given by

$$\begin{aligned}\mathbf{W}(100;c) \approx & \mathbf{E}^{ref}(100) \\ = & \begin{pmatrix} 5514.74692 & -294.963701 & -228.342535 & 407.812186 & -1101.27548 \end{pmatrix}. \end{aligned} \tag{15.6}$$

For consistency with the remainder of the thesis, the sensitivities $\mathbf{W}$ and $\mathbf{E}^{ref}$ are written here as matrices (in boldface), even though they are row vectors in this example because y is scalar.

The reference values $\eta^{ref}(100)$ and $\mathbf{E}^{ref}(100)$ have been obtained by using Colsol-DDE with stringent absolute and relative tolerances and relying on the implemented Internal Numerical Differentiation method. Verification of the reference values was done by manually implementing the combined system of nominal and variational DDE-IVP in DDE_SOLVER (Thompson and Shampine [246]) and solving this IVP with stringent tolerances. In this "manual" implementation, automatically generated derivatives of the model functions as provided by Tapenade (Hascoët and Pascual [140]) have been used. The results for the solution and for the sensitivities obtained by Colsol-DDE and DDE_SOLVER were identical in the leading 9 digits given above.

Please note that solving the variational DDE-IVP is sufficient for sensitivity computation in this example because the initial function is continuous, and because the initial function links continuously to the initial value. For DDE-IVPs with discontinuities in the initial function, jumps in the sensitivities would have to be taken into account, cf. Chapters 7 and 8.

Plots of the IVP solution and of the sensitivities with respect to all 5 parameters on the time interval $\mathcal{T} = [0, 200]$ are given in Figure 15.1. The nominal solution $y(t;c)$, displayed in Figure 15.1a, shows an oscillatory but aperiodic behavior. The sensitivities, displayed in the Figures 15.1b-15.1f, all show an oscillatory, aperiodic behavior with increasing amplitude. The oscillations with the largest amplitudes are observed in the sensitivity of the nominal IVP solution with respect to the parameter c_1, i.e. in the sensitivity with respect to the constant value of the initial function.

Convergence Behavior

The convergence behavior of the methods implemented in Colsol-DDE is investigated by varying the relative tolerance over several orders of magnitude: $\sigma_{tol}^{rel} = 10^{-2}$, and $\sigma_{tol}^{rel} = n \cdot 10^{-m}$, with $n \in \{1, \ldots, 9\}$ and $m \in \{3, \ldots, 11\}$. Sensitivities are computed by Internal Numerical Differentiation, and the error control mechanism of Colsol-DDE is only applied to the nominal IVP solution.

For all selected values of the relative tolerance, the IVP is solved and the sensitivities of the IVP solution with respect to the parameters c_i, $1 \leqslant i \leqslant 5$, are computed. This gives results $\eta(100)$ and $\mathbf{E}(100)$. The relative errors of these results are computed by

$$\epsilon_{rel}^{nom} = \frac{|\eta(100) - \eta^{ref}(100)|}{|\eta^{ref}(100)|} \tag{15.7a}$$

$$\epsilon_{rel}^{sens} = \max_{1 \leqslant i \leqslant 5} \left(\frac{|\mathbf{E}_{1,i}(100) - \mathbf{E}_{1,i}^{ref}(100)|}{|\mathbf{E}_i^{ref}(100)|} \right). \tag{15.7b}$$

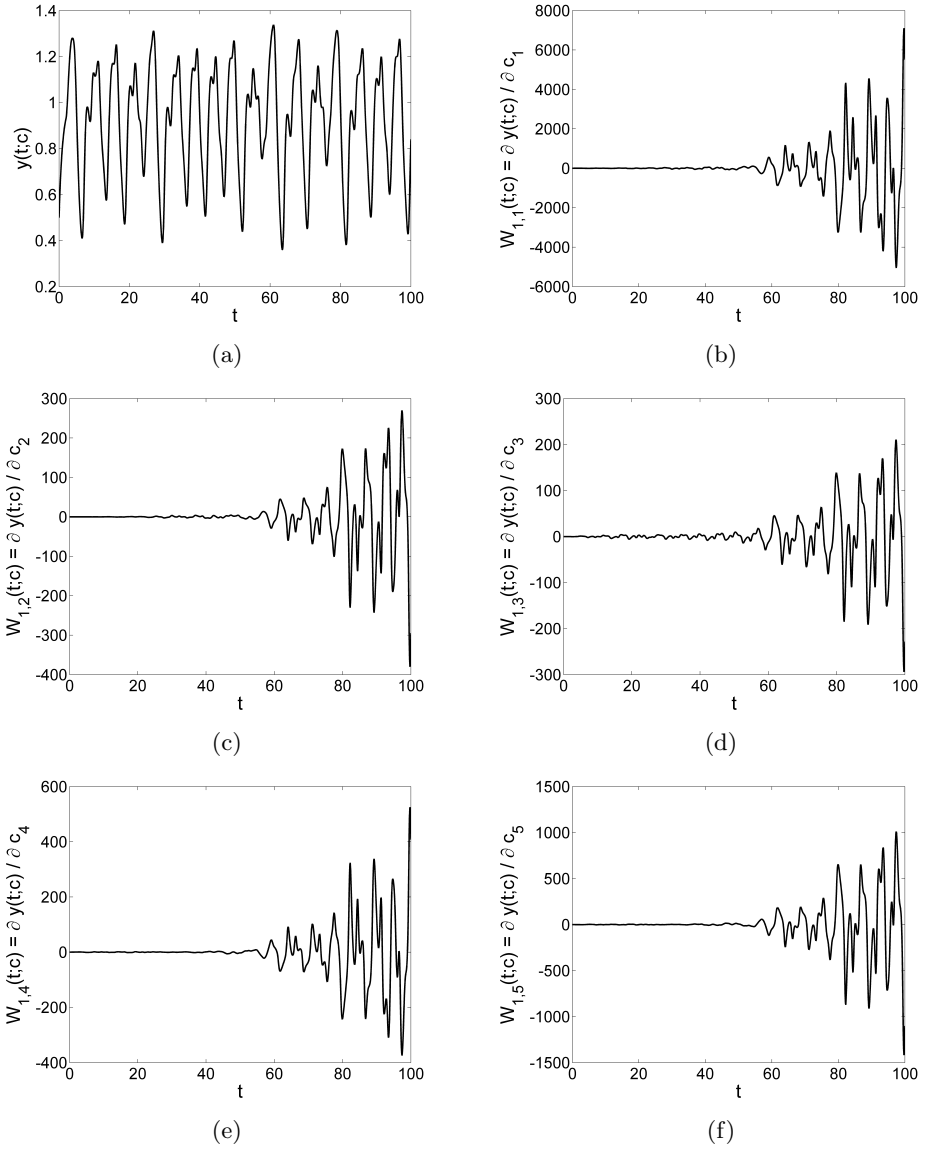


Figure 15.1.: Solution of the DDE-IVP (15.1), (15.3) and sensitivities: (a) nominal IVP solution $y(t;c)$, (b)-(f) sensitivities $\mathbf{W}_{1,i}(t;c) = \partial y(t;c)/\partial c_i$ for $1 \leqslant i \leqslant 5$.

The relative errors as a function of the relative tolerance are displayed in Figure 15.2. A very good error-tolerance proportionality is observed. In particular, error-tolerance proportionality is also observed for the sensitivities, although the error control strategy has only been applied to the nominal solution. Furthermore, the relative error in the sensitivities is only slightly larger than the error in the nominal solution.

Comparison of the Results to Literature Values

ZivariPiran [271] and ZivariPiran and Enright [273] have investigated the Mackey Glass DDE with the same parameter values. Using their software DDEM, they have obtained completely different results for the sensitivities. In particular, the sensitivity $\mathbf{W}_{1,1}(t,c) = \partial y(t;c)/\partial c_1$, approaches 0 for $t \to 100$ in their computations. This clearly contradicts the behavior observed in Figure 15.1b, and, in fact, also sensitivity approximations that are obtained by using finite difference, see ZivariPiran [271] and ZivariPiran and Enright [273]. The reason for these contradictory results is unclear. However, the results of the manual implementation of the nominal and variational DDE-IVP in DDE_SOLVER gives confidence to the implementation in Colsol-DDE and thus to the above-presented reference sensitivities.

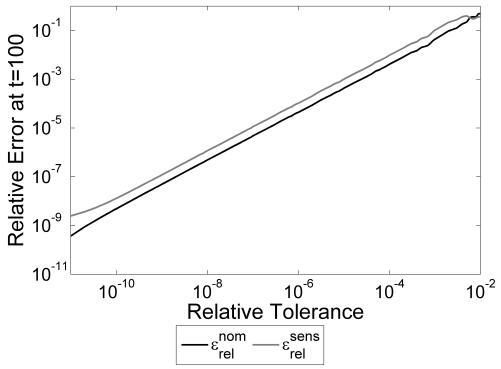


Figure 15.2.: Convergence of the results obtained with Colsol-DDE to the numerical reference values for the solution of the DDE-IVP (15.1), (15.3), and to the corresponding sensitivities. The black line displays the relative error ϵ_{rel}^{nom} in the nominal solution, and the gray line displays the relative error ϵ_{rel}^{sens} in the sensitivities.

15.1.2. Mechanics: An HDDE Model for Stick Balancing

Problem Definition

Consider the following differential equation system:

$$\dot{y}_1(t;c) = y_2(t;c) \tag{15.8a}$$
$$\dot{y}_2(t;c) = \sin(y_1(t;c)) + g(y(t;c), y(t-\tau(c);c), c, \zeta(t)) \tag{15.8b}$$

where

$$g(y(t;c), y(t-\tau(c);c), c, \zeta(t)) = \begin{cases} -\big(c_5 y_1(t-\tau(c);c) + c_6 y_2(t-\tau(c);c)\big) \cdot \cos(y_1(t;c)) & \text{for} \quad \zeta(t) = +1 \\ 0 & \text{for} \quad \zeta(t) = -1 \end{cases}. \tag{15.9}$$

There is one constant delay in the differential equation system (15.8):

$$\tau(t, y(t), c) \equiv \tau(c) = c_3. \tag{15.10}$$

Further, $\zeta(t)$ is the sign of the switching function σ:

$$\zeta(t) = \text{sign}(\sigma(t, y^-(t;c), c, y^-(t-\tau(c);c))). \tag{15.11}$$

The following state-dependent switching function is considered:

$$\sigma(t, y(t;c), c, y(t-\tau(c);c)) \equiv \sigma(y(t-\tau(c);c), c)$$
$$\sigma(y(t-\tau(c);c), c) = y_1(t-\tau(c);c)\Big(y_2(t-\tau(c);c) - c_4 y_1(t-\tau(c);c)\Big). \tag{15.12}$$

The differential equation system (15.8) is associated with the following initial condition:

$$y(t;c) = \begin{pmatrix} c_1 & c_2 \end{pmatrix}^T \quad \text{for} \quad t \leqslant 0. \tag{15.13}$$

This means that $t^{ini}(c) \equiv t^{ini} = 0$, $\phi(t,c) \equiv \phi(0,c) = y^{ini}(c) = (c_1, c_2)^T$. The final time is set to $t^{fin}(c) \equiv t^{fin} = 25$, i.e. the IVP (15.8), (15.13) is considered on the time interval $\mathcal{T} = [0, 25]$.

There are 6 parameters in the IVP: The parameters c_1 and c_2 in the initial condition, the parameter c_3 is the constant delay, parameter c_4 is used in the definition of the switching function, and the parameters c_5 and c_6 are used in the right-hand-side function.

The following values are used for the parameters:

$$c_1 = 0.01, \quad c_2 = 0.01, \quad c_3 = 0.5, \quad c_4 = 0.3, \quad c_5 = 1.5, \quad c_6 = 4. \tag{15.14}$$

$\mathbf{E}^{ref}_{*,1}(25) = \partial y(25;c)/\partial c_1$	$\mathbf{E}^{ref}_{*,2}(25) = \partial y(25;c)/\partial c_2$	$\mathbf{E}^{ref}_{*,3}(25) = \partial y(25;c)/\partial c_3$
-45.03397148	-26.29134854	-53.99468177
-10.79594265	-6.488152219	-12.87492139
$\mathbf{E}^{ref}_{*,4}(25) = \partial y(25;c)/\partial c_4$	$\mathbf{E}^{ref}_{*,5}(25) = \partial y(25;c)/\partial c_5$	$\mathbf{E}^{ref}_{*,6}(25) = \partial y(25;c)/\partial c_6$
1.957157801	-0.5925401568	0.4113500361
0.2206458197	0.02754767065	0.4612943382

Table 15.1.: Numerical reference sensitivities for the solution of the HDDE-IVP (15.8), (15.13).

Categorization

The IVP (15.8), (15.13) is an HDDE-IVP with one constant delay and with one state-dependent switching function.

Background and References

The differential equation (15.8) originates from the description of an inverted pedulum that is mounted on a moving cart. By moving the cart back and forth, the inverted pendulum can be balanced in the upright position. Under the assumption that the mass of the pendulum is small compared to the mass of the cart, the model of Sieber and Krauskopf [236] reads

$$\dot{y}_1(t;c) = y_2(t;c) \tag{15.15a}$$

$$\dot{y}_2(t;c) = \sin(y_1(t;c)) - G(y(t;c), y(t-\tau(c);c), c). \tag{15.15b}$$

Herein, $y_1(t;c)$ denotes the angular displacement of the pendulum compared to the upright position and $y_2(t;c)$ denotes the angular velocity. $G(y(t;c), y(t-\tau(c);c))$ represents the angular acceleration that is due to the movement of the cart. This acceleration depends on the current state and on the state of the system at the past time point $t-\tau(c)$.

The alternative "control law" $g(y(t;c), y(t-\tau(c);c), c, \zeta(t))$ given in equation (15.9), which is used in the differential equation (15.8), has recently been proposed by Simpson, Kuske, and Li [238]. Here, the external control by the movement of the cart is active only if the components of the past state vector $y(t-\tau(c);c)$ are located in certain regions of the phase space.

Numerical Reference Solution and Numerical Reference Sensitivities

By using a sequence of stringent tolerances, the methods implemented in Colsol-DDE show a reasonable convergence behavior for at least the leading 10 digits in the approximations of the components of the nominal solution and of the components of the sensitivity matrix. This motivates to give the following reference solution $\eta^{ref}(25)$:

$$y(25;c) \approx \eta^{ref}(25) = \begin{pmatrix} 0.2410376148 \\ 0.9931167622 \end{pmatrix}. \tag{15.16}$$

Further, it holds that

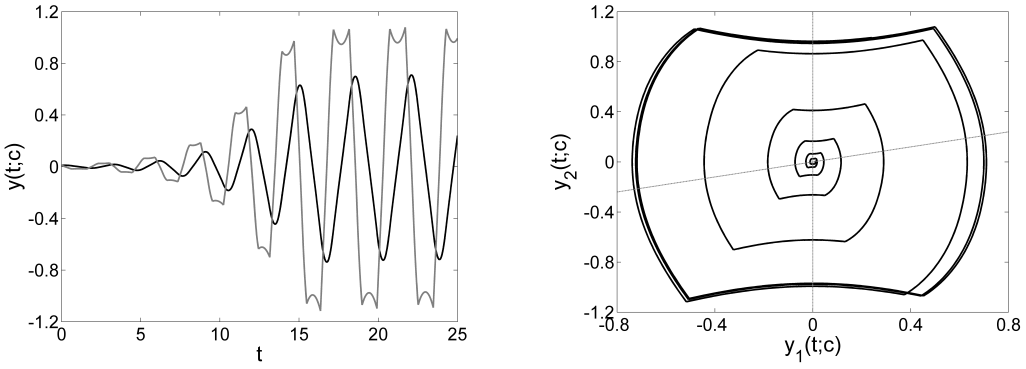
$$\mathbf{W}(25;c) \approx \mathbf{E}^{ref}(25) \tag{15.17}$$

and the reference sensitivities $\mathbf{E}^{ref}(25)$ are given in Table 15.1.

The reference solution $\eta^{ref}(25)$ has been validated by using DDE_SOLVER (see Thompson and Shampine [246]) with stringent tolerances.

Since Colsol-DDE is the only existing solver that features computation of sensitivities of HDDE-IVP solutions, validation of the reference sensitivities $\mathbf{E}^{ref}(25)$ is difficult. Here, a finite difference approach ("External Numerical Differentiation") has been used. This approach may only provide moderate accuracies as discussed in Subsection 8.2.1, however, it allowed verification of the leading 5-6 digits given in Table 15.1. This, together with the reasonable convergence behavior of Colsol-DDE, provides confidence in the correctness of the given reference sensitivities.

The reference solution $y(t)$ is plotted as a function of time in Figure 15.3a and in phase space in Figure 15.3b. Starting at the point $(0.01, 0.01)$ the state spirals outward in a clockwise direction until it eventually shows a periodic behavior.



(a) (b)

Figure 15.3.: Solution of the HDDE-IVP (15.8), (15.13) (a) as a function of time and (b) in phase space. In (a), the solution $y_1(t)$ is plotted as a black line, and $y_2(t)$ is plotted as a gray line. The thin gray lines in (b) represent the points in phase space where the switching function $\sigma(y(t - c_3; c))$ is zero. The control is active whenever the past states are either in the upper right part of the plot or in the lower left part.

The reference sensitivities are displayed in Figure 15.4. Note that discontinuities are present in the sensitivities of $y_2(t; c)$ (vertical lines), but not in the sensitivities of $y_1(t; c)$. The reason for this is that only the second component of the right-hand-side function is discontinuous in the zeros of the switching function.

Convergence Behavior

The convergence behavior of Colsol-DDE is investigated by varying the relative tolerance over several orders of magnitude: $\sigma_{tol}^{rel} = 10^{-2}$, and $\sigma_{tol}^{rel} = n \cdot 10^{-m}$, with $n \in \{1, \ldots, 9\}$ and $m \in \{3, \ldots, 10\}$. For the computation of the sensitivities, Internal Numerical Differentiation is used, and the error control is only applied to the nominal IVP solution.

For all values of the relative tolerance, the IVP is solved and the sensitivities with respect to the parameters c_i, $1 \leqslant i \leqslant 6$, are computed. For each value of σ_{rel}^{tol}, this yields a result $\eta(25)$ for the nominal solution and a result $\mathbf{E}(25)$ for the sensitivities. The relative errors in the nominal solution and in the sensitivities are then computed as follows:

$$\epsilon_{rel}^{nom} = \max_{1 \leqslant i \leqslant 2} \frac{|\eta_i(25) - \eta_i^{ref}(25)|}{|\eta_i^{ref}(25)|} \tag{15.18a}$$

$$\epsilon_{rel}^{sens} = \max_{1 \leqslant i \leqslant 2, 1 \leqslant j \leqslant 6} \left(\frac{|\mathbf{E}_{i,j}(25) - \mathbf{E}_{i,j}^{ref}(25)|}{|\mathbf{E}_{i,j}^{ref}(25)|} \right). \tag{15.18b}$$

The relative errors as functions of the relative tolerance are displayed in Figure 15.5. A good convergence behavior of the employed methods is observed for the nominal solution (black). Although the local errors in the sensitivities are not controlled, a good convergence behavior is also observed for the sensitivities (gray). The relative error in the sensitivities is typically one order of magnitude larger than the error in the nominal solution.

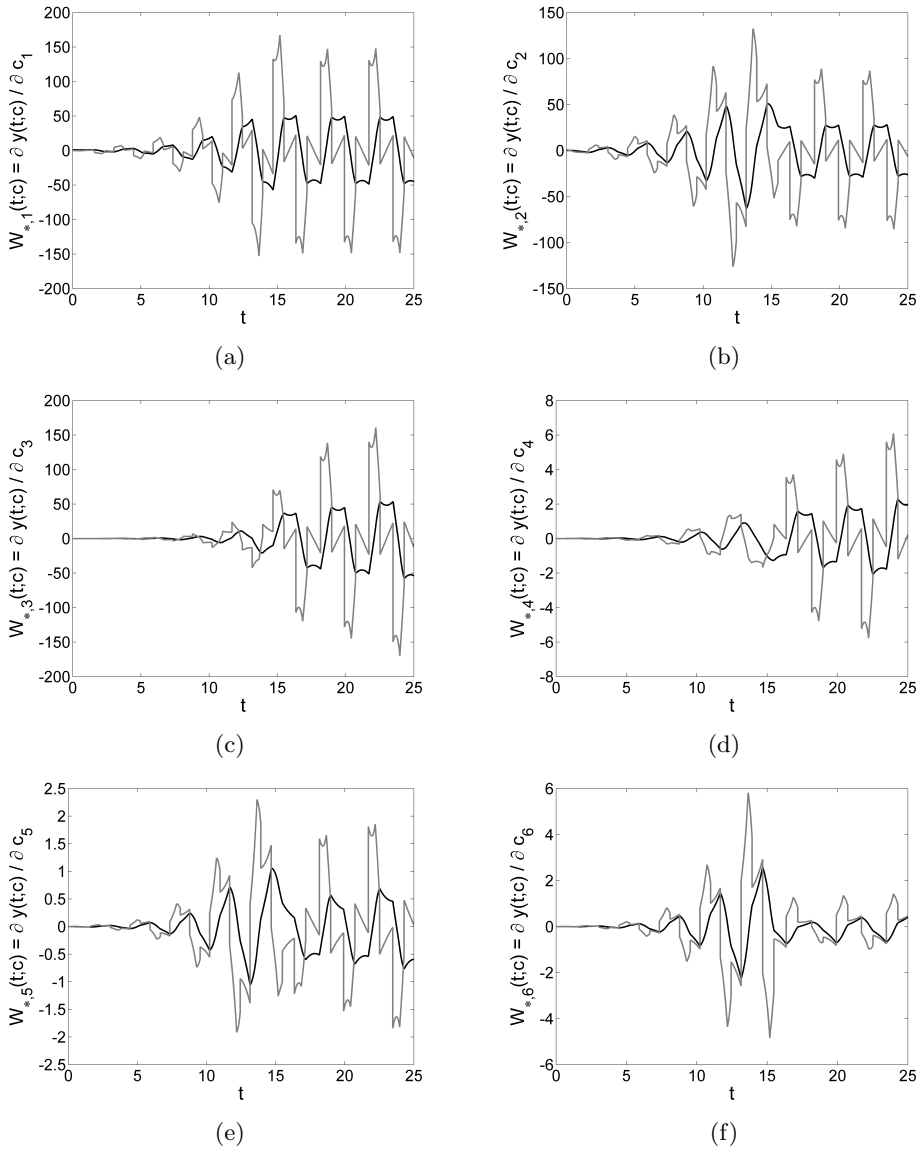


Figure 15.4.: Sensitivities of the solution of HDDE-IVP (15.8), (15.13): In all Figures (a)-(f), the black line shows the sensitivity of $y_1(t;c)$ and the gray line shows the sensitivity of $y_2(t;c)$.

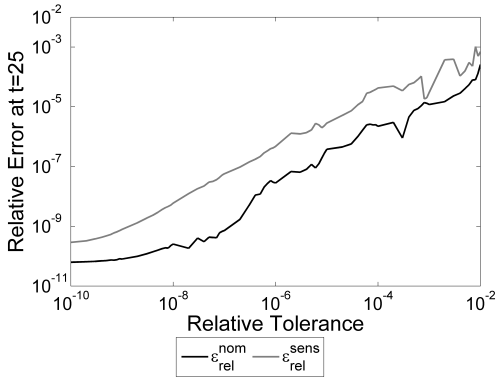


Figure 15.5.: Convergence of the results obtained with Colsol-DDE to the numerical reference values for the solution of the HDDE-IVP (15.8), (15.13) and to the corresponding sensitivities. The black line displays the relative error ϵ_{rel}^{nom} in the nominal solution, and the gray line displays the relative error ϵ_{rel}^{sens} in the sensitivities.

15.1.3. Computer Science: An IDDE Model for a Cellular Neural Network

Problem Definition

In this subsection, the following differential equation system is investigated:

$$\begin{aligned}
\dot{y}_1(t;c) = &- c_{10} \cdot y_1(t;c) + \sin(c_{11} \cdot t) \cdot g(y_1(t;c)) + \cos(c_{12} \cdot t) \cdot g(y_2(t;c)) \\
&+ \sin(c_{13} \cdot t) \cdot g(y_1(t - \tau_1(t,c);c)) \\
&+ \sin(t) \cdot g(y_2(t - \tau_2(t,c);c)) + c_{14} \cdot \sin(t) \quad \text{if} \quad \zeta_i(t) = \pm 1, \quad 1 \leqslant i \leqslant 20 \qquad (15.19a) \\
\dot{y}_2(t;c) = &- c_{15} \cdot y_2(t;c) + \frac{\cos(t) \cdot g(y_1(t;c))}{c_{16}} + \frac{\cos(c_{17} \cdot t) \cdot g(y_2(t;c))}{c_{18}} \\
&+ \cos(t) \cdot g(y_1(t - \tau_1(t,c);c)) + \cos(c_{19} \cdot t) \cdot g(y_2(t - \tau_2(t,c);c)) \\
&+ c_{20} \cdot \cos(t) \qquad \text{if} \quad \zeta_i(t) = \pm 1, \quad 1 \leqslant i \leqslant 20, \qquad (15.19b) \\
y(t;c) = &y^-(t;c) + \omega(y^-(t;c)) \quad \text{if} \quad \zeta_i(t) = 0 \quad \text{for at least one} \quad i \in \{1, 2, \ldots, 20\}. \qquad (15.19c)
\end{aligned}$$

where

$$g(x) = \frac{|x+1| - |x-1|}{2}. \tag{15.20}$$

The differential equation system (15.19) features two time-dependent delay functions:

$$\tau_1(t, y(t;c), c) \equiv \tau_1(t;c) = \frac{c_3 + \cos(t)}{c_4} \tag{15.21a}$$

$$\tau_2(t, y(t;c), c) \equiv \tau_2(t;c) = \frac{c_5 + \sin(t)}{c_6}. \tag{15.21b}$$

Further,

$$\zeta_i(t) = \operatorname{sign}(\sigma_i(t, y^-(t;c), c, \{y^\bullet(t - \tau_i(t;c);c))\}_{i=1}^2) \tag{15.22}$$

are the signs of the following simple time-dependent switching functions:

$$\sigma_i(t, y(t;c), c, \{y(t - \tau_i(t;c);c))\}_{i=1}^2) \equiv \sigma_i(t,c) = t - i \cdot c_7, \tag{15.23}$$

with $1 \leqslant i \leqslant 20$. The impulse function, which is evaluated in the zeros of these switching functions, is given by

$$\omega(t, y(t;c), c, \{y(t - \tau_i(t,c))\}_{i=1}^2, \zeta(t)) \equiv \omega(y(t;c)) \tag{15.24a}$$

$$\omega(y(t;c)) = \begin{pmatrix} c_8 y_1(t;c) \\ c_9 y_2(t;c) \end{pmatrix} \tag{15.24b}$$

The following initial condition is employed:

$$y(t;c) = \begin{pmatrix} c_1 & c_2 \end{pmatrix}^T \quad \text{for} \quad t \leqslant 0. \tag{15.25}$$

This means that $t^{ini}(c) \equiv t^{ini} = 0$, $\phi(t,c) \equiv \phi(0,c) = y^{ini}(c) = (c_1, c_2)^T$. The final time is set to $t^{fin}(c) \equiv t^{fin} = 41$, i.e. the considered time interval is $\mathcal{T} = [0, 41]$.

The values for the parameters that are considered here are the same as those in Corwin, Thompson, and White [71]:

$$\begin{aligned}
c_1 = -0.5, \quad c_2 = 0.5, \quad c_3 = 1, \quad c_4 = 2, \\
c_5 = 1, \quad c_6 = 2, \quad c_7 = 2, \quad c_8 = 0.2, \\
c_9 = 0.3, \quad c_{10} = 6, \quad c_{11} = 2, \quad c_{12} = 3, \\
c_{13} = 3, \quad c_{14} = 4, \quad c_{15} = 7, \quad c_{16} = 3, \\
c_{17} = 2, \quad c_{18} = 2, \quad c_{19} = 2, \quad c_{20} = 2.
\end{aligned} \tag{15.26}$$

Note that c_1 and c_2 define the initial conditions, and that c_3, c_4, c_5, and c_6 are used in the delay functions. For the specific parameter values used here, the delays vanish periodically. The parameter

c_7 occurs in the definition of the switching function. Since $c_7 = 2$ and because $[t^{ini}, t^{fin}] = [0, 41]$, there are in total 20 impulses. The parameters c_8, and c_9 define the impulse, and c_{10}, c_{11}, ..., c_{20} occur in the right hand side of the differential equation.

Categorization

The IVP (15.19), (15.24), (15.25) is an IDDE-IVP with two time-dependent delay functions and 20 simple time-dependent switching functions.

Please note that the function g defined in equation (15.20) is continuous, but not differentiable. In general, this requires to define additional state-dependent switching functions and to formulate the differential equation as a function of the signs of these switching functions. In this way, it can be ensured that the right-hand-side function is, for any fixed values of the switching function signs, sufficiently smooth in order to apply higher order numerical methods. The IVP then becomes an IHDDE-IVP.

However, for the specific initial conditions used here, it is known from Corwin, Thompson, and White [71] that $|y_1(t)| < 1$, and that $|y_2(t)| < 1$. Hence, implementation of additional switching function is dispensable here.

Background and References

The model (15.19) has its roots in the research on so-called *cellular neural networks*. Cellular neural networks have been introduced by Chua and Yang [65, 66] as a class of information processing systems. An important application area of cellular neural networks is image processing, see Egmont-Petersen, de Ridder, and Handels [85]. The specific model for a cellular neural network given in equation (15.19) takes into account time delays and impulsive effects and is due to Yang and Cao [267].

Numerical Reference Solution and Numerical Reference Sensitivities

By using a sequence of stringent tolerances, Colsol-DDE shows a reasonable convergence behavior in the leading $8 - 13$ digits in the approximations of the components of the nominal solution and of the components of the Wronskian matrix. The smallest number of significant digits is obtained for the entries in the first two columns of the Wronskian matrix, which have very small absolute values ($\approx 10^{-50}$). For the nominal solution components as well as for all other components of the sensitivity matrix, a reasonable convergence behavior is obtained for at least the leading 10 digits. Hence, a numerical reference solution $\eta^{ref}(41)$ with 10 digits is given:

$$y(41; c) \approx \eta^{ref}(41) = \begin{pmatrix} 0.06327602399 \\ -0.3804007754 \end{pmatrix}. \tag{15.27}$$

Further,

$$\mathbf{W}(41; c) \approx \mathbf{E}^{ref}(41), \tag{15.28}$$

and the components of $\mathbf{E}^{ref}(41)$ are given in Table 15.2 with 8 digits in the first two columns and with 10 digits in all other columns.

The reference solution $\eta^{ref}(41)$ has been validated by using ddesd (Shampine and Thompson [233]) with stringent tolerances.

Validation of the reference sensitivities is not easily possible because Colsol-DDE is the only existing solver designed for the computation of sensitivities of IDDE-IVP solutions. Therefore, only the leading 4-5 digits have been verified by using the External Numerical Differentiation approach. This is considered as a good indication for the correctness of the reference sensitivities given in Table 15.2.

A plot of the solution on the considered time interval $[0, 41]$ is given in Figure 15.6. Both state vector components show an oscillatory behavior as a function of time (Figure 15.6a). In the phase space plot (Figure 15.6b), the impulses lead to a "ragged" structure.

For brevity, only selected sensitivities are displayed in figures. The sensitivities with respect to c_1, i.e. $\mathbf{W}_{*,1}(t; c)$, are shown in Figure 15.7. Both components of the column $\mathbf{W}_{*,1}(t; c)$ rapidly approach zero. In order to resolve the initial dynamics, the sensitivities are therefore displayed on

$\mathbf{E}^{ref}_{*,1}(41) = \partial y(41;c)/\partial c_1$	$\mathbf{E}^{ref}_{*,2}(41) = \partial y(41;c)/\partial c_2$	$\mathbf{E}^{ref}_{*,3}(41) = \partial y(41;c)/\partial c_3$
$3.2129483 \cdot 10^{-50}$	$-5.7915006 \cdot 10^{-51}$	0.004577405219
$-1.6481695 \cdot 10^{-49}$	$2.9709082 \cdot 10^{-50}$	-0.05010246309
$\mathbf{E}^{ref}_{*,4}(41) = \partial y(41;c)/\partial c_4$	$\mathbf{E}^{ref}_{*,5}(41) = \partial y(41;c)/\partial c_5$	$\mathbf{E}^{ref}_{*,6}(41) = \partial y(41;c)/\partial c_6$
$-1.984800522 \cdot 10^{-4}$	-0.001477483886	$5.517575242 \cdot 10^{-4}$
$4.713910077 \cdot 10^{-4}$	0.01412557471	-0.006985226893
$\mathbf{E}^{ref}_{*,7}(41) = \partial y(41;c)/\partial c_7$	$\mathbf{E}^{ref}_{*,8}(41) = \partial y(41;c)/\partial c_8$	$\mathbf{E}^{ref}_{*,9}(41) = \partial y(41;c)/\partial c_9$
0.06667825669	0.002207972975	$5.147258058 \cdot 10^{-4}$
-0.1688573721	-0.003591162952	-0.003600313094
$\mathbf{E}^{ref}_{*,10}(41) = \partial y(41;c)/\partial c_{10}$	$\mathbf{E}^{ref}_{*,11}(41) = \partial y(41;c)/\partial c_{11}$	$\mathbf{E}^{ref}_{*,12}(41) = \partial y(41;c)/\partial c_{12}$
-0.02830142263	1.004468381	-0.01672278018
0.01033310197	-0.2548906213	-0.1911011073
$\mathbf{E}^{ref}_{*,13}(41) = \partial y(41;c)/\partial c_{13}$	$\mathbf{E}^{ref}_{*,14}(41) = \partial y(41;c)/\partial c_{14}$	$\mathbf{E}^{ref}_{*,15}(41) = \partial y(41;c)/\partial c_{15}$
-0.8837175467	0.003235348615	-0.009470848088
0.1502288185	-0.008400349003	0.06592171736
$\mathbf{E}^{ref}_{*,16}(41) = \partial y(41;c)/\partial c_{16}$	$\mathbf{E}^{ref}_{*,17}(41) = \partial y(41;c)/\partial c_{17}$	$\mathbf{E}^{ref}_{*,18}(41) = \partial y(41;c)/\partial c_{18}$
$-6.756497854 \cdot 10^{-4}$	0.05292386639	-0.002050396788
0.003472102391	-0.07979146305	0.01486079828
$\mathbf{E}^{ref}_{*,19}(41) = \partial y(41;c)/\partial c_{19}$	$\mathbf{E}^{ref}_{*,20}(41) = \partial y(41;c)/\partial c_{20}$	
0.04313277544	0.02516731477	
0.09416982507	-0.1733996897	

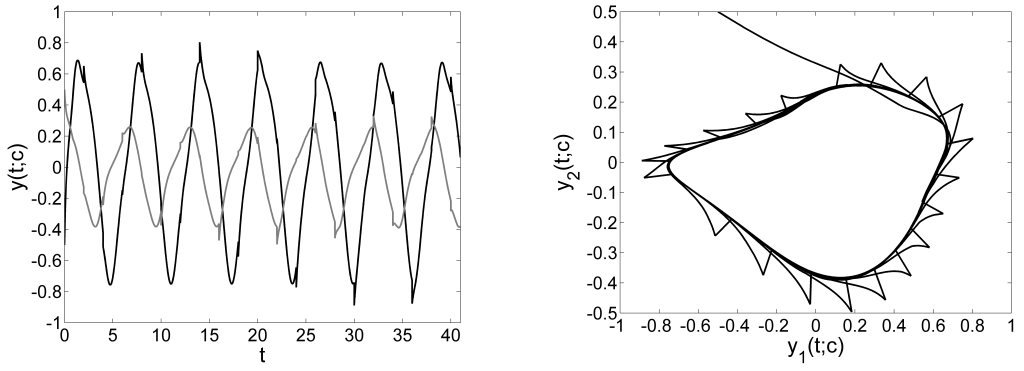
Table 15.2.: Numerical reference sensitivities for the solution of the IDDE-IVP (15.19), (15.24), (15.25).

the interval $[0, 3]$ (Figure 15.7a). Figure 15.7b zooms in on the discontinuity point $t = 2$. It can be seen that a small jump is applied to the sensitivities.

Figure 15.8 displays the sensitivity $\mathbf{W}_{1,3}(t;c) = \partial y_1(t;c)/\partial c_3$. The close up view in Figure 15.8b shows that this sensitivity is discontinuous at the time points of the root discontinuities ($t = 2$, $t = 4$, $t = 6$), but also at other time points, e.g. at $t \approx 6.9$. In fact, the point $t = 6.9$ is a zero of the propagation switching function $\sigma^{\alpha}_{1,s}(t,c) := t - \tau_1(t,c) - s$ for $s = 6$.

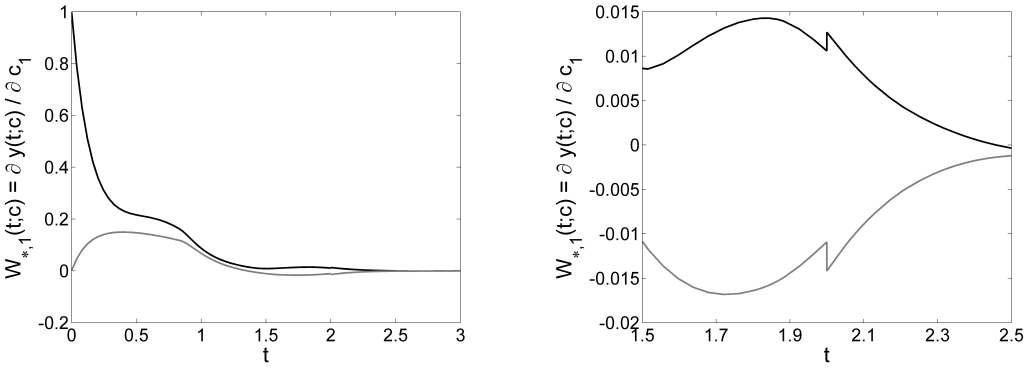
Figure 15.9 displays the sensitivities of the IVP solution with respect to the parameters c_7, c_8, and c_9. All these sensitivities have jumps at the zeros of the switching functions. The sensitivities with respect to c_7 have additional jumps at the time points of the propagated discontinuities. It is observed that the jumps in $\mathbf{W}_{*,7}(t;c)$ are much larger than those in $\mathbf{W}_{*,8}(t;c)$ and $\mathbf{W}_{*,9}(t;c)$.

Eventually, Figure 15.10 displays the sensitivities with respect to two parameters in the right-hand-side function, c_{10} and c_{13}. The sensitivities with respect to c_{10} appears to be "almost" periodic, with only small impulsive perturbations at the zeros of the switching functions. Contrariwise, the sensitivities with respect to the parameter c_{13} show oscillations of increasing amplitude, and the absolute values of these sensitivities are much larger than those of the sensitivities with respect to c_{10}.



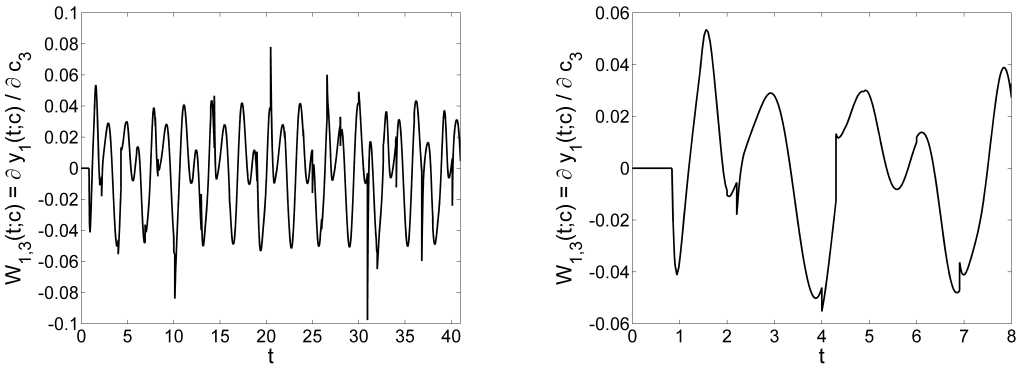
(a) (b)

Figure 15.6.: Solution of the IDDE-IVP (15.19), (15.24), (15.25) (a) as a function of time and (b) in phase space. In (a), the state vector component $y_1(t;c)$ is plotted as a black line, and the state vector component $y_2(t;c)$ is plotted as a gray line.



(a) (b)

Figure 15.7.: Sensitivities of the solution $y(t;c)$ of the IDDE-IVP (15.19), (15.24), (15.25) with respect to the parameter c_1, (a) on the time interval $[0,3]$, and (b) in the vicinity of the discontinuity point $t = 2$. In both cases, $\mathbf{W}_{1,1}(t;c)$ is displayed in black, and $\mathbf{W}_{2,1}(t;c)$ is displayed in gray.



(a) (b)

Figure 15.8.: Sensitivity of the component $y_1(t;c)$ of the solution of the IDDE-IVP (15.19), (15.24), (15.25) with respect to the parameter c_3, (a) in the (entire) time interval $[0,41]$, and (b) in the interval $[0,8]$.

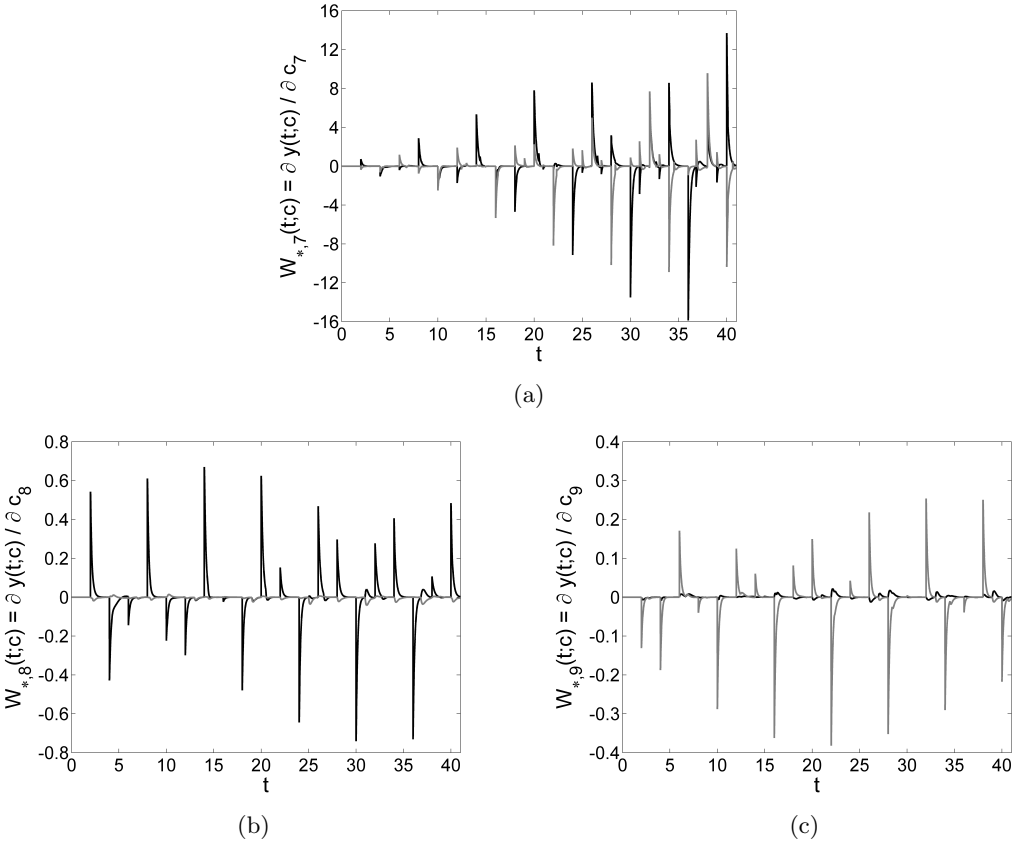


Figure 15.9.: Sensitivities of the solution $y(t;c)$ of the IDDE-IVP (15.19), (15.24), (15.25) with respect to the parameters c_7, c_8, and c_9. In all three figures, the sensitivity of $y_1(t;c)$ is displayed in black, and the sensitivity of $y_2(t;c)$ is displayed in gray.

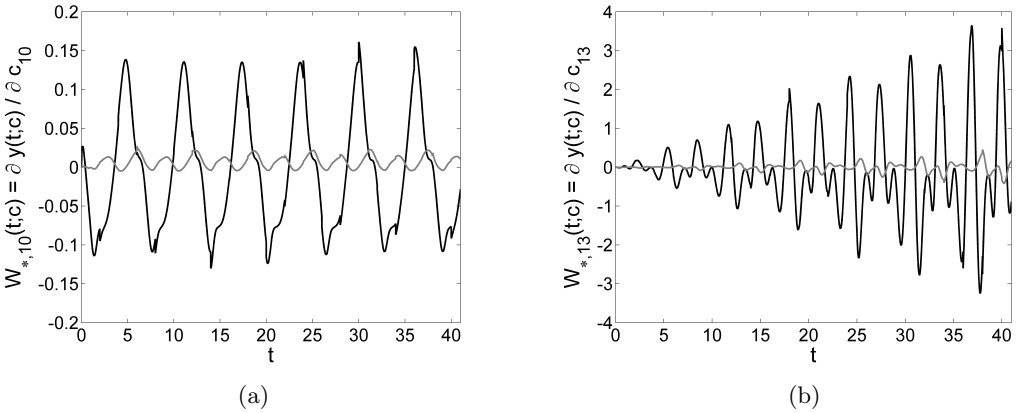


Figure 15.10.: Sensitivities of the solution $y(t;c)$ of the IDDE-IVP (15.19), (15.24), (15.25) with respect to the parameters c_{10} and c_{13}. In both cases, the sensitivity of $y_1(t;c)$ is displayed in black, and the sensitivity of $y_2(t;c)$ is displayed in gray.

15.2. Internal Numerical Differentiation vs. Alternative Approaches for Sensitivity Computation

This section is concerned with a comparison of the newly developed Internal Numerical Differentiation method to alternative approaches for sensitivity computation. As a first step, the employed test problem is defined.

Problem Definition

The following DDE variant of the Hodgin-Huxley model [148] for three coupled neurons is due to Orosz, Moehlis, and Murray [199] and is used a test problem for the investigations in this section:

$$\begin{aligned} y_1(t;c) =& c_{14} \cdot \Big[c_{15} - \varphi_{Na}(y_1(t;c), y_4(t;c), y_7(t;c), c) - \varphi_K(y_1(t;c), y_{10}(t;c), c) \\ & - \varphi_L(y_1(t;c), c) + c_{22}\big((y_2(t-c_{13}) - y_1(t)) + (y_3(t-c_{13}) - y_1(t))\big) \Big] \end{aligned} \tag{15.29a}$$

$$\begin{aligned} y_2(t;c) =& c_{14} \cdot \Big[c_{15} - \varphi_{Na}(y_2(t;c), y_5(t;c), y_8(t;c), c) - \varphi_K(y_2(t;c), y_{11}(t;c), c) \\ & - \varphi_L(y_2(t;c), c) + c_{22}\big((y_1(t-c_{13}) - y_2(t)) + (y_3(t-c_{13}) - y_2(t))\big) \Big] \end{aligned} \tag{15.29b}$$

$$\begin{aligned} y_3(t;c) =& c_{14} \cdot \Big[c_{15} - \varphi_{Na}(y_3(t;c), y_6(t;c), y_9(t;c), c) - \varphi_K(y_3(t;c), y_{12}(t;c), c) \\ & - \varphi_L(y_3(t;c), c) + c_{22}\big((y_1(t-c_{13}) - y_3(t)) + (y_2(t-c_{13}) - y_3(t))\big) \Big] \end{aligned} \tag{15.29c}$$

$$y_i(t;c) = \gamma_m(y_{i-3}(t;c), c) \cdot (1 - y_i(t;c)) - \beta_m(y_{i-3}(t;c), c) \cdot y_i(t;c) \quad \text{for} \quad i = 4,5,6 \tag{15.29d}$$

$$y_i(t;c) = \gamma_h(y_{i-6}(t;c), c) \cdot (1 - y_i(t;c)) - \beta_h(y_{i-6}(t;c), c) \cdot y_i(t;c) \quad \text{for} \quad i = 7,8,9 \tag{15.29e}$$

$$y_i(t;c) = \gamma_n(y_{i-9}(t;c), c) \cdot (1 - y_i(t;c)) - \beta_n(y_{i-9}(t;c), c) \cdot y_i(t;c) \quad \text{for} \quad i = 10,11,12 \tag{15.29f}$$

Herein, the following functions have been used:

$$\varphi_{Na}(z_1, z_2, z_3, c) = c_{16} \cdot z_2^3 \cdot z_3 \cdot (z_1 - c_{19}) \tag{15.30a}$$

$$\varphi_K(z_1, z_2, c) = c_{17} \cdot z_2^4 \cdot (z_1 - c_{20}) \tag{15.30b}$$

$$\varphi_L(z_1, c) = c_{18} \cdot (z_1 - c_{21}) \tag{15.30c}$$

and

$$\gamma_m(z, c) = \frac{c_{24} \cdot (z + c_{23})}{1 - \exp(-(z + c_{23})/c_{25})} \tag{15.31a}$$

$$\gamma_h(z, c) = c_{27} \cdot \exp(-(z + c_{26})/c_{28}) \tag{15.31b}$$

$$\gamma_n(z, c) = \frac{c_{30} \cdot (z + c_{29})}{1 - \exp(-(z + c_{29})/c_{31})} \tag{15.31c}$$

$$\beta_m(z, c) = c_{33} \cdot \exp(-(z + c_{32})/c_{34}) \tag{15.31d}$$

$$\beta_h(z, c) = \frac{c_{36}}{1 + \exp(-(z + c_{35})/c_{37})} \tag{15.31e}$$

$$\beta_n(z, c) = c_{39} \cdot \exp(-(z + c_{38})/c_{40}). \tag{15.31f}$$

The DDE (15.29) is associated with the following initial condition

$$y_i(t;c) = c_i \quad \text{for} \quad t \leqslant 0, \tag{15.32}$$

which means that $t^{ini}(c) \equiv t^{ini} = 0$ and $\phi(t,c) \equiv \phi(0;c) = y^{ini}(c)$. The final time is set to $t^{fin}(c) \equiv t^{fin} = 60$, i.e. the DDE-IVP (15.29), (15.32) is considered on the interval $\mathcal{T} = [0, 60]$.

The DDE-IVP has a 12-dimensional state vector and a 40-dimensional parameter vector. The state vector components $y_1(t;c)$, $y_2(t;c)$, and $y_3(t;c)$ represent the potential at the membrane of the neuron. The first 12 parameters represent the constant values of the initial function (which

are equal to the initial value). Parameter c_{13} is the delay and the parameter c_{22} stands for the coupling strength between the neurons. For the meaning of the other state vector components and parameters it is referred to Hodgin and Huxley [148] and to Orosz, Moehlis, and Murray [199].

The following parameter values are used:

$$\begin{aligned}
c_{01} = -66.0, \quad c_{02} = -57.0, \quad c_{03} = -55.8, \quad c_{04} = 0.043,\\
c_{05} = 0.12, \quad c_{06} = 0.54, \quad c_{07} = 0.32, \quad c_{08} = 0.37,\\
c_{09} = 0.069, \quad c_{10} = 0.49, \quad c_{11} = 0.44, \quad c_{12} = 0.73,\\
c_{13} = 5.511241875, \quad c_{14} = 1.0, \quad c_{15} = 20.0, \quad c_{16} = 120.0,\\
c_{17} = 36.0, \quad c_{18} = 0.3, \quad c_{19} = 50.0, \quad c_{20} = -77.0,\\
c_{21} = -54.4, \quad c_{22} = 0.03, \quad c_{23} = 40.0, \quad c_{24} = 0.1,\\
c_{25} = 10.0, \quad c_{26} = 65.0, \quad c_{27} = 0.07, \quad c_{28} = 20.0,\\
c_{29} = 55.0, \quad c_{30} = 0.01, \quad c_{31} = 10.0, \quad c_{32} = 65.0,\\
c_{33} = 4.0, \quad c_{34} = 18.0, \quad c_{35} = 35.0, \quad c_{36} = 1.0,\\
c_{37} = 10.0, \quad c_{38} = 65.0, \quad c_{39} = 0.125, \quad c_{40} = 80.0
\end{aligned} \tag{15.33}$$

Numerical Reference Solution and Numerical Reference Sensitivities

For the specific parameters given in equation (15.33), the DDE-IVP is solved with stringent tolerances and sensitivities are computed with the Internal Numerical Differentiation approach implemented in Colsol-DDE. The leading 8 digits of the obtained solution are verified by implementing the combined system of nominal DDE-IVP and variational DDE-IVP manually[1] in DDE_SOLVER by Thompson and Shampine [246]. Please note that validation of the sensitivities in this way is only possible because the initial function is continuous, and because the initial function links continuosly to the initial value.

The obtained reference solution $\eta^{ref}(60)$ with 8 valid digits is:

$$y(60;c) \approx \eta^{ref}(60) = \begin{pmatrix} -57.458788 \\ -38.031909 \\ -67.977636 \\ 0.11091611 \\ 0.81196903 \\ 0.034856126 \\ 0.37964947 \\ 0.065512347 \\ 0.30805708 \\ 0.43376271 \\ 0.74446898 \\ 0.50819294 \end{pmatrix}. \tag{15.34}$$

Further, for the sensitivities it holds that

$$\mathbf{W}(60;c) \approx \mathbf{E}(60), \tag{15.35}$$

and the reference sensitivities $\mathbf{E}(60)$ are given in Table 15.3. For the sake of brevity, only the leading 4 digits of the sensitivities are given.

A plot on the interval $[0, 60]$ is given in Figure 15.11. All 12 state vector components show a periodic behavior, and the peaks in the membrane potential of the three neurons ("neuron spikes") are equally spaced.

15.2.1. Comparison to External Numerical Differentiation

The goal in this subsection is to compare the efficiency and accuracy of the newly developed Internal Numerical Differentiation method for DDE-IVPs (Section 8.2) to classical finite difference approximations of the sensitivities ("External Numerical Differentiation", see Subsection 8.2.1).

[1] The "manual" implementation employed automatically generated subroutines for the derivatives of the model functions provided by Tapenade (see Hascoët and Pascual [140]).

$\mathbf{E}^{ref}_{*,1}(60;c)$	$\mathbf{E}^{ref}_{*,2}(60;c)$	$\mathbf{E}^{ref}_{*,3}(60;c)$	$\mathbf{E}^{ref}_{*,4}(60;c)$	$\mathbf{E}^{ref}_{*,5}(60;c)$	$\mathbf{E}^{ref}_{*,6}(60;c)$	$\mathbf{E}^{ref}_{*,7}(60;c)$	$\mathbf{E}^{ref}_{*,8}(60;c)$
-6.495E-02	4.736E-01	2.313E-02	-6.188E-01	3.470E+01	-3.228E-01	7.639E+00	1.841E+01
4.650E-01	-8.589E+00	-3.726E-01	6.025E+00	-6.295E+02	1.976E+00	-8.035E+01	-3.221E+02
-1.163E-02	4.660E-01	1.695E-02	-1.966E-01	3.451E+01	-2.314E-01	2.804E+00	1.806E+01
-6.831E-04	4.267E-03	2.194E-04	-6.349E-03	3.120E-01	-3.065E-03	7.766E-02	1.660E-01
5.273E-03	-9.754E-02	-4.229E-03	6.834E-02	-7.149E+00	2.252E-02	-9.115E-01	-3.659E+00
-4.704E-05	1.902E-03	6.914E-05	-7.991E-04	1.409E-01	-9.440E-04	1.141E-02	7.371E-02
1.083E-04	-8.180E-04	-3.925E-05	1.035E-03	-6.000E-02	5.740E-04	-1.279E-02	-3.189E-02
1.141E-04	-2.006E-03	-8.740E-05	1.448E-03	-1.470E-01	4.901E-04	-1.922E-02	-7.531E-02
-2.065E-04	7.934E-03	2.886E-04	-3.403E-03	5.875E-01	-4.022E-03	4.836E-02	3.077E-01
7.831E-06	1.630E-04	4.864E-06	2.168E-05	1.212E-02	-3.971E-05	-4.252E-05	6.215E-03
1.477E-04	-2.433E-03	-1.071E-04	1.830E-03	-1.782E-01	6.073E-04	-2.416E-02	-9.134E-02
1.791E-04	-7.430E-03	-2.691E-04	3.080E-03	-5.503E-01	3.707E-03	-4.408E-02	-2.880E-01

$\mathbf{E}^{ref}_{*,9}(60;c)$	$\mathbf{E}^{ref}_{*,10}(60;c)$	$\mathbf{E}^{ref}_{*,11}(60;c)$	$\mathbf{E}^{ref}_{*,12}(60;c)$	$\mathbf{E}^{ref}_{*,13}(60;c)$	$\mathbf{E}^{ref}_{*,14}(60;c)$	$\mathbf{E}^{ref}_{*,15}(60;c)$	$\mathbf{E}^{ref}_{*,16}(60;c)$
2.463E+00	-3.395E+01	-1.343E+02	-6.030E+00	-4.913E+00	2.831E+01	3.875E+00	4.323E-01
-1.539E+01	3.111E+02	2.430E+03	3.505E+01	4.023E+01	-3.724E+02	-4.682E+01	-5.824E+00
1.532E+00	-1.019E+01	-1.334E+02	-4.713E+00	-2.773E+00	2.045E+01	2.786E+00	2.911E-01
2.383E-02	-3.494E-01	-1.207E+00	-5.652E-02	-4.334E-02	2.689E-01	3.729E-02	4.005E-03
-1.752E-01	3.528E+00	2.759E+01	3.998E-01	4.588E-01	-4.222E+00	-5.318E-01	-6.627E-02
6.247E-03	-4.140E-02	-5.445E-01	-1.923E-02	-1.129E-02	8.327E-02	1.134E-02	1.187E-03
-4.333E-03	5.685E-02	2.322E-01	1.081E-02	8.627E-03	-5.727E-02	-1.264E-02	-5.211E-04
-3.811E-03	7.501E-02	5.674E-01	8.731E-03	1.012E-02	-8.196E-02	-1.148E-02	-1.558E-03
2.667E-02	-1.769E-01	-2.271E+00	-8.188E-02	-4.764E-02	3.417E-01	3.986E-02	5.420E-03
1.962E-04	1.625E-03	-4.677E-02	-9.093E-04	-1.794E-03	1.420E-02	4.190E-03	1.495E-04
-4.765E-03	9.513E-02	6.879E-01	1.076E-02	1.262E-02	-1.068E-01	-1.432E-02	-8.833E-04
-2.450E-02	1.595E-01	2.127E+00	7.560E-02	4.367E-02	-3.118E-01	-3.880E-02	-4.635E-03

$\mathbf{E}^{ref}_{*,17}(60;c)$	$\mathbf{E}^{ref}_{*,18}(60;c)$	$\mathbf{E}^{ref}_{*,19}(60;c)$	$\mathbf{E}^{ref}_{*,20}(60;c)$	$\mathbf{E}^{ref}_{*,21}(60;c)$	$\mathbf{E}^{ref}_{*,22}(60;c)$	$\mathbf{E}^{ref}_{*,23}(60;c)$	$\mathbf{E}^{ref}_{*,24}(60;c)$
-2.782E+00	-1.839E+01	5.012E-01	1.793E+00	1.163E+00	1.531E+02	1.064E+01	1.706E+03
3.489E+01	2.589E+02	-7.023E+00	-2.179E+01	-1.404E+01	-2.365E+03	-1.399E+02	-2.255E+04
-1.965E+00	-9.275E+00	3.326E-01	1.825E+00	8.358E-01	1.106E+02	7.366E+00	1.184E+03
-2.639E-02	-1.653E-01	4.617E-03	1.759E-02	1.119E-02	1.410E+00	1.064E-01	1.701E+01
3.970E-01	2.938E+00	-7.935E-02	-2.495E-01	-1.595E-01	-2.686E+01	-1.587E+00	-2.556E+02
-8.006E-03	-3.767E-02	1.356E-03	7.604E-03	3.401E-03	4.515E-01	3.244E-02	5.164E+00
7.690E-03	-3.491E-02	-2.003E-04	-1.518E-02	-3.792E-03	-2.779E-01	-1.805E-02	-2.811E+00
9.287E-03	5.580E-02	-1.795E-03	-6.740E-03	-3.443E-03	-5.512E-01	-3.230E-02	-5.186E+00
-2.991E-02	-2.837E-01	6.850E-03	7.876E-03	1.196E-02	1.859E+00	1.283E-01	2.073E+01
-2.657E-03	2.343E-02	1.812E-04	5.563E-03	1.257E-03	3.623E-02	3.181E-03	4.985E-01
8.068E-03	5.287E-02	-2.351E-04	-7.093E-03	-4.297E-03	-6.889E-01	-3.815E-02	-6.006E+00
2.790E-02	2.286E-01	-5.295E-03	-1.390E-02	-1.164E-02	-1.753E+00	-1.173E-01	-1.885E+01

$\mathbf{E}^{ref}_{*,25}(60;c)$	$\mathbf{E}^{ref}_{*,26}(60;c)$	$\mathbf{E}^{ref}_{*,27}(60;c)$	$\mathbf{E}^{ref}_{*,28}(60;c)$	$\mathbf{E}^{ref}_{*,29}(60;c)$	$\mathbf{E}^{ref}_{*,30}(60;c)$	$\mathbf{E}^{ref}_{*,31}(60;c)$	$\mathbf{E}^{ref}_{*,32}(60;c)$
3.365E+01	-1.482E+00	4.235E+02	-2.074E-01	-9.429E+00	-1.717E+04	-2.303E+01	7.424E+00
-4.386E+02	1.930E+01	-5.514E+03	2.732E+00	1.134E+02	2.053E+05	2.763E+02	-9.706E+01
2.318E+01	-9.505E-01	2.716E+02	-1.578E-01	-6.239E+00	-1.143E+04	-1.517E+01	5.092E+00
3.383E-01	-1.370E-02	3.914E+00	-1.995E-03	-8.971E-02	-1.634E+02	-2.194E-01	7.459E-02
-4.980E+00	2.198E-01	-6.281E+01	3.010E-02	1.289E+00	2.334E+03	3.138E+00	-1.097E+00
1.047E-01	-3.875E-03	1.107E+00	-6.437E-04	-2.540E-02	-4.656E+01	-6.172E-02	2.253E-02
-5.903E-02	-1.194E-02	3.411E+00	-4.792E-04	2.262E-02	4.511E+01	5.420E-02	-1.321E-02
-1.017E-01	2.985E-03	-8.528E-01	1.074E-03	2.808E-02	5.377E+01	6.672E-02	-2.309E-02
4.017E-01	-3.024E-02	8.641E+00	-5.113E-03	-1.009E-01	-1.781E+02	-2.482E-01	8.783E-02
1.068E-02	-5.600E-04	1.600E-01	-5.981E-05	2.044E-03	3.897E+00	4.713E-03	2.326E-03
-1.220E-01	2.203E-03	-6.293E-01	1.120E-03	3.659E-02	7.430E+01	8.396E-02	-2.648E-02
-3.688E-01	1.513E-02	-4.323E+00	2.512E-03	9.991E-02	1.830E+02	2.431E-01	-8.107E-02

$\mathbf{E}^{ref}_{*,33}(60;c)$	$\mathbf{E}^{ref}_{*,34}(60;c)$	$\mathbf{E}^{ref}_{*,35}(60;c)$	$\mathbf{E}^{ref}_{*,36}(60;c)$	$\mathbf{E}^{ref}_{*,37}(60;c)$	$\mathbf{E}^{ref}_{*,38}(60;c)$	$\mathbf{E}^{ref}_{*,39}(60;c)$	$\mathbf{E}^{ref}_{*,40}(60;c)$
-3.341E+01	-3.079E+00	-9.145E-01	9.962E+00	-3.219E+00	-3.347E+00	2.142E+03	7.615E-02
4.368E+02	4.160E+01	1.221E+01	-1.343E+02	4.274E+01	3.965E+01	-2.537E+04	-9.329E-01
-2.292E+01	-2.115E+00	-5.457E-01	1.030E+01	-2.113E+00	-2.258E+00	1.445E+03	4.852E-02
-3.357E-01	-3.035E-02	-8.277E-03	1.057E-01	-2.989E-02	-3.206E-02	2.052E+01	7.051E-04
4.937E+00	4.612E-01	1.402E-01	-1.467E+00	4.847E-01	4.505E-01	-2.883E+02	-1.073E-02
-1.014E-01	-8.291E-03	-2.225E-03	4.207E-02	-8.618E-03	-9.180E-03	5.875E+00	1.996E-04
5.946E-02	4.864E-03	-8.724E-03	-1.846E-01	-1.904E-02	9.377E-03	-6.002E+00	-1.894E-04
1.039E-01	1.090E-02	5.893E-04	-1.053E-01	7.548E-03	9.783E-03	-6.261E+00	-3.081E-04
-3.952E-01	-3.707E-02	-1.561E-02	1.704E-02	-4.839E-02	-3.471E-02	2.221E+01	7.215E-04
-1.047E-02	-7.870E-04	-3.520E-04	-2.969E-03	-9.540E-04	5.374E-04	-3.440E-01	-4.132E-05
1.192E-01	9.472E-03	-9.081E-04	-1.809E-01	5.640E-03	1.273E-02	-8.146E+00	-5.864E-04
3.648E-01	3.376E-02	8.691E-03	-1.645E-01	3.368E-02	3.627E-02	-2.322E+01	-7.609E-04

Table 15.3.: Numerical reference sensitivities for the solution of the DDE-IVP (15.29), (15.32). All values are given in "scientific e-notation"; e.g., $-6.495E-02$ means $-6.495 \cdot 10^{-2}$.

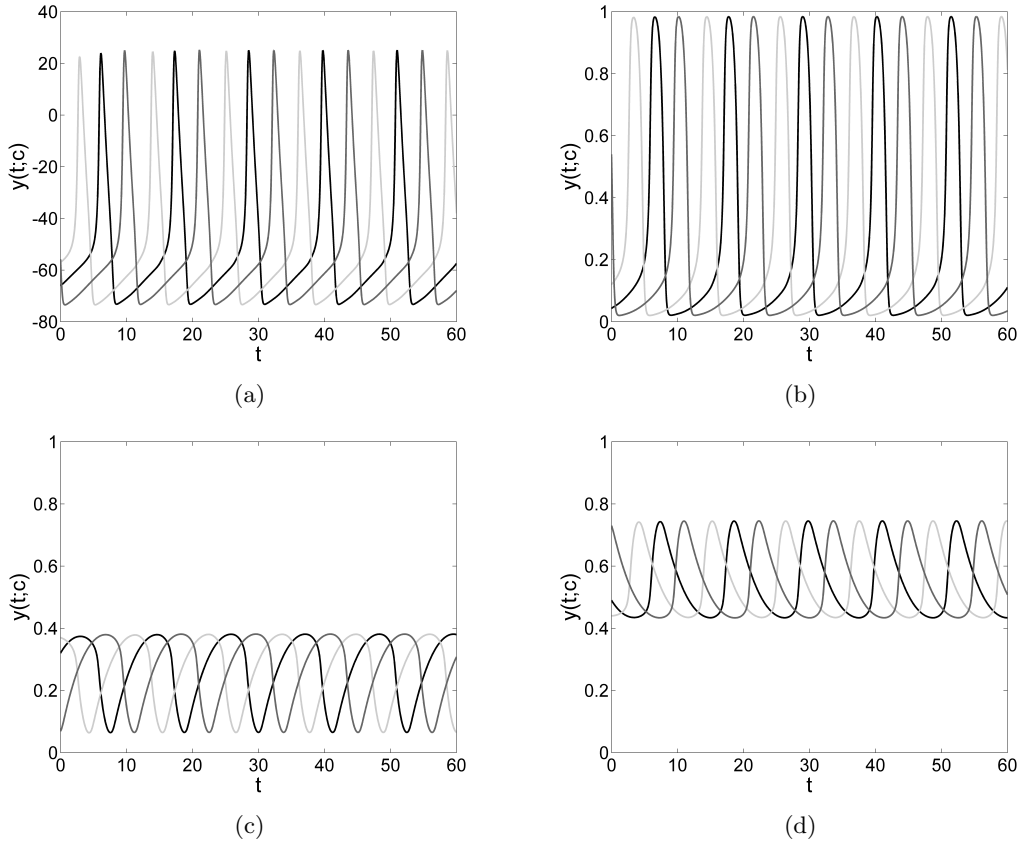


Figure 15.11.: Solution of the DDE-IVP (15.29), (15.32): (a) $y_1(t;c)$ (black), $y_2(t;c)$ (dark gray), $y_3(t;c)$ (light gray); (b) $y_4(t;c)$ (black), $y_5(t;c)$ (dark gray), $y_6(t;c)$ (light gray); (c) $y_7(t;c)$ (black), $y_8(t;c)$ (dark gray), $y_9(t;c)$ (light gray); (d) $y_{10}(t;c)$ (black), $y_{11}(t;c)$ (dark gray), $y_{12}(t;c)$ (light gray).

Consider the situation that the sensitivities should be computed with a moderate relative accuracy of 10^{-2}. On the one hand, the realization of Internal Numerical Differentiation in Colsol-DDE is used. The error control strategy is applied only to the nominal solution. The relative tolerance is set to $\sigma_{tol}^{rel} = 10^{-3}$ and the absolute tolerance is set to $\sigma_{tol}^{abs} = 10^{-8}$.

On the other hand, External Numerical Differentiation is realized by calling Colsol-DDE 41 times, once for computing an approximation of the nominal solution $y(60;c)$, and 40 times for computing approximations of $y(60;c+\epsilon_{fd}^i e_i)$, where e_i is the unit vector in the i-th direction ($i = 1, 2, \ldots, 40$) and ϵ_{fd}^i is the variational parameter of the finite difference. The variational parameter is chosen to be proportional to the absolute value of the parameter: $\epsilon_{fd}^i = \tilde{\epsilon}_{fd}|c_i|$. Three different values for $\tilde{\epsilon}_{fd}$ are used: 10^{-4}, 10^{-6}, and 10^{-8}. For all integrations, the relative and absolute tolerance are set as follows: $\sigma_{tol}^{rel} = 10^{-4}$, $\sigma_{tol}^{abs} = 10^{-8}$.

Accuracy of Sensitivity Computation

The accuracy of an approximation $\mathbf{E}(60)$ of the sensitivity matrix, obtained either with Internal Numerical Differentiation or with External Numerical Differentiation, is assessed by computing the relative error for each of the 480 components of the sensitivity matrix:

$$\boldsymbol{\Delta}\mathbf{E}_{i,j}^{rel} = \frac{|\mathbf{E}_{i,j}(60) - \mathbf{E}_{i,j}^{ref}(60)|}{|\mathbf{E}_{i,j}^{ref}(60)|}. \tag{15.36}$$

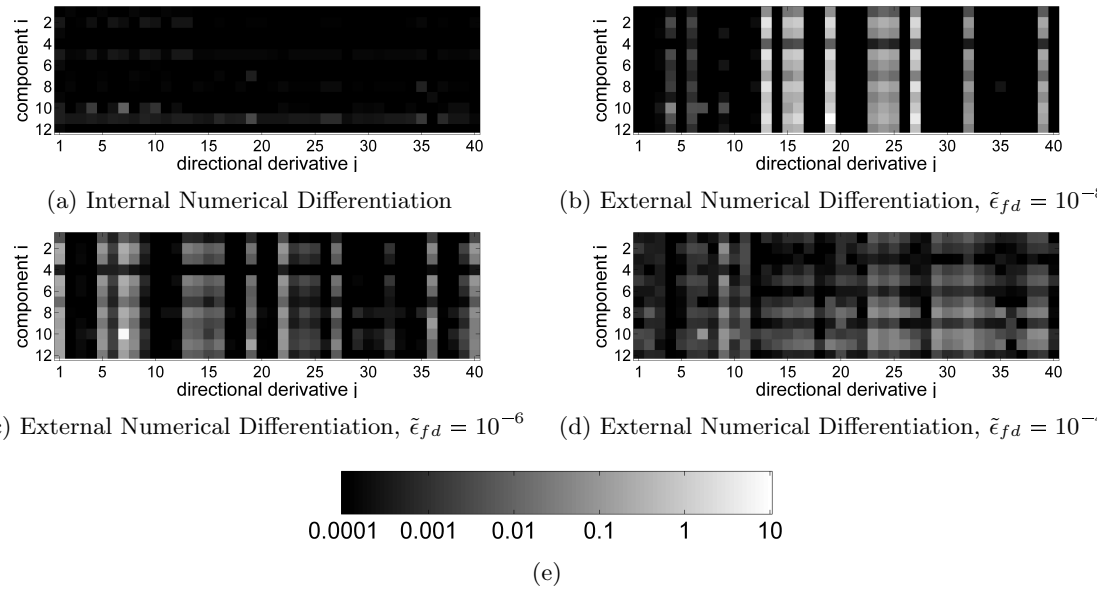
The results are displayed in Figure 15.12. Clearly, the relative errors obtained with Internal Numerical Differentiation are very small compared to the relative errors obtained with External Numerical Differentiation. More precisely, the largest relative error obtained with Internal Numerical Differentiation occurs in $\mathbf{E}_{10,7}(60;c) = \partial y_{10}(60;c)/\partial c_7$ and is given by $\boldsymbol{\Delta}\mathbf{E}_{10,7}^{rel} \approx 9.77 \cdot 10^{-3}$.

Contrariwise, the largest relative errors obtained with External Numerical Differentiation are approximately 16.4, 8.03, and $9.84 \cdot 10^{-2}$ for $\tilde{\epsilon}_{fd} = 10^{-8}$, $\tilde{\epsilon}_{fd} = 10^{-6}$, and $\tilde{\epsilon}_{fd} = 10^{-4}$, respectively. This means that errors of more than 100% are obtained for the two smaller values of $\tilde{\epsilon}_{fd}$, and for $\tilde{\epsilon}_{fd} = 10^{-4}$ the relative error is still one order of magnitude larger than with Internal Numerical Differentiation.

Please note that the dependency of the relative errors on the variational parameter $\tilde{\epsilon}_{fd}$ in the External Numerical Differentiation approach is different for individual components of the sensitivity matrix. For example, the element $\mathbf{\Delta E}^{rel}_{12,13}$ appears as bright gray in Figure 15.12b, indicating a large relative error. For increasing $\tilde{\epsilon}_{fd}$, the error becomes smaller (darker gray in Figures 15.12c and 15.12d). The opposite behavior is observed, e.g., for the element $\mathbf{\Delta E}^{rel}_{10,38}$. Yet other elements, e.g. $\mathbf{\Delta E}^{rel}_{10,7}$, are largest for the intermediate value $\tilde{\epsilon}_{fd} = 10^{-6}$.

The poor performance of External Numerical Differentiation is rooted in the fact that integration methods with adaptive components involve discrete decisions, e.g. because of stepsize rejections. As a consequence, the integration result obtained with a variable-stepsize IVP solver generally depends discontinuously on the values of the parameters (recall the discussion in Section 8.2.1). For the particular choice $\epsilon_{fd} = 10^{-8}$, Colsol-DDE takes one additional integration step (compared to the nominal solution) for solving the DDE-IVPs for variations in the parameters c_{13}, c_{15}, c_{16}, c_{19}, c_{23}, c_{24}, c_{25}, c_{27}, c_{32}, and c_{39}. This causes the large errors in the corresponding columns (see Figure 15.12b).

It is remarked that the poor performance of the External Numerical Differentiation approach is not related to the use of the specific solver Colsol-DDE. Instead, External Numerical Differentiation typically leads to poor sensitivity approximations also for other variable-stepsize IVP solvers. In particular, this effect is observed for dde23 and RADAR5 in Lenz, Schlöder, and Bock [173].



(a) Internal Numerical Differentiation

(b) External Numerical Differentiation, $\tilde{\epsilon}_{fd} = 10^{-8}$

(c) External Numerical Differentiation, $\tilde{\epsilon}_{fd} = 10^{-6}$

(d) External Numerical Differentiation, $\tilde{\epsilon}_{fd} = 10^{-4}$

(e)

Figure 15.12.: Accuracy of sensitivity computation with Internal Numerical Differentiation and with External Numerical Differentiation: Relative errors in the entries $\mathbf{\Delta E}^{rel}_{i,j}$ of the sensitivity matrix, (a) for sensitivity computation with Internal Numerical Differentiation, and (b)-(d) for sensitivity computation with External Numerical Differentiation for various values of the finite difference variational parameter $\tilde{\epsilon}_{fd}$. The relative errors are displayed as grayscale values, where darker gray indicates smaller relative errors. White corresponds to relative error of 10 (1000%) or more, black corresponds to a relative error of 10^{-4} (0.01%) or less, see the color bar in (e).

Efficiency of Sensitivity Computation

The combined computation of the nominal IVP solution and of the sensitivities with the Internal Numerical Differentiation approach realized in Colsol-DDE takes 3.36s. Contrariwise, numerically solving 41 IVPs for the External Numerical Differentiation approach needs 15.1s. This means

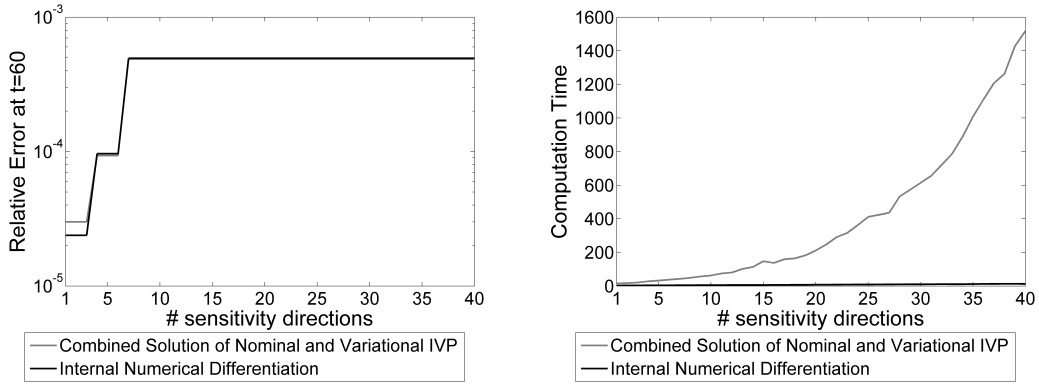
that Internal Numerical Differentiation reduces the computation time by roughly 80%, while also providing the more accurate sensitivity approximation. This clearly demonstrates the advantages of the developed Internal Numerical Differentiation method over External Numerical Differentiation.

15.2.2. Comparison to Manual Implementation of Variational IVP

In this subsection, the Internal Numerical Differentiation approach realized in Colsol-DDE is compared to another classical approach for sensitivity computation, namely to manual implementation of the combined nominal and variational DDE-IVP. Please recall, from the discussion of the validation of the reference sensitivities, that this approach is applicable here because the initial function is continuous and because the initial function links continuously to the initial value, see equation (15.32). For DDE-IVPs with a discontinuity at the initial time, with discontinuous initial functions, and also for HDDE-IVPs and IHDDE-IVPs, it is generally necessary to take into account possible jumps in the sensitivities, see Chapter 7.

The two approaches are implemented as follows. First, the combined system of nominal and variational DDE-IVP is defined as an augmented IVP and solved by Colsol-DDE. This implementation makes use of model function derivatives generated by Tapenade (Hascoët and Pascual [140]). The time derivative $\dot{y}(t - c_{13})$ that appears in the variational DDE is approximated by evaluating the right-hand-side function of the nominal DDE-IVP at the past time point, which can be achieved in Colsol-DDE by defining a second constant delay with value $2c_{13}$. Second, Internal Numerical Differentiation as implemented in Colsol-DDE is used. The error control strategy is thereby applied to both the nominal IVP solution and to the computation of sensitivities, in order to have a better comparability to the manual implementation of the variational DDE-IVP. All settings of Colsol-DDE are the same for the implementation of the two approaches; in particular, the relative and absolute tolerance are set as follows: $\sigma_{tol}^{rel} = 10^{-3}$, $\sigma_{tol}^{abs} = 10^{-8}$.

The accuracy and efficiency of the two approaches are assessed by regarding the relative errors in the computed sensitivities and the computation time as a function of the number of *forward sensitivity directions* (i.e. the number of columns of $\mathbf{W}(t; c)$ that are computed). Figure 15.13a demonstrates that the accuracy of the computed sensitivities is nearly identical for the two approaches (the two lines overlap almost completely). The computation times are, however, drastically different. With Internal Numerical Differentiation, the computation time increases only very mildly from $2.9s$ to $13s$ when the number of sensitivity directions is increased from 1 to 40. Hence, the black line in Figure 15.13b coincides almost exactly with the horizontal axis. Contrariwise, the computation time for the manual implementation of the combined system of nominal and variational DDE-IVP increases from $15s$ for 1 sensitivity direction to $1500s$ for 40 sensitivity directions.



(a) (b)

Figure 15.13.: Comparison of Internal Numerical Differentiation (black) to manual implementation of the combined system of nominal and variational DDE-IVP (gray). Figure (a) displays the maximum relative error in any of the $12 \cdot n_{fd}$ components of the sensitivity matrix (n_{fd} := number of forward sensitivity directions). Figure (b) displays the computation time.

The observed effects are explained as follows. When implementing the combined nominal and variational DDE-IVP manually, the dimension of the state vector is $n_y = 12 \cdot (n_{fd} + 1)$, where n_{fd}

represents the number of forward sensitivity directions. The Lobatto IIIA collocation method in Colsol-DDE requires, in every integration step, solution of a $2n_y$–dimensional nonlinear equation system; in particular, for $n_{fd} = 60$, the size of the system is 984. Further, the Newton-type method used for solving the equation system occasionally requires a recomputation and decomposition of the Jacobian matrix. The costs of these matrix computations and decompositions are the major cause of the increase in computation time seen in Figure 15.13b.

Internal Numerical Differentiation as realized in Colsol-DDE exploits the fact that the computation of the sensitivities can be decoupled from the nominal solution, such that the computational effort shrinks to solving $(n_{fd} + 1)$ equation systems of dimension 12, see Subsection 9.1.5. Direct Internal Numerical Differentiation further exploits that the equation systems for the sensitivities are linear in the unknowns. Hence, it is sufficient to compute and decompose a 12×12 matrix, and the decomposition can be used for solving all n_{fd} linear equation systems that arise in the sensitivity computation.

These results show that structure exploitation is crucial for an efficient computation of sensitivities, in particular for implicit integration methods such as those realized in Colsol-DDE.

15.3. Comparison of Different Realizations of Internal Numerical Differentiation

This section presents and discusses numerical results obtained with different realizations of Internal Numerical Differentiation. First, forward and adjoint mode of Internal Numerical Differentiation are compared. Second, it is demonstrated on a practical example that accurate sensitivity computation sometimes requires to couple the Internal Numerical Differentiation approach with an error control strategy for the sensitivities.

15.3.1. Forward and Adjoint Mode of Internal Numerical Differentiation

Equivalence of Forward and Adjoint Internal Numerical Differentiation

On a given mesh, forward and adjoint Internal Numerical Differentiation are equivalent, i.e. theoretically they yield exactly the same result (see Subsection 8.3.5). This is validated for the HDDE-IVP (15.8), (15.13), which models the motion of an inverted pendulum mounted on a moving cart.

Please recall the investigation of the convergence behavior of Colsol-DDE for this example. Figure 15.5 displays, as a function of the relative tolerance, the relative errors ϵ_{rel}^{nom} and ϵ_{rel}^{sens} in the approximations $\eta(25)$ and $\mathbf{E}(25)$ of the nominal solution and of the sensitivities, respectively. Thereby, the approximation $\mathbf{E}(25)$ of the sensitivity matrix has been obtained by forward Internal Numerical Differentiation.

Alternatively, the sensitivity matrix $\mathbf{W}(25; c)$ can be approximated by adjoint Internal Numerical Differentiation, which yields a result that is denoted by $\tilde{\mathbf{E}}(25)$. The relative difference of the results obtained with forward and adjoint Internal Numerical Differentiation can be computed by

$$\tilde{\epsilon}_{rel}^{sens} = \max_{1 \leqslant i \leqslant 2, 1 \leqslant j \leqslant 6} \left(\frac{|\tilde{\mathbf{E}}_{i,j}(25) - \mathbf{E}_{i,j}(25)|}{|\mathbf{E}_{i,j}^{ref}(25)|} \right), \tag{15.37}$$

where $\mathbf{E}^{ref}(25)$ is the reference result for the sensitivity matrix given in Table 15.1.

Figure 15.14 shows $\tilde{\epsilon}_{rel}^{sens}$ as a function of the relative tolerance. For all values of the relative tolerance, an excellent agreement of forward and adjoint Internal Numerical Differentiation is obtained. The small deviations, which never exceed 10^{-12}, are expected due to the use of floating point arithmetic, see Subsection 8.3.5.

Similar good agreements have been found for a large number of test problems, including some with impulses and state-dependent delays. Since forward and adjoint sensitivity computation are done in almost completely disjoint parts of the Colsol-DDE source code, the good agreement in the results of the two approaches is a very strong evidence for the correctness of the implementation.

Comparison of Efficiency

It is known from theory that adjoint approaches for sensitivity computation are computationally more efficient than forward approaches when there are many parameters (cf. Subsection 8.3.1). In

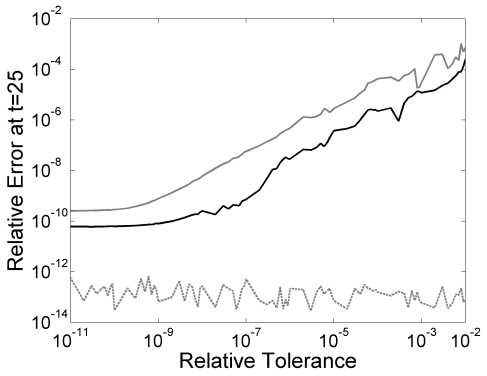


Figure 15.14.: Equivalence of forward and adjoint Internal Numerical Differentiation. The solid black line displays the relative error ϵ_{rel}^{nom} in the nominal solution and the solid gray line displays the relative error ϵ_{rel}^{sens} in the sensitivities, cf. Figure 15.5. In addition, the dashed gray line displays the relative difference $\tilde{\epsilon}_{rel}^{sens}$ of the results obtained with forward and adjoint Internal Numerical Differentiation.

the following, this is demonstrated for the DDE-IVP (15.29), (15.32) which has a 12-dimensional state vector and a 40-dimensional parameter vector. For the sensitivity computations with both forward and adjoint Internal Numerical Differentiation, the DDE-IVP is solved with the following tolerance values: $\sigma_{tol}^{rel} = 10^{-3}$ and $\sigma_{tol}^{abs} = 10^{-8}$. The error control strategy is applied only to the nominal solution.

Table 15.4 shows the computation times needed with forward and adjoint Internal Numerical Differentiation. The full sensitivity matrix can be obtained by computing the directional sensitivities for $n_y = 12$ adjoint directions or for $n_c = 40$ forward directions. As expected for a problem with $n_c > n_y$, adjoint Internal Numerical Differentiation is more efficient and reduces the computation time by about 12%.

In general, it depends on the context whether forward or adjoint Internal Numerical Differentiation is more efficient. For example, if only the first row of the sensitivity matrix is of interest, i.e. $\mathbf{W}_{1,*}(t^{fin}; c)$, then one has to choose between 1 adjoint sensitivity direction and 40 forward sensitivity directions. Here, the adjoint mode is about 29% faster. Contrariwise, if only the first column of the sensitivity matrix is of interest, $\mathbf{W}_{*,1}(t^{fin}, c)$, then one has to choose between 1 forward sensitivity direction and 12 adjoint sensitivity directions. Here, the forward mode is about 51% faster.

Mode	Number of sensitivity directions	computation time [s]
forward	1	1.44
forward	40	3.36
adjoint	1	2.40
adjoint	12	2.95

Table 15.4.: Comparison of forward and adjoint Internal Numerical Differentiation: computation times

15.3.2. Sensitivity Computation with and without Error Control

In Section 15.1 the convergence of the results of Colsol-DDE was investigated for the DDE-IVP (15.1), (15.3) and for the HDDE-IVP (15.8), (15.13). The computed sensitivities converged to the reference result although the error control strategy had been applied only to the nominal solution. Further, the relative error in the sensitivities was only slightly larger than in the nominal solution. These observations are typical for sensitivity computation by Internal Numerical Differentiation. However, in some situations the relative error in the sensitivities can also be much larger than the relative error in the nominal solution. In the following, an example for such a situation is given. Further, it is demonstrated that a remedy is given by making the sensitivity computation subject

to the error control strategy.

Problem Definition

The so-called *repressilator* is a gene regulatory network of three genes that inhibit each other, see Elowitz and Leibler [89]. Orosz, Moehlis, and Murray [199] have proposed an extension of the model for the repressilator, which introduces an additional control gene with time-delayed transcription:

$$\dot{y}_1(t;c) = -\,y_1(t;c) + \alpha \cdot g(y_6(t;c)) \tag{15.38a}$$
$$\dot{y}_2(t;c) = -\,y_2(t;c) + \alpha \cdot g(y_7(t;c)) \tag{15.38b}$$
$$\dot{y}_3(t;c) = -\,y_3(t;c) + \alpha \cdot g((1-\eta)y_5(t;c) + \eta y_8(t;c)) \tag{15.38c}$$
$$\dot{y}_4(t;c) = -\,y_4(t;c) + \alpha^* \cdot g(y_6(t-\tau)) \tag{15.38d}$$
$$\dot{y}_5(t;c) = -\,c \cdot y_5(t;c) + c \cdot y_1(t;c) \tag{15.38e}$$
$$\dot{y}_6(t;c) = -\,c \cdot y_6(t;c) + c \cdot y_2(t;c) \tag{15.38f}$$
$$\dot{y}_7(t;c) = -\,c \cdot y_7(t;c) + c \cdot y_3(t;c) \tag{15.38g}$$
$$\dot{y}_8(t;c) = -\,\beta^* \cdot y_8(t;c) + \beta^* \cdot y_4(t;c) \tag{15.38h}$$

Herein the function $g(x)$ is given by

$$g(x) = \frac{1}{1+x^n} + f_0. \tag{15.39}$$

The state vector components y_1, y_2, y_3, and y_4 represent the concentrations of mRNA containing information of the four genes. Further, the state vector components y_5, y_6, y_7, and y_8 represent the concentrations of the proteins translated from the information encoded in the mRNA.

The following initial condition is used:

$$y(t;c) = \begin{pmatrix} 0.2582 & 9.2097 & 7.0079 & 1.9009 & 8.6734 & 4.1847 & 2.3194 & 1.5617 \end{pmatrix}^T \text{ for } t \leqslant 0. \tag{15.40}$$

This means that $t^{ini}(c) \equiv t^{ini} = 0$ and that $\phi(t,c) \equiv \phi(0) = y^{ini}$. The final time is set to $t^{fin}(c) \equiv t^{fin} = 200$, such that the considered interval is $\mathcal{T} = [0, 200]$.

Kuhn [168] has investigated the DDE-IVP (15.38), (15.40) and, by applying optimization techniques, has found that the solution rapidly approaches a steady state for the following values:

$$\begin{aligned} \alpha = 215.52, \quad \alpha^* = 215.58, \quad c = 0.2069, \quad \beta^* = 0.109633 \\ \eta = 0.609648, \quad f_0 = 0.001, \quad n = 2, \quad \tau = 8.6409. \end{aligned} \tag{15.41}$$

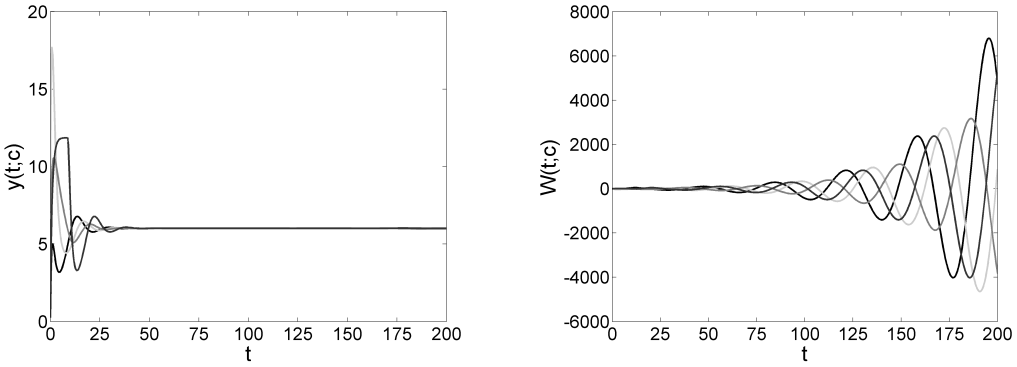
Numerical Reference Solution and Numerical Reference Sensitivities

Reference values for the solution and the sensitivity of the solution with respect to the scalar parameter c, both with 8 valid digits, are given:

$$y(200;c) \approx \eta^{ref}(200) = \begin{pmatrix} 6.0306857 \\ 5.9914350 \\ 6.0284712 \\ 5.9972995 \\ 6.0131210 \\ 6.0026783 \\ 6.0267559 \\ 6.0031892 \end{pmatrix}, \quad \mathbf{W}(200;c) \approx \mathbf{E}^{ref}(200) = \begin{pmatrix} 4718.6092 \\ 880.91300 \\ -3834.3053 \\ 5247.4681 \\ 5337.4239 \\ -2008.5146 \\ -1033.8767 \\ 437.11806 \end{pmatrix}. \tag{15.42}$$

For consistency with the remainder of the thesis, the sensitivities $\mathbf{W}$ and $\mathbf{E}^{ref}$ are written here as matrices (in boldface), even though they are column vectors in this example because c is scalar.

All state vector components rapidly approach a steady state as discussed in Kuhn [168]. Contrariwise, the sensitivities show oscillations with increasing amplitude. For a plot of the first four components of $y(t;c)$ and $\mathbf{W}(t;c)$, see Figure 15.15.



(a) (b)

Figure 15.15.: Solution of the DDE-IVP (15.38), (15.40), and sensitivities: (a) nominal IVP solution, $y_1(t;c)$ in black, $y_2(t;c)$ in dark gray, $y_3(t;c)$ in medium gray, and $y_4(t;c)$ in light gray; (b) sensitivities, $\mathbf{W}_{1,1}(t;c)$ in black, $\mathbf{W}_{1,2}(t;c)$ in dark gray, $\mathbf{W}_{1,3}(t;c)$ in medium gray, and $\mathbf{W}_{1,4}$ in light gray.

Comparison

The DDE-IVP (15.38), (15.40) is numerically solved with Colsol-DDE with relative tolerance $\sigma_{tol}^{rel} = 10^{-3}$ and absolute tolerance $\sigma_{tol}^{rel} = 10^{-8}$. Sensitivities are computed by Internal Numerical Differentiation in two ways. On the one hand, the error control strategy is applied only to the nominal solution, which gives approximations $\eta^1(t)$ and $\mathbf{E}^1(t;c)$. On the other hand, the error control strategy is applied to both the nominal solution and the sensitivities, which gives approximations $\eta^2(t)$ and $\mathbf{E}^2(t;c)$.

The obtained solutions and sensitivities as well as the relative errors of all components, defined by

$$(\epsilon_{rel}^{nom})_i = \frac{|\eta_i^1(200) - \eta_i^{ref}(200)|}{|\eta_i^{ref}(200)|} \tag{15.43a}$$

$$(\epsilon_{rel}^{sens})_i = \frac{|\mathbf{E}_i^1(200) - \mathbf{E}_i^{ref}(200)|}{|\mathbf{E}_i^{ref}(200)|}, \tag{15.43b}$$

are given in Table 15.5. The relative errors are given in percent, i.e. multiplied by a factor 100. Without error control on the sensitivities, the nominal solution is approximated with relative errors of less than 10^{-3}. This corresponds well to the chosen relative tolerance. However, the relative errors in the computed sensitivities are more than four orders of magnitude larger. For three components (highlighted in boldface), the error is larger than 100%, and not even the sign of the result is correct. Contrariwise, with error control on the sensitivities, the relative errors in the sensitivities is less than 0.1% in all components.

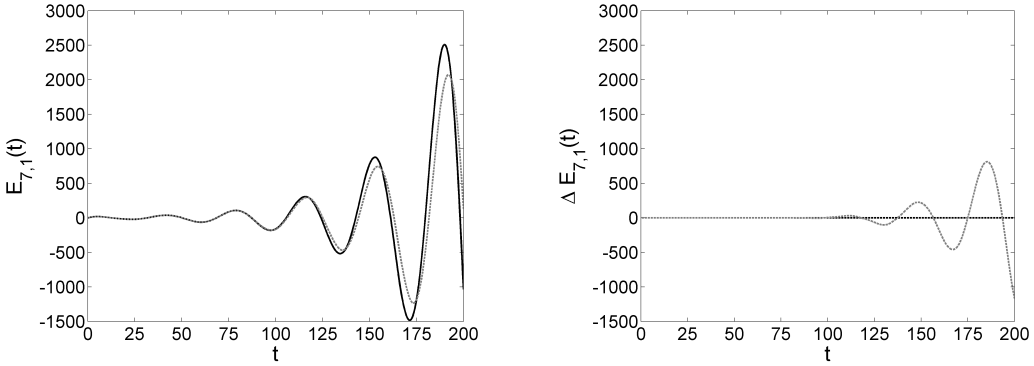
For the sensitivity component $\mathbf{W}_{7,1} = \partial y_7(t;c)/\partial c$, Figure 15.16a displays the reference solution $\mathbf{E}_{7,1}^{ref}(t)$ and the approximations $\mathbf{E}_{7,1}^1(t)$ and $\mathbf{E}_{7,1}^2(t)$. Without error control, the computed sensitivity $\mathbf{E}_{7,1}^1(t)$ (dashed gray line) deviates signifcantly from the reference sensitivity $\mathbf{E}_{7,1}^{ref}(t)$ (solid black line). With error control, the computed sensitivity $\mathbf{E}_{7,1}^2(t)$ (dashed black line) is completely overlayed by the solid black line representing the reference sensitivity. The errors in both $\mathbf{E}_{7,1}^1(t)$ and $\mathbf{E}_{7,1}^2(t)$ are displayed in Figure 15.16b.

The explanation for this behavior is that the numerical solution of the nominal DDE-IVP has reached a steady state for $t \gtrsim 50$. Hence, Colsol-DDE – as an efficient variable-stepsize solver – takes large integration steps. However, these large integration steps are unsuitable for an accurate computation of the sensitivities.

It should be noted that the computation of error-controlled sensitivities leads only to a moderate increase of the number of integration steps and of the computation time. More precisely, 129 integration steps and $0.064s$ of computation time are needed if the error control is applied only to the nominal solution, and 316 integration steps and $0.21s$ of computation time are needed if the error control is applied to both the nominal solution and to the sensitivities.

$\eta^1(200)$	$100 \cdot \epsilon_{rel}^{nom}$	$\mathbf{E}^1(200)$	$100 \cdot \epsilon_{rel}^{sens}$
6.0150579	0.26	5142.9684	9.0
6.0007368	0.16	**-1012.3459**	**215**
6.0277592	0.012	-2028.8327	47.1
5.9970402	0.0043	2983.9745	43.1
6.0047092	0.14	4330.0101	18.9
6.0113466	0.14	-2522.1552	25.6
6.0223561	0.073	**133.62396**	**113**
6.0077961	0.077	**-491.35630**	**212**
$\eta^2(200)$	$100 \cdot \epsilon_{rel}^{nom}$	$\mathbf{E}^2(200)$	$100 \cdot \epsilon_{rel}^{sens}$
6.0307066	0.00035	4718.9988	0.0083
5.9913891	0.00077	880.42225	0.056
6.0285092	0.00063	-3834.0230	0.0074
5.9972526	0.00078	5247.0746	0.0075
6.0131070	0.00023	5337.4061	0.00033
6.0026613	0.00028	-2008.7611	0.012
6.0267833	0.00045	-1033.6090	0.026
6.0031660	0.00039	436.87860	0.055

Table 15.5.: Sensitivity computation with and without error control. The upper part of the table gives, in the columns 1 and 3, the results $\eta^1(200)$ and $\mathbf{E}^1(200)$ that are obtained without error control on the sensitivities. The columns 2 and 4 contain the relative errors of these results. The lower part of the table gives the corresponding results $\eta^2(200)$ and $\mathbf{E}^2(200)$ and relative errors for the computations with error control on the sensitivities.



(a) (b)

Figure 15.16.: Sensitivity of the component $y_7(t;c)$ of the solution of the DDE-IVP (15.38), (15.40) with respect to the scalar parameter c: (a) The reference sensitivity $\mathbf{E}_{7,1}^{ref}(t)$ is displayed as a solid black line. The approximation $\mathbf{E}_{7,1}^{1}(t)$, computed without error control on the sensitivity, is displayed as a dashed gray line. The approximation $\mathbf{E}_{7,1}^{2}(t)$, computed with error control on the sensitivity, is displayed as a dashed black line and is completely overlayed by the reference sensitivity. (b) Absolute error $\Delta\mathbf{E}_{7,1}(t;c) = \mathbf{E}_{7,1}^{ref}(t;c) - \mathbf{E}_{7,1}^{j}(t;c)$ in the computed sensitivities for $j = 1, 2$. The absolute error of $\mathbf{E}_{7,1}^{1}(t)$ (no error control on sensitivities) is displayed as a dashed gray line, and the absolute error of $\mathbf{E}_{7,1}^{2}(t)$ (with error control on sensitivities) is displayed as a dashed black line.

16. Parameter Estimation

> *Stimmen nun die mittelst den Formeln erhaltenen Werte mit denen durch Experimente erhaltenen in genügendem Umfange überein, hat man folgendes erreicht: 1. Eine Erklärung der betreffenden Erscheinungen [...]; 2. einen Ausgangspunkt für [...] numerische Berechnungen [...], und schließlich 3. gewisse Möglichkeiten zu Voraussagungen [...].*
>
> Sievert, in his paper "Zur theoretisch-mathematischen Behandlung des Problems der biologischen Strahlenwirkung" [237], describing an essential motivation for mathematical modeling and parameter estimation.

This chapter presents numerical results for parameter estimation problems in the context of delay differential equations (DDEs) and hybrid discrete-continuous delay differential equations (HDDEs).

Numerical Results Presented in This Chapter

Parameter estimation results are presented for three models.

First, an HDDE model is considered that has been proposed by Sievert [237] for the damaging effect of radioactive radiation on cells. This model is used as a basis for the formulation of a non-smooth parameter estimation problem. On this non-smooth problem, the convergence behavior of the derivative-based Gauss-Newton method is tested. For this purpose, initial guesses are generated for which the differentiability assumption in the local contraction theorem (Theorem 11.6) is violated. Nevertheless, convergence of the method is observed in practice.

Second, parameter estimation results are presented for the DDE model of the crosstalk of the IL-6 and GM-CSF signaling pathways, see Section 3.2. In comparison to an ODE model developed by Sommer et al. [240], the DDE model is smaller (14-dimensional instead of 16-dimensional state vector) but it fits the experimental data better.

Finally, parameter estimation results are presented for the HDDE model of the voting behavior of the viewers of the German TV singing competition "Unser Star für Baku", which was aired in 2012. A very good agreement of the model and the results in the TV show is obtained. A statistical analysis of the solution of the parameter estimation problem reveals that the "laziness" of the TV viewers and the time delay in the voting procedure have significant effects on the results displayed in the livescore.

Organization of This Chapter

The chapter is subdivided into three sections. Section 16.1 investigates the convergence behavior of the Gauss-Newton method for the application to a non-smooth parameter estimation problem. Section 16.2 presents parameter estimation results for the DDE model of the crosstalk of the IL-6 and GM-CSF signaling pathways. The results for the parameter estimation of the HDDE model for the voting behavior of the viewers of the TV show "Unser Star für Baku" are presented in the concluding Section 16.3.

16.1. Non-Smooth Least-Squares Problems: Convergence Behavior of Gauss-Newton Method

16.1.1. An HDDE Model for the Irradiation of Cells

In this section, the convergence behavior of the Gauss-Newton method is investigated for the application to a non-smooth least-squares problem. The considered differential equation model is

the earliest hybrid-discrete continuous delay differential equation (HDDE) known to the author, which is due to Sievert [237]:

$$\dot{y}(t;c) = \begin{cases} -c_6 y(t;c) + c_7 \cdot (c_5 - y(t-\tau(c);c)) & \text{for } \zeta(t) = (-1,+1,-1)^T \\ -c_6 y(t;c) + c_8 \cdot (c_5 - y(t-\tau(c);c)) & \text{for } \zeta(t) = (+1,+1,-1)^T \\ c_7 \cdot (c_5 - y(t-\tau(c);c)) & \text{for } \zeta(t) = (-1,-1,\pm 1)^T,\ \zeta(t) = (-1,+1,+1)^T \\ c_8 \cdot (c_5 - y(t-\tau(c);c)) & \text{for } \zeta(t) = (+1,-1,\pm 1)^T,\ \zeta(t) = (+1,+1,+1)^T \end{cases} \tag{16.1}$$

Herein, the constant but parameter-dependent delay is given by

$$\tau(t, y(t), c) \equiv \tau(c) = c_4. \tag{16.2}$$

Further, $\zeta(t) = (\zeta_1(t), \zeta_2(t), \zeta_3(t))^T$, and $\zeta_i(t)$ are the signs of switching functions:

$$\zeta_i(t) = \operatorname{sign}(\sigma_i(t, y^-(t;c), c, y^-(t-\tau(c);c))) \quad \text{for} \quad i = 1,2,3. \tag{16.3}$$

The three switching functions are defined as follows:

$$\sigma_1(t, y(t;c), c, y(t-\tau(c);c)) \equiv \sigma_1(y(t-\tau(c);c), c) = c_5 - y(t-\tau(c);c) \tag{16.4a}$$
$$\sigma_2(t, y(t;c), c, y(t-\tau(c);c)) \equiv \sigma_2(t,c) = t - c_2 \tag{16.4b}$$
$$\sigma_2(t, y(t;c), c, y(t-\tau(c);c)) \equiv \sigma_3(t,c) = t - c_3. \tag{16.4c}$$

The first of these switching functions is state-dependent while the other two are simple time-dependent. The following initial condition is used:

$$y(t;c) = c_1 \quad \text{for} \quad t \leqslant 0. \tag{16.5}$$

This means that $t^{ini}(c) \equiv t^{ini} = 0$, $\phi(t,c) \equiv \phi(0,c) = y^{ini}(c) = c_1$. The final time is set to $t^{fin}(c) \equiv t^{fin} = 300$.

16.1.2. Background

The HDDE (16.1) has been proposed by Sievert [237] as a heuristic model to describe the influence of radioactive radiation on biological cells. The fundamental assumption underlying the model has been described by Sievert [237] as follows:

> "If 'something', X, changes in the cell due to a harmful external force, then reconstructing internal forces become active that aim at bringing X back to its nominal value".

Here, the nominal – or "ideal" value – is represented by the parameter c_5. The reconstructing internal forces are subject to a time delay $\tau(c) = c_4$. If the past state $y(t-\tau(c);c)$ is larger than c_5, then $\zeta_1(t) = -1$, and the reconstructing force is proportional to the parameter c_7. If the past state $y(t-\tau(c);c)$ is smaller than c_5, then $\zeta_1(t) = +1$, and the reconstructing force is proportional to c_8.

The zeros of the simple time-dependent switching functions σ_2 and σ_3 characterize the beginning and end of the time interval in which a damaging radioactive radiation is applied to the cell. More precisely, the damaging influence is proportional to c_6, and it is applied for $t \in [c_2, c_3]$.

16.1.3. Measurement Data

For the investigations in this section, artificial measurement data are generated. Therefore, the HDDE-IVP (16.1), (16.5) is solved with Colsol-DDE for the following parameter values:

$$\begin{aligned} c_1^\star = 90, \quad c_2^\star = 50, \quad c_3^\star = 70, \quad c_4^\star = 30 \\ c_5^\star = 100, \quad c_6^\star = 0.06, \quad c_7^\star = 0.025, \quad c_8^\star = 0.02. \end{aligned} \tag{16.6}$$

In the context of the parameter estimation problem setup below, the parameter values $c^\star$ are considered as the *correct parameters*, cf. Chapter 10.1.

The time points $t_j = j - 1$, $j = 1, \ldots, 301$, are used as measurement times. For each measurement time, a measurement value is obtained by adding a Gaussian random number with standard deviation $\sigma = 5$ (variance $\sigma^2 = 25$) to the HDDE-IVP solution, i.e.

$$\eta_j = y(t_j; c^\star) + \epsilon_j, \quad \epsilon_j \sim \mathcal{N}(0, 25). \tag{16.7}$$

The HDDE-IVP solution $y(t; c^\star)$ and the generated measurement data are displayed in Figure 16.1.

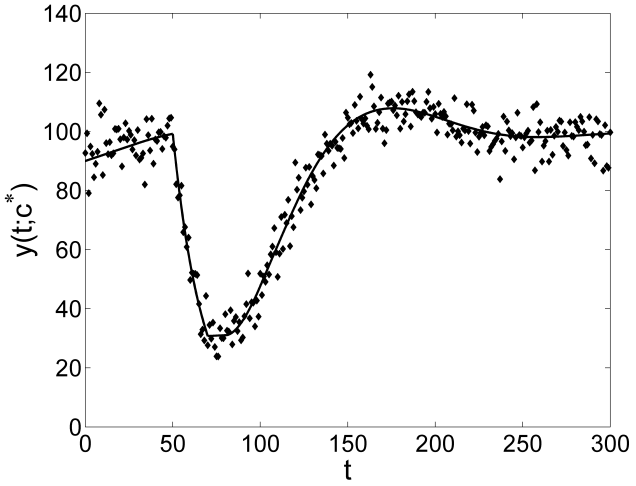


Figure 16.1.: Solution $y(t; c^\star)$ of the HDDE-IVP (16.1), (16.5), displayed as a solid line, and simulated measurement data with standard deviation $\sigma = 5$, displayed as diamonds.

16.1.4. Setup of Optimization Problem

The least-squares optimization problem

$$\min_c \|F_1(c)\|_2^2 \tag{16.8}$$

is considered, where the components of $F_1(c)$ are defined as follows:

$$F_{1,i}(c) := \frac{\eta_i - y(t_i; c)}{\sigma_i} \quad \text{for} \quad i = 1, \ldots, 301. \tag{16.9}$$

Therein, $y(t_i; c)$ represents the solution of the HDDE-IVP (16.1), (16.5) at the measurement time t_i for given parameters c. Further, $\sigma_i = 5$ for $i = 1, \ldots, 300$ are the standard deviations of the simulated measurement data.

16.1.5. Solution of the Optimization Problem

The solution of the optimization problem is given in the fourth column of Table 16.1 ("estimated values"). Table 16.1 also contains the square roots of the diagonal elements of the matrix $\tilde{\mathbf{V}}_c$ (see equation (12.17)). These diagonal elements can be used as a measure for the uncertainty in the parameter estimates. However, they are only approximations of the standard deviations of the parameters, because the objective function is a nonlinear function of the parameters, and furthermore non-differentiable (as described in the following).

16.1.6. Smoothness Considerations

Smoothness of the Sensitivities

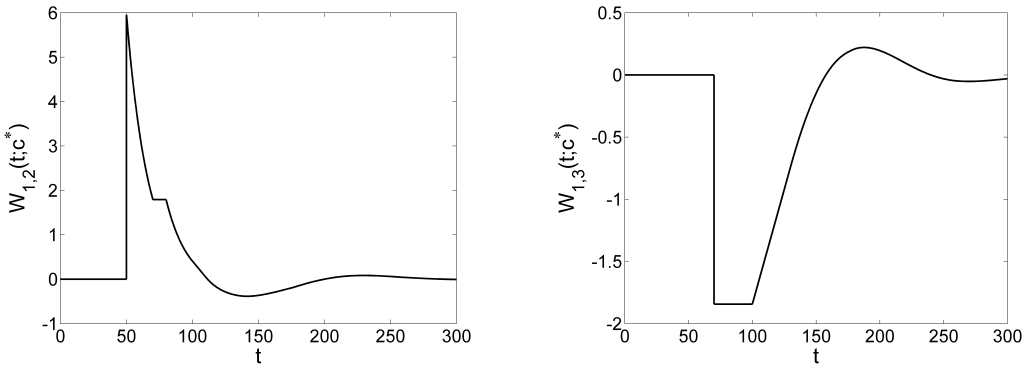
The sensitivities of the HDDE-IVP solution $y(t; c)$ with respect to the parameters c_2 and c_3, i.e.

$$\mathbf{W}_{1,2}(t; c) = \frac{\partial y(t; c)}{\partial c_2} \quad \text{and} \quad \mathbf{W}_{1,3}(t; c) = \frac{\partial y(t; c)}{\partial c_3}, \tag{16.10}$$

Parameter	Description	Correct Value $c_i^\star$	Est. Value $\hat{c}_i$	"Std. Dev."	"Std Dev. [%]"
c_1	initial value	90	93.3	1.3	1.3
c_2	start of irradiation	50	49.50	0.52	1.0
c_3	end of irradiation	70	70.4	1.1	1.5
c_4	time delay	30	30.31	0.62	2.1
c_5	optimal value	100	99.63	0.88	0.9
c_6	intensity of irradiation	0.06	0.0561	0.0033	5.9
c_7	reconstr. force for too large values	0.025	0.0255	0.0053	20.8
c_8	reconstr. force for too small values	0.02	0.01980	0.00047	2.4

Table 16.1.: Parameter estimation results for the irradiation of biological cells. The second column gives the description of the parameters and the third column gives the correct values $c_i^\star$ that were used for the generation of measurement data, cf. equation (16.6). The fourth column gives the result of the parameter estimation, i.e. the "estimated values" $\hat{c}_i$. The fifth column gives the square roots of the diagonal elements of the matrix $\tilde{\mathbf{V}}_c$ (see equation (12.17)), which are approximations of the standard deviations of the parameters. Eventually, the sixth column gives the relative values of the so-obtained standard deviations (in %).
In order to obtain an approximate 90% confidence interval, the diagonal elements given in the fifth column have to be multiplied by the quantile of the χ^2-distribution with 8 degrees of freedom, i.e. $q(\chi_8^2, 0.9) \approx 3.49$.
All approximations of absolute standard deviations are given with two digits precision. The estimated values are given with an according number of digits. All approximations of relative standard deviations are given with one digit after the decimal point.

are discontinuous at the points $t = c_2$ and $t = c_3$, respectively. Furthermore, the sensitivity $\mathbf{W}_{1,2}(t;c)$ is non-differentiable at $t = c_3$ and at $t = c_2 + c_4$, and the sensitivity $\mathbf{W}_{1,3}(t;c)$ is non-differentiable at $t = c_3 + c_4$. This is illustrated in Figure 16.2 for $c = c^\star$, where $c_2^\star = 50$, $c_3^\star = 70$, $c_2^\star + c_4^\star = 80$, and $c_3^\star + c_4^\star = 100$. The sensitivities with respect to other parameters have time



(a) (b)

Figure 16.2.: Sensitivity of the solution $y(t;c^\star)$ of the HDDE-IVP (16.1), (16.5), (a) with respect to the parameter c_2 and (b) with respect to the parameter c_3.

points of non-differentiability, too. For example, the sensitivity of the HDDE-IVP solution with respect to the delay,

$$\mathbf{W}_{1,4}(t,c) = \frac{\partial y(t;c)}{\partial c_4}, \tag{16.11}$$

is non-differentiable at $t = c_4$, $t = c_2$, $t = c_3$, $t = c_2 + c_4$, and $t = c_3 + c_4$. For $c = c^\star$, the non-differentiabilities are located at $t = 30$, $t = 50$, $t = 70$, $t = 80$, and $t = 100$, see Figure 16.3.

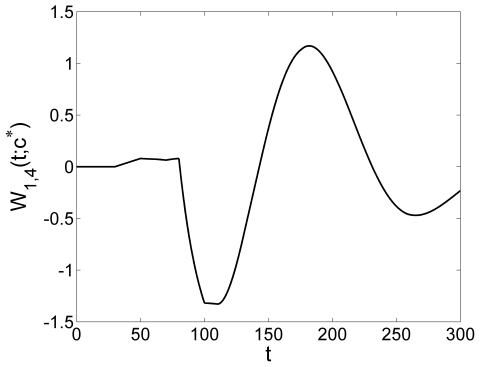


Figure 16.3.: Sensitivity of the solution $y(t; c^\star)$ of the HDDE-IVP (16.1), (16.5) with respect to the parameter c_4.

Smoothness of the Optimization Problem

Formally, the derivatives of the individual components $F_{1,i}(c)$ of the function $F_1(c)$ are given by

$$\frac{dF_{1,i}(c)}{dc} = -\frac{1}{\sigma_i}\mathbf{W}(t_i; c). \tag{16.12}$$

Hence, discontinuities and non-differentiabilities in $\mathbf{W}(t; c)$ directly lead to a non-smooth dependence of $F_1(c)$ on the parameters c. More precisely, the function $F_1(c)$ is non-differentiable at an evaluation point c whenever at least one of the measurement times t_i is identical to one of the time points of discontinuity of $\mathbf{W}(t; c)$. This is the case whenever c_2 or c_3 assume integer values, because the measurement times are $t_i = 0, 1, \ldots, 300$. Furthermore, the function $F_1(c)$ fails to be twice continuously differentiable whenever at least one of the measurement times t_i is identical to one of the time points of non-differentiability of $\mathbf{W}(t, c)$.

16.1.7. Numerical Investigation of Convergence Behavior

The lack of smoothness of the function F_1 raises the suspicion that derivative-based optimization methods such as the Gauss-Newton method may show a very poor convergence behavior. For example, a necessary condition for the application of the local contraction theorem (Theorem 11.6) is that $F_1(c)$ is a continuously differentiable function on a ball in parameter space that contains both the initial guess c^0 and the solution $\hat{c}$ of the optimization problem. For the particular parameter estimation problem considered here, this implies that the initial guess c^0 should be such that

$$c_2^0 \in [49, 50] \quad \text{and} \quad c_3^0 \in [70, 71], \tag{16.13}$$

because $t = c_2$ and $t = c_3$ are time points of discontinuity of $\mathbf{W}(t, c)$, and because 49, 50, 70, and 71 are measurement times.

Furthermore, it should be noted that the use of the restrictive monotonicity test as a globalization strategy for the Gauss-Newton method (see Section 11.3) looses its theoretical justification in the case of non-smooth least-squares problems. In particular, the descent property of the natural level function (equation (11.31)) described in Lemma 11.11 may be lost.

In order to test the convergence in the numerical practice, 10 initial guesses are generated by adding random numbers to the estimated parameter values $\hat{c}$. More precisely, the parameters $\hat{c}_1, \ldots, \hat{c}_5$ are perturbed by normally distributed random numbers with standard deviation 10 (variance 100), and the parameters $\hat{c}_6, \hat{c}_7$, and $\hat{c}_8$ are perturbed by normally distributed random numbers with standard deviation 0.01 (variance 0.0001):

$$c_i^0 \sim \mathcal{N}(\hat{c}_i, 100) \quad \text{for} \quad i = 1, \ldots, 5 \tag{16.14a}$$

$$c_i^0 \sim \mathcal{N}(\hat{c}_i, 0.0001) \quad \text{for} \quad i = 6, 7, 8. \tag{16.14b}$$

The use of a smaller standard deviation for the initial guesses of the latter three parameters is motivated by the smaller absolute values of the corresponding components of $\hat{c}$.

The generated initial guesses are given in Table 16.2. Please note that none of the initial guesses is such that the conditions (16.13) are fulfilled. Accordingly, the local contraction theorem (Theorem 11.6) does not apply, and it is not guaranteed that the method converges.

	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
initial guess 1	101.44	40.62	71.44	36.02	103.76	0.03279	0.03453	0.00144
initial guess 2	83.41	44.06	73.48	24.27	101.40	0.05675	0.02587	0.04207
initial guess 3	90.21	43.50	75.34	40.78	97.65	0.05539	0.02044	0.02216
initial guess 4	96.56	56.89	87.56	34.71	93.46	0.05854	0.02622	0.01371
initial guess 5	96.03	47.56	49.06	47.61	93.55	0.04386	0.02868	0.00637
initial guess 6	85.91	41.10	83.99	38.98	98.83	0.04576	0.03883	0.01561
initial guess 7	102.27	38.78	80.05	31.44	104.03	0.05468	0.03451	0.01680
initial guess 8	94.30	50.74	84.81	11.77	88.23	0.06638	0.02206	0.02993
initial guess 9	82.35	29.89	68.46	28.13	105.05	0.06238	0.02339	0.01114
initial guess 10	97.17	37.42	99.52	42.54	86.80	0.04565	0.02281	0.01542

Table 16.2.: Initial guesses used for testing the convergence behavior of a damped Gauss-Newton method applied to a non-smooth parameter estimation problem. Each row corresponds to one initial guess c^0.

The practical performance of the damped Gauss-Newton method realized in ParamEDE is, however, very satisfactory. For all 10 initial guesses, the method converges to the solution $\hat{c}$ given in Table 16.1. The number of executed iterations are given in Table 16.3. For 9 of 10 initial guesses, the method converges in at most 12 iterations. Furthermore, for initial guess 2 and 7, the method uses exclusively full-step iterations, i.e. no globalization strategy is needed in order to solve the problem.

	No. of iterations	No. of full-step iterations
initial guess 1	66	9
initial guess 2	7	7
initial guess 3	7	6
initial guess 4	9	5
initial guess 5	12	6
initial guess 6	7	5
initial guess 7	7	7
initial guess 8	8	6
initial guess 9	11	6
initial guess 10	9	6

Table 16.3.: Number of iterations and number of full-step iterations needed to solve the parameter estimation problem for the irradiation of cells.

Discussion and Conclusion

The good convergence behavior of the damped Gauss-Newton method is possibly related to the fact that the objective function is differentiable except for a set of measure zero, and that the solution $\hat{c}$ is not contained in this set. Starting from an initial guess (which is also not contained in the set of points of non-differentiability), the increments of the Gauss-Newton method are such that the iterates eventually end up in a domain where the conditions (16.13) are fulfilled. Then, in the neighborhood of the solution $\hat{c}$, the assumptions of the local contraction theorem are fulfilled, and the method converges with full step iterations.

The results of this section show that derivative-based optimization methods can be quite successful for solving non-smooth problems, even though classical convergence theory does not apply. A more detailed analysis of the reasons of the observed good convergence behavior is subject to future work.

16.2. Crosstalk of the Signaling Pathways of IL-6 and GM-CSF

This section gives parameter estimation results for the DDE model describing the interaction ("crosstalk") of the signaling pathways of IL-6 and GM-CSF (see Section 3.2).

16.2.1. Measurement Data

Cells were exposed to four different experimental settings:

(a) stimulation with IL-6

(b) stimulation with IL-6 and with a blocking antibody for the GM-CSF receptor complex

(c) stimulation with both IL-6 and GM-CSF

(d) no stimulation.

For each setting, three independent experiments were done, and the concentration of pSTAT-3 in the cytoplasm was measured at 11 time points: $0, 5, 10, 15, 20, 25, 30, 45, 60, 90, 120$. For the stimulation with IL-6, also the concentration of SOCS-3 was measured at these time points. The SOCS-3 measurements take into account both "free" SOCS-3 molecules and those SOCS-3 molecules that are bound in deactivated IL-6 and GM-CSF receptor complexes.

Detailed information on the experimental design, on the measurement techniques, and on the postprocessing of obtained measurement data can be found in Sommer et al. [240].

16.2.2. Setup of Optimization Problem

Both the ODE model and the DDE model described in Section 3.2 contain 12 parameters. For the ODE model, the parameters are $p, \alpha_r, b, \alpha_{SK}, \alpha_{STAT3+}, \delta_{SOCS3}, \delta_r, \alpha_{STAT3}, \nu, \mu_1, \mu_2$, and γ. For the DDE model, ν, μ_1, and μ_2 are replaced by τ_1, τ_2, and κ. In addition, a scaling parameter has to be introduced for the parameter estimation because the measurements of SOCS-3 are in abitrary units, i.e. no absolute values for the concentrations are available. Hence, in total, 13 parameters need to be estimated.

Let the 13-dimensional parameter vector be denoted by c. Then a least-squares parameter estimation problem of the following form is considered:

$$\min_c \|F_1(c)\|_2^2. \tag{16.15}$$

The function $F_1(c)$ has 150 components. The first 120 components correspond to the measurements of pSTAT-3 that are available (4 different stimulations, 3 independent experiments for each stimulation, 10 measurement times for each experiment and each stimulation). The measurements at $t = 0$ are not used as least-squares terms, but are instead exploited in order to determine the initial concentration of pSTAT-3. The remaining 30 components of $F_1(c)$ correspond to the measurements of SOCS-3 (3 independent experiments for the stimulation with IL-6, 10 measurement times for each experiment).

The first 120 components of $F_1(c)$ are weighted differences between the pSTAT-3 measurement values and $y_9(t_j; c)$, i.e. of the 9-th component of the IVP solution at the corresponding measurement time. The remaining 30 components of $F_1(c)$ are weighted differences between the SOCS-3 measurement values (scaled with the factor c_{13}) and $y_{14}(t_j; c) + y_{15}(t_j; c) + y_{16}(t_j; c)$. The summation over the three state vector components is necessary because the measurement technique does not distinguish between "free" SOCS-3 molecules and SOCS-3 molecules that are bound in the deactivated IL-6 and GM-CSF receptor complexes. For all components of the function $F_1(c)$, sample standard deviations are used as weighting factors.

Please note that the problem considered here is a *multi-experiment parameter estimation problem*: 4 IVP solutions are required for the evaluation of $F_1(c)$, one for each of the 4 stimulations. The IVP solutions differ by the values of u_{IL6} and u_{GMCSF} that are used in the right-hand-side of the differential equation system (see equation (3.16)), and further by the employed initial values. For the first experiment, the stimulation with IL-6, one has $u_{IL6} = 0.004$, $u_{GMCSF} = 0$, and the

Par.	Notation of Section 3.2	Description	Estimated Value $\hat{c}_i$	"Std. Dev."	"Std. Dev. [%]"
c_1	p	production rate of IL-6 and GM-CSF	$5.00 \cdot 10^{-7}$	$3.9 \cdot 10^{-8}$	7.8
c_2	α_r	activation rate of IL-6 and GM-CSF receptors	3000	undet.	undet.
c_3	b	blockade of GM-CSF receptor (overstimulation)	100000	undet.	undet.
c_4	α_{SK}	activation rate of SK by active GM-CSF receptor	8000	undet.	undet.
c_5	α_{STAT3+}	additional STAT-3 activation rate on active IL-6 receptor due to presence of SK	30000	undet.	undet.
c_6	δ_{SOCS3}	degradation rate of SOCS-3	0.0224	0.0022	9.8
c_7	δ_r	deactivation rate of IL-6 and GM-CSF receptors by SOCS-3	920	200	21.7
c_8	α_{STAT3}	STAT-3 activation rate on active IL-6 receptor	0.4104	0.0070	1.7
c_9	ν	import rate of pSTAT-3 into nucleus	0.0378	0.0015	4.0
c_{10}	τ_1	time that pSTAT-3 remains in nucleus	5.00	undet.	undet.
c_{11}	τ_2	time delay between import of pSTAT-3 into nucleus and production of SOCS-3	10.11	0.95	9.4
c_{12}	κ	κ/ν gives the number of produced SOCS-3 per pSTAT-3 that is imported into the nucleus	0.093	0.019	20.4
c_{13}	—	scaling constant for SOCS-3 measurements	1.20	0.21	17.5

Table 16.4.: Parameter estimation results for the crosstalk of the IL-6 and GM-CSF signaling pathways. The first column gives the generic parameter symbol c_i and the second column gives the parameter symbol used in Section 3.2. The third column recalls the meaning of the parameter. The fourth column gives the estimated values $\hat{c}_i$ of the parameters, and the fifth column gives the square roots of the diagonal elements of the matrix $\tilde{\mathbf{V}}_c$ (see equation (12.17)). These are approximations of the standard deviations of the parameters. Eventually, the last column gives the relative values of the so-obtained standard deviation (in %). The mark "undet." in the last two columns indicates those parameters that are undetermined (i.e., these parameters have very large standard deviations).
All standard deviations (fifth column) have been rounded to two digits. The estimated values are given with an according number of digits. Relative standard deviations (sixth column) are given with one digit after the decimal point.

following initial values[1]:

$$y^{ini} = (0, 7.958 \cdot 10^{-4}, 0, 0, 4 \cdot 10^{-7}, 0.007958, 0, 0.1454, 4.511 \cdot 10^{-4}, 0, 0.033, 0, 0, 0, 0)^T. \quad (16.16)$$

For the second experiment, the stimulation with IL-6 and with a blocking antibody for the GM-CSF receptor complex, one has $u_{IL6} = 0.004$, $u_{GMCSF} = 0$ and $y_2^{ini} = 0$. This means that the action of the blocking antibody is simulated by setting the initial value for the GM-CSF receptor complex to 0. The initial values of all other components are the same as in equation (16.16). For the third experiment, the double stimulation with IL-6 and GM-CSF, one has $u_{IL6} = 0.004$, $u_{GMCSF} = 0.004$, and y^{ini} as given in equation (16.16). Finally, for the fourth experiment, one has $u_{IL6} = 0$, $u_{GMCSF} = 0$ and y^{ini} as given in equation (16.16).

For all computations with the DDE model, a constant initial function is used, and the values are identical to y^{ini}, i.e. formally $\phi(t, c) \equiv \phi(0, c) = y^{ini}$.

For a derivation of the employed initial values, it is referred to Sommer et al. [240].

16.2.3. Parameter Estimation Results

For the DDE model, the estimated values $\hat{c}$ of the parameters given in Table 16.4 have been obtained by using ParamEDE. For the ODE model, estimated values $\check{c}$ of the parameters can be found in Sommer et al. [240]. Both models contain five parameters that are unidentifyable ("undetermined") because they have locally (in the vicinity of the estimated parameter values) no measurable influence on the IVP solution.

Among the identifyable parameters of the DDE model is $c_{11} = \tau_2$, the time delay between the import of pSTAT-3 into the nucleus and the formation of SOCS-3. This time delay is given by $\tau_2 = 10.11 \pm 0.95$ (minutes). Furthermore, also the two parameters c_9 and c_{12} are identifyable.

[1] For the sake of notational consistency with the ODE model, a 16-dimensional state vector is used in the DDE model, although the 10-th and 13-th component (which represent nuclear pSTAT-3 and SOCS-3 mRNA in the ODE model) are not needed.

Hence, the ratio $c_{12}/c_9 = \kappa/\nu$ is identifyable, which is approximately 2.46. This gives the number of SOCS-3 molecules that are produced per pSTAT-3 molecule that is imported into the nucleus.

The Figures 16.4 and 16.5 show the fit of the DDE model to the experimental data. In addition, also the fit of the ODE model developed by Sommer et al. [240] is displayed. The two models differ only slightly, and both show a good agreement with the data. However, a computation of the least-squares sums reveals that the DDE model fits the data slightly better (decrease of approximately 20%). This is remarkable since the DDE model is "simpler" in the sense that the dimension of the differential equation system is reduced from 16 to 14 (see Section 3.2). This shows exemplarily that DDEs can be an interesting alternative to ODEs for modeling biological processes.

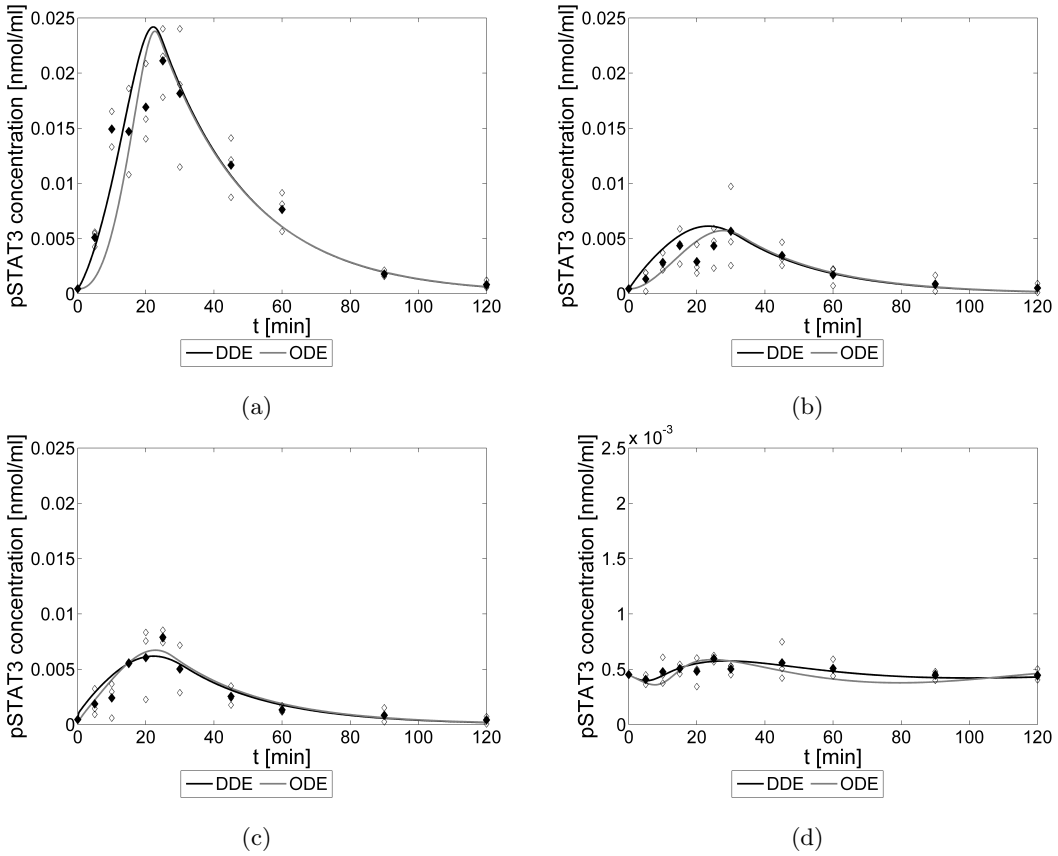
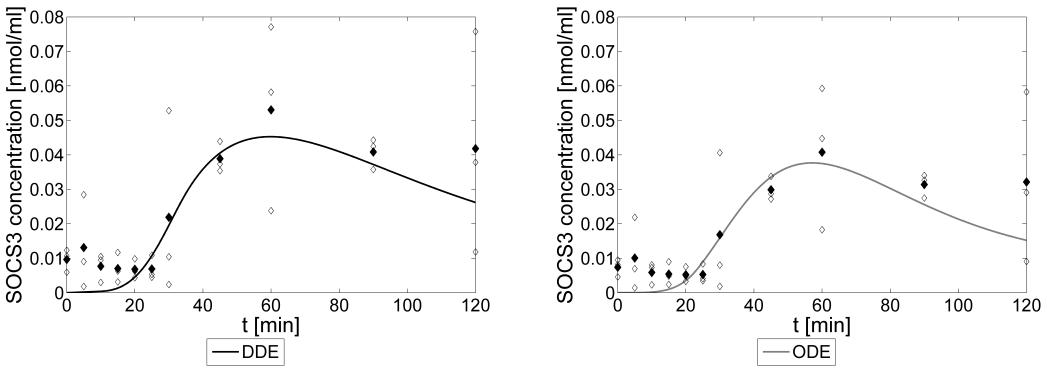


Figure 16.4.: Parameter estimation results for the crosstalk of the IL-6 and GM-CSF signaling pathways: Fit of the ODE-IVP solution and of the DDE-IVP solution to the pSTAT-3 measurement data. Figure (a) shows the results for the stimulation with IL-6, (b) shows the results for the stimulation with IL-6 and simultaneous blockade of the GM-CSF receptor. Further, (c) shows the result for the double stimulation with both IL-6 and GM-CSF, and (d) shows the results for unstimulated cells. Please note that a different scaling has been chosen for the vertical axis in Figure (d).
In all figures, the component $y_9(t;\hat{c})$ of the DDE-IVP solution (for the estimated values $\hat{c}$ of the parameters, see Table 16.4) is displayed as a solid black line. The component $y_9(t;\check{c})$ of the ODE-IVP solution (for the estimated values $\check{c}$ of the parameters of the ODE-model, see Sommer et al. [240]) is displayed as a solid gray line.
For each of the four different stimulations, three independent experiments were done. The measurement values obtained in the individual experiments are shown as small diamonds with black outline and white filling, the means of the three measurements are displayed as larger and solid diamonds.



(a) (b)

Figure 16.5.: Parameter estimation results for the crosstalk of the IL-6 and GM-CSF signaling pathways: Fit of the ODE-IVP solution and of the DDE-IVP solution to the SOCS-3 measurement data. Figure (a) shows the result for the DDE model. The solid black line represents $y_{14}(t;\hat{c}) + y_{15}(t;\hat{c}) + y_{16}(t;\hat{c})$ for the estimated values $\hat{c}$ of the parameters of the DDE model. Figure (b) shows the result for the ODE model. Here, the solid gray line represents $y_{14}(t;\check{c}) + y_{15}(t;\check{c}) + y_{16}(t;\check{c})$ for the estimated values $\check{c}$ of the parameters of the ODE model.
In both figures, the measurements values obtained in the three independent experiments are displayed as small diamonds with black outline and white filling, and the means of the three values are displayed as larger and solid diamonds. Please note that the measurement data displayed in (a) and (b) incorporate slightly different scaling factors $\hat{c}_{13}$ and $\check{c}_{13}$.

16.3. "Unser Star für Baku"

In this section, parameter estimation results are presented for the HDDE model of the voting behavior of the viewers of the TV singing competition "Unser Star für Baku" (see Section 3.3).

16.3.1. Measurement Data

Measurement data have been obtained from the 3rd episode of the show. In this episode, 10 candidates were competing. The candidates on the ranks 1-8 were allowed to return in the 4th episode, while the last 2 candidates had to leave the competition. The time interval between the last commerical break and the end of the voting time has a length of $390s$. During this time interval, the percentage of votes for each of the 10 candidates is recorded every 5 seconds[2]. In total, this gives $79 \cdot 10 = 790$ measurement values.

16.3.2. Setup of Optimization Problem

Since an episode is considered in which 8 of 10 candidates are selected for the next round of the competition, an HDDE model with a 11-dimensional state vector is constructed as explained in Section 3.3. Thereby, the state vector components $y_1(t;c), \ldots, y_{10}(t;c)$ represent the percentages of votes for the 10 candidates and $y_{11}(t;c)$ represents the total number of votes.

There are 24 parameters in the model equations. At first, there are 10 initial values, denoted by $c_1, \ldots, c_{10}$, which represent the percentages of votes that the candidates have at the beginning of the considered time interval. There are further 10 parameters $c_{11}, \ldots, c_{20}$, which represent the sizes of the fan-bases for each of the 10 candidates (these parameters have been called k_i in Section 3.3). Eventually, there is the laziness parameter c_{21}, the time delay c_{22}, the panic factor c_{23}, and the duration of the panic c_{24}, which have been called λ, τ, ρ, and δ in Section 3.3, respectively.

A least-squares approach for parameter estimation is used, i.e. the following minimization prob-

[2]The show can, at the time of submission of this thesis, still be watched online under `http://www.unser-star-fuer-baku.tv/videos/`

lem is considered:

$$\min_c \|F_1(c)\|_2^2. \tag{16.17}$$

Thereby, the i-th component of $F_1(c)$ takes the form

$$F_{1,i}(c) = \frac{\eta_{j,k} - y_j(t_k; c)}{\sigma_{j,k}}, \tag{16.18}$$

for one combination of indices j, k of the candidate and of the measurement time. Explicitly, the optimization problem thus takes the following form:

$$\min_c \sum_{j=1}^{10} \sum_{k=1}^{79} \left(\frac{\eta_{j,k} - y_j(t_k; c)}{\sigma_{j,k}} \right)^2. \tag{16.19}$$

The function $y(t; c)$ denotes the solution of an HDDE-IVP with the following initial condition:

$$y(t; c) = \big(c_1, \ldots, c_{10}, 100000\big)^T \quad \text{for} \quad t \leqslant 0. \tag{16.20}$$

Hence, it is assumed that a (guessed) total number of 100000 votes has been received during the show so far. Furthermore, the use of the constant initial function implies that no votes have been received a short time before the considered interval during the commercial break. This assumption is motivated by the fact that it yields the simplest possible parameterization of the initial function.

The denominators in equation (16.19) are chosen as $\sigma_{j,k} = 0.1$ for all j and k. This value corresponds to the fact that the percentage values are given with an accuracy of 0.1%.

16.3.3. Parameter Estimation Results

The estimated values $\hat{c}$ of the parameters, which are given in the third column of Table 16.5, have been obtained by using ParamEDE. For these values of the parameters, the HDDE-IVP solution fits the data taken from the TV show very good, see Figure 16.6. For 602 of the 790 data points, the deviation is less than 0.1%, which is the accuracy of the results displayed in the livescore. Furthermore, the largest deviation of the simulated results (solid lines) to the data (diamonds) is $\approx 0.26\%$ and occurs for candidate "Shelly" ($y_3(t; c)$, see Figure 16.6c), at $t = 300$.

In the fourth column of Table 16.5, the diagonal elements of the matrix $\tilde{\mathbf{V}}_c$ computed by means of equation (12.17) are given. These diagonal elements are approximations of the standard deviations of the parameters. Large relative standard deviations (fifth column of Table 16.5) are obtained for the fan-bases of all candidates, and the values are particularly large for the two candidates "Roman" and "Yana"; hence, c_{11} and c_{12} are characterized as undetermined in Table 16.5. The reason for this is the so-called "laziness", i.e. the viewers tend to vote for a candidate only if he or she is currently on one of the two loser ranks (rank 9 or rank 10), or in immediate danger of dropping to a loser rank. This is never the case for the candidates "Roman" and "Yana" during the last 390 seconds of the show, and hence, the size of their fan-base cannot be estimated from the data.[3]

Much smaller relative standard deviations are obtained for the delay and for the laziness of TV viewers. This indicates the important role that the delay and the laziness play for the dynamics observed in the livescore. Furthermore, it can be concluded from the values of the relative standard deviations that delay and laziness have a larger influence on the results than the size of the fan-bases of the candidates.

The smallest relative standard deviations are obtained for the parameters $c_1, \ldots, c_{10}$. This shows that the constant values of the initial function have a major influence on the HDDE-IVP solution. This is an undesirable effect, because the percentages of votes for the candidates prior to the considered time interval are unknown and might not have been constant. Future investigations on the livescore may therefore consider different parameterizations of the initial function.

[3] It can only be suspected that "Roman" and "Yana" have many fans, because otherwise they would not have received so many votes prior to the last 390 seconds of the show.

Parameter	Description	Estimated Value $\hat{c}_i$	"Std. Dev."	"Std. Dev." [%]
c_1	initial value "Roman"	13.309	0.018	0.1
c_2	initial value "Yana"	11.609	0.017	0.1
c_3	initial value "Shelly"	10.008	0.021	0.2
c_4	initial value "Sebastian"	9.361	0.031	0.3
c_5	initial value "Leonie"	9.416	0.015	0.2
c_6	initial value "Celine"	9.582	0.023	0.2
c_7	initial value "Ornella"	9.462	0.028	0.3
c_8	initial value "Umut"	9.468	0.016	0.2
c_9	initial value "Katya"	9.506	0.012	0.1
c_{10}	initial value "Rachel"	8.281	0.033	0.4
c_{11}	fan-base "Roman"	58	undet.	undet.
c_{12}	fan-base "Yana"	35	undet.	undet.
c_{13}	fan-base "Shelly"	49	21	42.9
c_{14}	fan-base "Sebastian"	2.05	0.74	36.1
c_{15}	fan-base "Leonie"	1.00	0.37	37.0
c_{16}	fan-base "Celine"	15.7	6.6	42.0
c_{17}	fan-base "Ornella"	12.7	5.3	41.7
c_{18}	fan-base "Umut"	1.65	0.61	37.0
c_{19}	fan-base "Katya"	0.87	0.31	35.6
c_{20}	fan-base "Rachel"	2.59	0.93	35.9
c_{21}	laziness	16.92	0.93	5.5
c_{22}	delay	35.69	0.49	1.4
c_{23}	panic factor	8.5	3.3	38.8
c_{24}	panic duration	331	19	5.7

Table 16.5.: Parameter estimation results for the voting behavior of the viewers of the TV singing competition "Unser Star für Baku". The second column gives the description for the 24 parameters together with the names of the 10 competing candidates. The third column gives the estimated values $\hat{c}$ of the parameters, which have been obtained by using ParamEDE. The fourth column gives the square roots of the diagonal elements of the matrix $\tilde{\mathbf{V}}_c$ (see equation (12.17)), which can be used as approximations of the standard deviations of the parameters, see Subsection 12.2.1. The parameters c_{11} and c_{12} have very large standard deviations and are thus characterized as "undetermined". Eventually, the fifth column gives the relative values of the so-computed standard deviations (in %).
All absolute standard deviations are given with two digits precision. The estimated values are given with a corresponding number of digits. All relative values of the standard deviations are given with one digit after the decimal point.

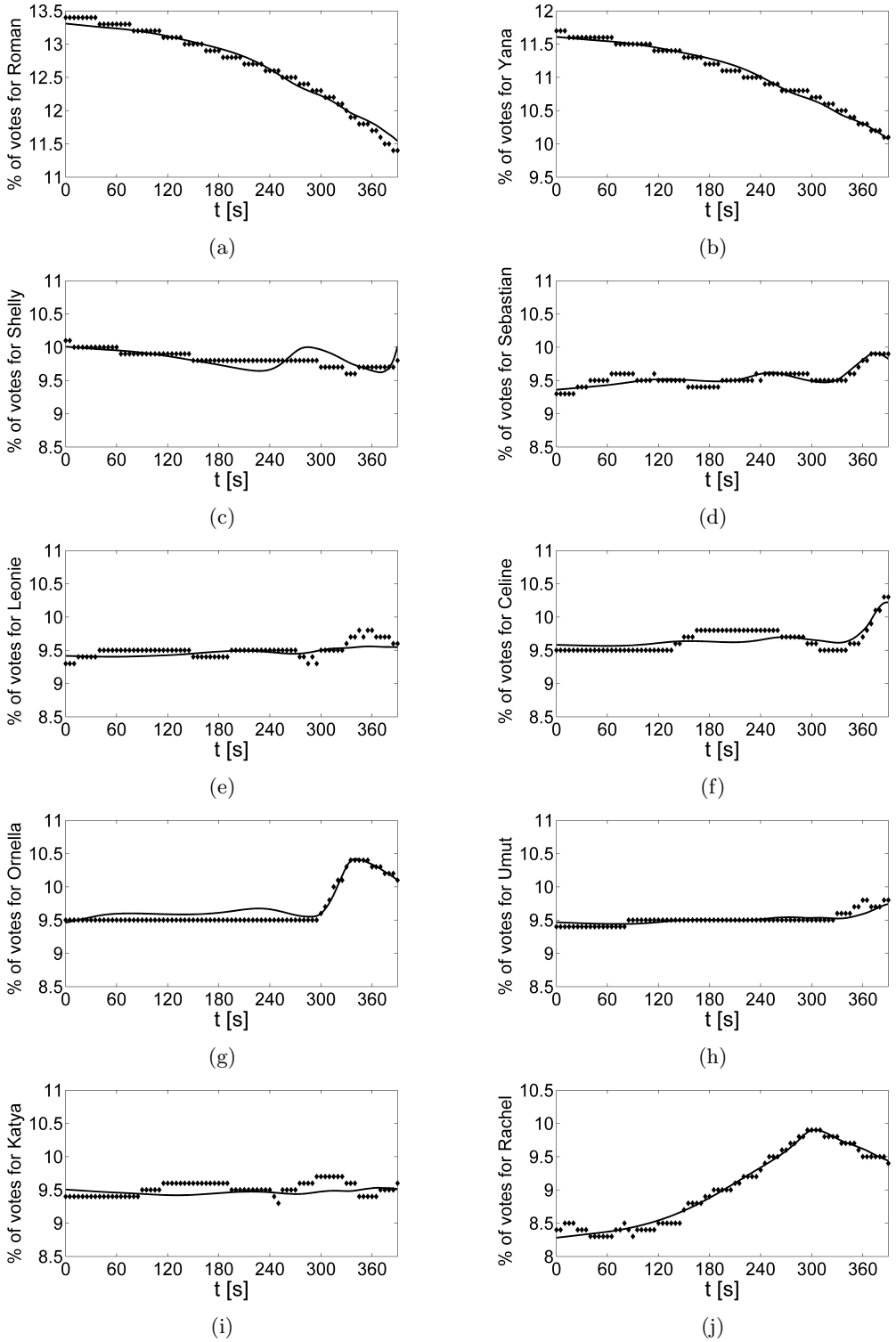


Figure 16.6.: Parameter estimation results for the voting behavior of the viewers of the TV singing competition "Unser Star für Baku": Fit of the components $y_i(t;\hat{c})$ of the HDDE-IVP solution for the estimated values $\hat{c}$ of the parameters (solid lines) to data taken from the show (diamonds). The Figures (a)-(h) display the results for all 10 candidates.

Summary & Outlook

In this concluding chapter, the main contributions of this thesis are summarized and ideas for future research are given.

This thesis has introduced and treated a new class of differential equations called impulsive hybrid discrete-continuous delay differential equations (IHDDEs). These are differential equations with time delays and with discontinuities in the right-hand-side function of the differential equation and/or in the state itself. Delay functions and switching functions have been considered that depend on the state, which implies that the time points of discontinuity in y (and in its time derivatives) are defined implicitly as functions of the state itself. A possible further generalization of the considered problem is to consider dependencies of the right-hand-side function on the time derivative of the state in the past, which is the case in so-called "delay differential equations of neutral type".

In this thesis, initial value problems (IVPs) were formulated for IHDDEs, and a concept of a solution of an IHDDE-IVP was established. The solution concept involved the requirement that there are only finitely many zeros of switching functions, and further that there are only finitely many sign changes of the propagation switching functions. These requirements exclude, for example, Filippov solutions. A proposition for future research is to consider Filippov solutions of IHDDEs, and to develop the corresponding existence, uniqueness, and differentiability theory.

For the specific solution concept used in this thesis, theoretical results on existence of solutions, uniqueness of solutions, and differentiability of solutions with respect to parameters were given. All theorems given in this context made assumptions that allow to transform complicated IVPs into sequences of IVPs in ordinary differential equations (ODEs). The proofs of the theorems were then approached with the method of steps, combined with an exploitation of tailored consistency, distinctness, or regularity assumptions that ensure uniqueness (or differentiability) of the IVP solution in the neighborhood of discontinuity points. This technique could also be used for deriving higher order differentiability results of IHDDE-IVP solutions, but the derivation of the corresponding "higher-order variational IVPs" as well as the formulation of suitable regularity assumptions is a very technical issue.

In this thesis, the modified standard approach was introduced as a novel concept for numerically solving IVPs in differential equations with time delays. The modified standard approach formalizes the idea of using extrapolations beyond discontinuities in the past. It was shown that the numerical solution of a continuous Runge-Kutta method – realized in the framework of the modified standard approach – converges to the exact solution. It should be possible to develop convergence theorems similar to the presented one for other continuous one-step and continuous multi-step methods, which is an interesting subject for future research.

This work has investigated a "first differentiate, then discretize" and a "first discretize, then differentiate" approach for the numerical computation of forward sensitivities in IHDDEs. An analysis of the two approaches at the example of continuous Runge-Kutta methods has revealed that the presence of time delays destroys commutativity of the discretization and differentiation operators as it is known from ordinary differential equation theory. A key role is played by the time derivative of the state in the past. Straightforward differentiation of the continuous representation may lead to a scheme for sensitivity computation that has a lower convergence order than the scheme that is used for solving the nominal IVP. In order to obtain identical convergence orders for the nominal solution and for the computation of sensitivities, different approximations of the time derivative of the state in the past are favorable, for example an evaluation of the right-hand-side function at the past time point.

An extension of the concept of Internal Numerical Differentiation was proposed in this thesis, and practical numerical schemes were presented that realize Internal Numerical Differentiation for IHDDEs. Both a forward scheme and an adjoint scheme were presented. By construction, the developed forward and adjoint schemes "fit exactly" to each other, meaning that the same result is obtained for the sensitivities. A topic for future research in this context is to investigate the relation of the developed discrete adjoint scheme to a "continuous adjoint", i.e. to the solution of

a suitably defined adjoint IHDDE-IVP.

The performance of the newly developed methods for the numerical solution of IVPs in differential equations with time delays and for the computation of sensitivities was analysed. For several challenging IVPs, reference values for the solution and for the sensitivities were given, and the convergence of the results of a variable-stepsize method in the limit of small relative tolerances was studied. Furthermore, it was shown that localization of the time point of a propagated discontinuity is more efficient with the modified standard approach than with the standard approach. Internal Numerical Differentiation was compared to two classical approaches for sensitivity computation. In comparison to finite differences ("External Numerical Differentiation"), Internal Numerical Differentiation yielded more accurate sensitivities (relative error decreased by one order of magnitude), while reducing the computation time by about 80%. In comparison to a solution of the combined system of nominal and variational IVP, Internal Numerical Differentiation provided sensitivities of the same accuracy at only 1% of the computation time. In additional numerical investigations, the forward and adjoint variant of Internal Numerical Differentiation were compared, and the superior efficiency of the developed discrete adjoint scheme for problems with many parameters was demonstrated. Eventually, it was shown on a practical example that accurate computation of sensitivities may require the use of a newly developed error control strategy for sensitivity computation in IHDDEs.

This thesis has further addressed the task of estimating parameters in the model functions of IHDDEs. A single shooting parameterization of IHDDE-constrained least-squares parameter estimation problems was considered, and a damped Generalized Gauss-Newton method for the solution of the resulting finite-dimensional nonlinear constrained least-squares problem was proposed and realized. Parameter estimation problems in differential equations with time delays and switches are non-smooth optimization problems. Nevertheless, the damped Generalized Gauss-Newton method (as an example of a derivative-based optimization method) showed a very good convergence behavior on a test case with artificial measurement data. Furthermore, the developed methods were successfully used for solving two parameter estimation problems in systems with time delays and with real-world experimental data. This demonstrates the suitability of the methods for practical problems. A proposition for future research is to develop – on the basis of the methods proposed in this work – boundary value problems approaches for parameter estimation in IHDDEs. The use of such approaches could be advantageous, e.g. for the treatment of unstable problems.

Among the applications discussed in this thesis is the voting behavior of the viewers of the TV singing competition "Unser Star für Baku", that was broadcast in Germany in 2012. A differential equation model for the voting behavior was developed that incorporated time delays and switches in the right-hand-side function. A good fit of the developed model to data taken from the TV show was obtained, and the analysis of the solution revealed that the time delay in the voting procedure and the "laziness" of the TV viewers have a significant impact on the dynamics observed in the livescore.

Eventually, this thesis contains detailed descriptions of two newly developed software packages, Colsol-DDE and ParamEDE. Colsol-DDE numerically solves IHDDE-IVPs with the modified standard approach and realizes forward and adjoint Internal Numerical Differentiation for computing the sensitivities with respect to parameters. ParamEDE realizes a damped Generalized Gauss-Newton method for solving nonlinear constrained parameter estimation problems in IHDDEs, in which the restrictive monotonicity test is used a globalization strategy. In ParamEDE, Colsol-DDE is used as a building block for solving IVPs and for computing sensitivities.

Acknowledgments

Warum gibst du dieses wahnwitzige Unternehmen nicht einfach auf, Onkel Donald?

When I was young, I was a big fan of "Lustige Taschenbücher" ("Donald Duck pocket books" in English). Of all the stories I read, I still remember one that began with Tick, Trick, and Track (Huey, Dewey, and Louie in the English original) admiring a person called "Thor Donnerkeil". This guy claims, in his books, that he has lived through a huge number of almost unbelievable adventures. In order to earn the respect of his nephews, Donald then embarks on even more fantastic and even more dangerous adventures. As the story goes, Donald fails with each and every single one of his undertakings, and his nephews – worried about Donald's well-being – ask him the above quoted question. In English, it reads: "Why don't you give up on this insane undertaking, Uncle Donald?"

Writing this thesis also felt, occasionally, as an insane undertaking that I might never be able to finish. Now that I have reached the last page of this work, it is time to thank the people that have accompanied and helped me throughout the last years.

At first, I wish to thank *Prof. Dr. Dres. h.c. Hans Georg Bock*, whose inspiring lectures first dragged me into the field of numerical simulation and optimization. He directed my way during the time of this PhD project and has supported my research ever since my first days in his group as a diploma student. For this, I am very grateful.

I am deeply grateful to *Dr. Johannes Schlöder*, who has done an outstanding job as my mentor. His assistance and guidance has been of invaluable importance for this work. Whenever I had questions, he was there to listen to my problems and making suggestions for the next steps in my work. I further cordially thank him for proofreading parts of this thesis.

While I was working on Chapter 5 of this thesis, I originally planned to refer, for the proof of Theorem 5.18, to the convergence result for the standard approach as it is given in Bellen and Zennaro [26]. Eventually, however, I decided that the generalization to discontinuous initial functions, to IHDDEs, and the fact that I was considering extrapolations beyond past discontinuities (what I call the modified standard approach) deserve a formal proof. I then digged deeper and deeper into the theory, but I was finding new questions rather than answers. That was the time when I contacted *Prof. Dr. Alfredo Bellen*, and I am deeply grateful for his help, for the very nice e-mail correspondance, and also for his invitation to the Dobbiaco Summer School 2013. It has been a pleasure to discuss the convergence proof with both him and with *Dr. Stefano Maset* personally on that occasion.

Prof. Dr. Dres. h.c. Hans Georg Bock and *Dr. Johannes Schlöder* have also made an important contribution to this thesis by creating and managing the *Simulation and Optimization* working group, which has been a highly inspiring environment for carrying out this PhD project. Many group members have contributed to this thesis work by creating a friendly atmosphere, baking many delicious cakes for the coffee breaks, and, of course, by having time for both scientific and non-scientific discussions. Specifically, I wish to mention *Dr. Dörte Beigel*, *Kathrin Hatz*, *Christian Hoffmann*, *Robert Kircheis*, *Dr. Mario Mommer*, *Dr. Chaw Pa Pa Oo*, *Andreas Sommer*, and *Leonard Wirsching*.

I am particularly grateful to *Andreas Sommer* for answering all my question on the "crosstalk model", and also for his help in the construction of the model variant that uses time delays. I further give my sincere thanks to *Dr. Dörte Beigel*, *Jürgen Gutekunst*, *Christian Hoffmann*, and *Andreas Sommer* for their efforts in proofreading individual chapters of this thesis.

Thinking about the work group, also some former and associated members come to my mind. Even though they are not "technically" part of the group, they are around frequently enough to make me feel that they actually belong here. I therefore thank *Dr. Tanja Binder*, *Martin Felis*, *Dr. Alexandra Herzog*, and *Dr. Christoph Zimmer* for many discussions and the close contact during the last years.

I am very much obliged to a few special members of the Simulation and Optimization work group, namely to the system administrator *Thomas Kloepfer*, to the long-time secretary *Margret Rothfuss*, and to *Anja Vogel*, who has recently taken over the job as secretary after Margret's retirement. It is thanks to the effort of these three friendly and hard-working people that things are always well-organized and going smoothly in the work group.

This thesis was carried out at the *Faculty of Mathematics and Computer Science* and at the *Interdisciplinary Center for Scientific Computing* (IWR) of *Heidelberg University*. In addition, I have been a fellow of the *Heidelberg Graduate School for Mathematical and Computational Methods for the Sciences (HGS MathComp)*. I have experienced these institutions as an excellent surrounding for my research. Financial support from *Heidelberg University*, from the *SBCancer Network* of the *Helmholtz Alliance on Systems Biology*, and from *HGS MathComp* is gratefully acknowledged.

I am particularly grateful for the funding that I received for presenting my work at several international conferences and workshops at various places around the world. These trips have allowed me to get an overview of the current state-of-the-art in my research field, which has greatly influenced this thesis. I also wish to thank the many fantastic people that I met on these occasions, because they have contributed to the very pleasent atmosphere at the meetings. Therefore, if you ever come to read these lines and remember meeting me in a snake-serving restaurant in Hanoi, in the city center of Dresden, at the archeological sites of Samos, on a boat trip along the coast of Vancouver, at the beach of Palma de Mallorca, or during a mountain hike in South Tyrol, then these thanks go out to you.

During my last year as a PhD student, I also had the honor to be one of the fellows speakers of the *HGS MathComp*. In this context, I want to express my gratitude to *Katharyn Fletcher*, *Dr. Caroline Krauter* and *Rahul Nair* for being such a great team of fellows speakers, and for the joint efforts in the organization of the Annual Colloquium 2012 and of the Fellows Seminars.

I would also like to thank all those people that have been part of the *football group of HGS MathComp*. I always enjoyed playing with you – and against you, except if you are among those who won all their tacklings against me. Special thanks go to *Dr. Volkmar Reinhardt*, *Jens Fangerau*, and *Pavel Hron*, who have organised and coordinated the matches in order to give all of us the weekly chance to turn frustrations about non-functional computer codes into celebrated football goals.

When I compare this PhD project to a challenging adventure, then it belongs to the truth that I would never have been able to start it in the first place without the help of my family, and, in particular, of my parents *Ruth and Manfred Lenz*. I am very grateful for the love and support that I have received ever since my first days until now.

My very special thanks go to *Elmar Korth*, who has accompanied me in my life as a very close friend ever since elementary school. I thank him for the many extensive walks that we shared in the last decades, and for always being there to exchange thoughts on all topics that come to our minds.

It remains to give my biggest thanks of all to my fiancé *Dr. Melanie Schwingel*. I thank her for her love and her invaluable emotional support, especially during the final weeks and months of the writing of this thesis. More than once, she has been my "Anti-Tick-Trick-Track" – believing in me, telling me to keep going, regardless of how insane this undertaking seemed to me.

The final quote of this thesis should thus be one that tells you to keep fighting, also against your own doubts whether your efforts will eventually pay off. I found one in the lyrics of a song of one of my favorite bands, and it goes like this:

Manchmal muss man, um zu siegen,
erst sich selbst im Kampf bezwingen,
seine Schwächen überwinden,
jeden Zweifel niederringen.

Subway to Sally, "Die Schlacht".

Bibliography

[1] J. Albersmeyer. Effiziente Ableitungserzeugung in einem adaptiven BDF–Verfahren. Diploma thesis, Ruprecht–Karls–Universität Heidelberg, 2005.

[2] J. Albersmeyer. *Adjoint-based Algorithms and Numerical Methos for Sensitivity Generation and Optimal Control of Large-Scale Dynamic Systems*. PhD thesis, Ruprecht–Karls–Universität Heidelberg, 2010.

[3] J. Albersmeyer and H.G. Bock. Sensitivity Generation in an Adaptive BDF-Method. In H. G. Bock, E. A. Kostina, X. H. Phu, and R. Rannacher, editors, *Modeling, Simulation and Optimization of Complex Processes: Proceedings of the International Conference on High Performance Scientific Computing, March 6–10, 2006, Hanoi, Vietnam*, pages 15–24. Springer, 2008.

[4] M. Alexe and A. Sandu. Forward and adjoint sensitivity analysis with continuous explicit Runge–Kutta schemes. *Applied Mathematics and Computation*, 208:328–346, 2009.

[5] H. Amann. *Gewöhnliche Differentialgleichungen*. de Gruyter, Berlin, New York, 1995.

[6] A. Anokhin, L. Berezansky, and E. Braverman. Exponential Stability of Linear Delay Impulsive Differential Equations. *Journal of Mathematical Analysis and Applications*, 193:923–941, 1995.

[7] E. Baake and J. P. Schlöder. Modelling the fast fluorescence rise of photosynthesis. *Bulletin of Mathematical Biology*, 54:999–1021, 1992.

[8] V. Bahl and A. A. Linninger. Modeling of Continuous-Discrete Processes. In *Hybrid Systems: Computation and Control*, pages 387–402. Springer, 2001.

[9] C. T. H. Baker, G. A. Bocharov, J. M. Ford, P. M. Lumb, S. J. Norton, C. A. H. Paul, T. Junt, P. Krebs, and B. Ludewig. Computational approaches to parameter estimation and model selection in immunology. *Journal of Computational and Applied Mathematics*, 184:50–76, 2005.

[10] C. T. H. Baker, G. A. Bocharov, and C. A. H. Paul. Mathematical Modelling of the Interleukin-2 T-cell System: A Comparative Study of Approaches Based on Ordinary and Delay Differential Equations. *Journal of Theoretical Medicine*, 2:117–128, 1997.

[11] C. T. H. Baker, G. A. Bocharov, C. A. H. Paul, and F. A. Rihan. Modelling and Analysis of Time-Lags in Cell Proliferation. Technical report, University of Manchester, Manchester, UK, 1997.

[12] C. T. H. Baker, J. C. Butcher, and C. A. H. Paul. Experience of STRIDE applied to delay differential equations. Technical report, University of Manchester, Manchester, UK, 1992.

[13] C. T. H. Baker and C. A. H. Paul. Pitfalls in Parameter Estimation For Delay Differential Equations. *SIAM Journal on Scientific Computing*, 18:305–314, 1997.

[14] C. T. H. Baker and F. A. Rihan. Sensitivity Analysis of Parameters in Modelling With Delay-Differential Equations. Technical report, University of Manchester, Manchester, UK, 1999.

[15] G. Ballinger and X. Liu. Existence, Uniqueness and Boundedness Results for Impulsive Delay Differential Equations. *Applicable Analysis*, 74:71–93, 2000.

[16] H. T. Banks, J. A. Burns, and E. M. Cliff. Parameter Estimation and Identification for Systems with Delays. *SIAM Journal on Control and Optimization*, 19:791–828, 1981.

[17] H. T. Banks and P. K. Daniel Lamm. Estimation of Delays and other Parameters in Nonlinear Functional Differential Equations. *SIAM Journal on Control and Optimization*, 21:895–915, 1983.

[18] H. T. Banks, D. Robbins, and K. Sutton. Theoretical foundations for traditional and generalized sensitivity functions for nonlinear delay differential equations. *Mathematical Biosciences and Engineering*, CRSC-TR12-14:1–34, 2012.

[19] Y. Bard. *Nonlinear Parameter Estimation.* Academic Press, San Diego, 1974.

[20] 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, Ruprecht–Karls–Universität Heidelberg, 1999.

[21] I. Bauer, H. G. Bock, and J. P. Schlöder. DAESOL - a BDF-code for the numerical solution of differential algebraic equations. Technical report, Ruprecht–Karls–Universität Heidelberg, Heidelberg, D, 1999.

[22] S. Becker. Untersuchungen zur Kompatibilitätskonstanten Kappa des verallgemeinerten Gauß-Newton-Verfahrens für allgemeine nichtlineare, beschränkte Ausgleichsprobleme. Diploma thesis, Ruprecht–Karls–Universität Heidelberg, 1999.

[23] D. Beigel. *Efficient goal-oriented global error estimation for BDF-type methods using discrete adjoints.* PhD thesis, Ruprecht–Karls–Universität Heidelberg, 2012.

[24] A. Bellen and N. Guglielmi. Solving neutral delay differential equations with state-dependent delays. *Journal of Computational and Applied Mathematics*, 229:350–362, 2009.

[25] A. Bellen and M. Zennaro. Stability Properties of Interpolants for Runge–Kutta Methods. *SIAM Journal on Numerical Analysis*, 25:411–432, 1988.

[26] A. Bellen and M. Zennaro. *Numerical Methods for Delay Differential Equations.* Oxford Science Publications, Oxford, 2003.

[27] R. Bellman. On the Computational Solution of Differential-Difference Equations. *Journal of Mathematical Analysis and Applications*, 2:108–110, 1961.

[28] R. Bellman and K. L. Cooke. Differential-Difference Equations. Technical report, The RAND Corporation, Santa Monica, US-CA, 1963.

[29] R. E. Bellman, J. D. Buell, and R. E. Kalaba. Numerical Integration of a Differential-Difference Equation With a Decreasing Time-Lag. *Communications of the ACM*, 8:227–228, 1965.

[30] L. T. Biegler. *Nonlinear Programming.* Society for Industrial and Applied Mathematics and the Mathematical Optimization Society, Philadelphia, 2010.

[31] M. Binder, N. Sulaimanov, C. Clausznitzer, M. Schulze, C. M. Hüber, S. M. Lenz, J. P. Schlöder, M. Trippler, R. Bartenschlager, V. Lohmann, and L. Kaderali. Replication Vesicles are Load- and Choke-Points in the Hepatitis C Virus Lifecycle. *PLoS Pathogens*, 9(8):1–21, 2013.

[32] T. Binder. Parameter Estimation for Dynamic Processes Described with DDEs. Diploma thesis, Ruprecht–Karls–Universität Heidelberg, 2007.

[33] G. A. Bocharov and A. A. Romanyukha. Numerical Treatment of the parameter identification problem for delay-differential systems arising in immune response modelling. *Applied Numerical Mathematics*, 15:307–326, 1994.

[34] M. Bocher. Linear Differential Equations with Discontinuous Coefficients. *Annals of Mathematics*, 6:49–63, 1905.

[35] H. G. Bock. Zur numerischen Behandlung zustandsbeschränkter Steuerungsprobleme mit Mehrzielmethode und Homotopieverfahren. *Zeitschrift für Angewandte Mathematik und Mechanik*, 67:T266–T268, 1977.

[36] 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.

[37] H. G. Bock. Numerische Behandlung von zustandsbeschränkten und Chebyshev-Steuerungsproblemen. Technical Report R106/81/11, Carl Cranz Gesellschaft, Heidelberg, 1981.

[38] 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, 1983.

[39] H. G. Bock. *Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen*. PhD thesis, Universität Bonn, 1985.

[40] H. G. Bock, E. A. Kostina, and O. Kostyukova. Covariance Matrices for Parameter Estimates of Costrained Parameter Estimation Problems. *SIAM Journal of Matrix Analysis and Applications*, 29:629–642, 2007.

[41] H. G. Bock, E. A. Kostina, and J. P. Schlöder. On the Role of Natural Level Functions to Achieve Global Convergence for Damped Newton Methods. In M. J. D. Powell and S. Scholtes, editors, *System Modelling and Optimization - Methods, Theory and Applications. 19th IFIP TC7 Conference on System Modelling and Optimization*, pages 51–74. Kluwer, 2000.

[42] H. G. Bock, E. A. Kostina, and J. P. Schlöder. Numerical Methods for Parameter Estimation in Nonlinear Differential Algebraic Equations. *GAMM Mitteilungen*, 30:376–408, 2007.

[43] H. G. Bock and J. P. Schlöder. Numerical Solution of Retarded Differential Equations with Statedependent Time Lags. *Zeitschrift für Angewandte Mathematik und Mechanik*, 61:T269–T271, 1981.

[44] H. G. Bock and J. P. Schlöder. Numerical Computation of Optimal Controls in the Presence of State–Dependent Time Lags. In *Proceedings 9th IFAC World Congress Automatic Control*, pages 108–113. Pergamon Press, 1984.

[45] H. G. Bock, J. P. Schlöder, and V. H. Schulz. Numerik großer Differentiell-Algebraischer Gleichungen - Simulation und Optimierung. In H. Schuler, editor, *Prozeß-Simulation*, chapter 2, pages 35–80. VCH, 1995.

[46] P. Bogacki and L. F. Shampine. A 3(2) Pair of Runge–Kutta Formulas. *Applied Mathematics Letters*, 2:321–325, 1989.

[47] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. *Numerical Optimization*. Springer, Berlin, Heidelberg, New York, 2006.

[48] D. W. Brewer. The differentiability with respect to a parameter of the solution of a linear abstract Cauchy problem. *SIAM Journal on Mathematical Analysis*, 13:607–620, 1982.

[49] I. N. Bronstein, K. A. Semendjajew, G. Musiol, and H. Mühlig. *Taschenbuch der Mathematik*. Verlag Harri Deutsch, Thun, Frankfurt am Main, 2001.

[50] C. G. Broyden. The Convergence of a Class of Double-rank Minimization Algorithms. 2. The New Algorithm. *IMA Journal of Applied Mathematics*, 6:222–231, 1970.

[51] R. Bulirsch. Einführung in die Flugbahnoptimierung Teil II: Die Mehrzielmethode zur numerischen Lösung von nichtlinearen Randwertproblemen und Aufgaben der optimalen Steuerung. Technical report, Carl Cranz Gesellschaft e.V., 1971.

[52] J. A. Burns and P. D. Hirsch. A Difference Equation Approach to Parameter Estimation for Differential-Delay Equations. *Applied Mathematics and Computation*, 7:281–311, 1980.

[53] K. Burrage, J. C. Butcher, and F. H. Chipman. An implementation of singly-implicit Runge–Kutta methods. *BIT*, 20:326–340, 1980.

[54] J. C. Butcher. The adaptation of STRIDE to delay differential equations. *Applied Numerical Mathematics*, 9:415–425, 1992.

[55] J. C. Butcher. *Numerical Methods for Ordinary Differential Equations.* Wiley, Chichester, 2008.

[56] A. Callender, D. R. Hartree, and A. Porter. Time–Lag in a Control System. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 235:415–444, 1936.

[57] M. Calvo, J. I. Montijano, and L. Rández. On the solution of discontinuous IVPs by adaptive Runge–Kutta codes. *Numerical Algorithms*, 33:163–182, 2003.

[58] Y. Cao, S. Li, and L. R. Petzold. Adjoint sensitivity analysis for differential-algebraic equations: algorithms and software. *Journal of Computational and Applied Mathematics*, 149:171–191, 2002.

[59] M. Caracotsios and W. E. Stewart. Sensitivity analysis of initial value problems with mixed ODE's and algebraic equations. Technical report, University of Wisconsin, Madison, US-WI, 1984.

[60] G. R. Carmichael, A. Sandu, and F. A. Potra. Sensitivity analysis for atmospheric chemistry models via automatic differentiation. *Atmospheric Environment*, 31:475–489, 1997.

[61] F. E. Cellier. *Combined Continuous/Discrete System Simulation by Use of Digital Computers: Techniques and Tools.* PhD thesis, ETH Zürich, 1979.

[62] M.-P. Chen, J. S. Yu, and J. H. Shen. The persistence of nonoscillatory solutions of delay differential equations under impulsive perturbations. *Computers and Mathematics with Applications*, 27:1–6, 1994.

[63] Y. Chen, Q. Hu, and J. Wu. Second-order differentiability with respect to parameters for differential equations with adaptive delays. *Frontiers of Mathematics in China*, 5:221–286, 2010.

[64] N. H. Choksy. Time Lag Systems – A Bibliography. *IRE Transactions on Automatic Control*, 1:66–70, 1960.

[65] L. O. Chua and L. Yang. Cellular Neural Networks: Applications. *IEEE Transactions on Circuits and Systems*, 35:1273–1290, 1988.

[66] L. O. Chua and L. Yang. Cellular Neural Networks: Theory. *IEEE Transactions on Circuits and Systems*, 35:1257–1272, 1988.

[67] M. D. Compere. *Simulation of Engineering Systems Described by High-Index DAE and Discontinuous ODE Using Single Step Methods.* PhD thesis, University of Texas at Austin, 2001.

[68] K. L. Cooke and P. van den Driessche. Analysis of an SEIRS epidemic model with two delays. *Journal of Mathematical Biology*, 35:240–260, 1996.

[69] J. Cortés. Discontinuous Dynamical Systems. *IEEE Control Systems Magazine*, 28:36–73, 2008.

[70] S. P. Corwin, D. Sarafyan, and S. Thompson. DKLAG6: A Code Based on Continuously Imbedded Sixth Order Runge–Kutta Methods for the Solution of State Dependent Functional Differential Equations. *Applied Numerical Mathematics*, 24:319–330, 1997.

[71] S. P. Corwin, S. Thompson, and S. M. White. Solving ODEs and DDEs with Impulses. *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 3:139–149, 2008.

[72] C. W. Cryer. Numerical Methods for Functional Differential Equations. In *Delay and Functional Differential Equations and Their Applications*, pages 17–101. Academic Press, 1972.

[73] P. C. Das and R. R. Sharma. On optimal controls for measure delay–differential equations. *SIAM Journal on Control*, 9:43–61, 1971.

[74] W. C. Davidon. Variable metric method for minimization. *SIAM Journal on Optimization*, 1:1–17, 1991.

[75] J. E. Dennis Jr., D. M. Gay, and R. E. Welsch. An Adaptive Nonlinear Least-Squares Algorithm. *ACM Transactions on Mathematical Software*, 7:348–368, 1981.

[76] P. Deuflhard and F. Bornemann. *Numerische Mathematik 2: Gewöhnliche Differentialgleichungen*. de Gruyter, Berlin, New York, 2008.

[77] R. P. Dickinson and R. J. Gelinas. Sensitivity Analysis of Ordinary Differential Equations – A Direct Method. *Journal of Computational Physics*, 21:123–143, 1976.

[78] L. Dieci and L. Lopez. Sliding motion in Filippov differential systems: theoretical results and a computational approach. *SIAM Journal on Numerical Analysis*, 47:2023–2051, 2009.

[79] L. Dieci and L. Lopez. Numerical solution of discontinuous differential systems: Approaching the discontinuity surface from one side. *Applied Numerical Mathematics*, 67:98–110, 2013.

[80] J. R. Dormand and P. J. Prince. A family of embedded Runge–Kutta formulae. *Journal of Computational and Applied Mathematics*, 6:19–26, 1980.

[81] E. P. Dougherty, J. Hwang, and H. Rabitz. Further developments and applications of the Greens function method of sensitivity analysis in chemical kinetics. *The Journal of Chemical Physics*, 71:1794–1808, 1979.

[82] R. D. Driver. *Ordinary and Delay Differential Equations*. Springer, New York, Heidelberg, Berlin, 1977.

[83] A. M. Dunker. The decoupled direct method for calculating sensitivity coefficients in chemical kinetics. *The Journal of Chemical Physics*, 81:2385–2393, 1984.

[84] P. Eberhard and C. Bischof. Automatic differentiation of numerical integration algorithms. *Mathematics of Computation*, 68:717–731, 1999.

[85] M. Egmont-Petersen, D. de Ridder, and H. Handels. Image processing with neural networks – a review. *Pattern Recognition*, 35:2279–2301, 2002.

[86] E. Eich-Soellner and C. Führer. *Numerical Methods in Multibody Dynamics*. B.G. Teubner, Stuttgart, 1998.

[87] D. Ellison. Efficient Automatic Integration of Ordinary Differential Equations with Discontinuities. *Mathematics and Computers in Simulation*, XXIII:12–20, 1981.

[88] L. M. Ellwein, H. T. Tran, C. Zapata, V. Novak, and M. S. Olufsen. Sensitivity Analysis and Model Assessment: Mathematical Models for Arterial Blood Flow and Blood Pressure. *Cardiovascular Engineering*, 8:94–108, 2008.

[89] M. B. Elowitz and S. Leibler. A synthetic oscillatory network of transcriptional regulators. *Nature*, 403:335–338, 2000.

[90] L. E. El'sgol'ts. *Differential Equations*. Hindustan Publishing Corporation, Delhi, 1961.

[91] L. E. El'sgol'ts. *Qualitative Methods in Mathematical Analysis*. Americal Mathematical Society, Providence, 1964.

[92] L. E. El'sgol'ts and S. B. Norkin. *Introduction to the Theory and Application of Differential Equations with Deviating Arguments*. Academic Press, New York, 1973. translated by J. L. Casti.

[93] K. Engelborghs, T. Luzyanina, and D. Roose. Numerical Bifurcation Analysis of Delay Differential Equations Using DDE-BIFTOOL. *ACM Transactions on Mathematical Software*, 28:1–21, 2002.

[94] W. H. Enright. A new error-control for initial value solvers. *Applied Mathematics and Computation*, 31:288–301, 1989.

[95] W. H. Enright. Continuous numerical methods for ODEs with defect control. *Journal of Computational and Applied Mathematics*, 125:159–170, 2000.

[96] W. H. Enright and H. Hayashi. A delay differential equation solver based on a continuous Runge–Kutta method with defect control. *Numerical Algorithms*, 16:349–364, 1997.

[97] W. H. Enright, T. E. Hull, and B. Lindberg. Comparing numerical methods for stiff systems of ODE's. *BIT*, 15:10–48, 1975.

[98] W. H. Enright, K. R. Jackson, S. Nørsett, and P. G. Thomson. Effective Solution of Discontinuous IVPs Using a Runge–Kutta Formula Pair with Interpolants. *Applied Mathematics and Computation*, 27:313–335, 1988.

[99] W. H. Enright, K. R. Jackson, S. P. Nørsett, and P. G. Thomson. Interpolants for Runge–Kutta Formulas. *ACM Transactions on Mathematical Software*, 12:193–218, 1986.

[100] W. H. Enright and L. Yan. The reliability/cost trade-off for a class of ODE solvers. *Numerical Algorithms*, 53:239–260, 2010.

[101] M. Ernst. Lösung von Anfangswertproblemen bei verzögerten Differentialgleichungen mit SolvIND. Bachelor thesis, Ruprecht–Karls–Universität Heidelberg, 2011.

[102] W. F. Feehery, J. E. Tolsma, and P. I. Barton. Efficient sensitivity analysis of large–scale differential–algebraic systems. *Applied Numerical Mathematics*, 25:41–54, 1997.

[103] A. Feldstein and K. W. Neves. High Order Methods for State-Dependent Delay Differential Equations with Nonsmooth Solutions. *SIAM Journal on Numerical Analysis*, 21:844–863, 1984.

[104] A. F. Filippov. Differential Equations with Discontinuous Right Hand Side. *Matematicheskii sbornik*, 93:99–128, 1960.

[105] A. F. Filippov. *Differential Equations with Discontinuous Right Hand Side*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1988.

[106] W. B. Fite. Properties of the solutions of certain functional differential equations. *Transactions of the American Mathematical Society*, 22:311–319, 1921.

[107] R. Fletcher. A new approach to variable metric algorithms. *The Computer Journal*, 13:317–322, 1970.

[108] R. Fletcher. *Practical Methods of Optimization*. John Wiley & Sons, Chichester, 1999.

[109] R. Fletcher and M. J. D. Powell. A rapidly convergent descent method for minimization. *The Computer Journal*, 6:163–168, 1963.

[110] O. Forster. *Analysis II*. Vieweg, Wiesbaden, 2006.

[111] S. Galán, W. F. Feehery, and P. I. Barton. Parametric sensitivity functions for hybrid discrete/continuous systems. *Applied Numerical Mathematics*, 31:17–47, 1999.

[112] C. W. Gear and O. Østerby. Solving Ordinary Differential Equations with Discontinuities. *ACM Transactions on Mathematical Software*, 10:23–44, 1984.

[113] C. W. Gear and T. Vu. Smooth numerical solutions of ordinary differential equations. In P. Deuflhard and E. Hairer, editors, *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, pages 2–12. Birkhäuser, 1983.

[114] C. Geiger and C. Kanzow. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer, Berlin, Heidelberg, New York, 2002.

[115] L. Genik and P. van den Driessche. An Epidemic Model with Recruitment-Death Demographics and Discrete Delays. In *Differential Equations with Applications to Biology*, volume 21 of *Fields Institute Communications*, pages 237–249. Americal Mathematical Society and The Fields Institute, 1999.

[116] L. Glass and M. C. Mackey. Mackey-Glass equation. *Scholarpedia*, 5:6908, 2010.

[117] D. Goldfarb. A Family of Variable-Metric Methods Derived by Variational Means. *Mathematics of Computation*, 24:23–26, 1970.

[118] K. Gopalsamy and B. G. Zhang. On Delay Differential Equations with Impulses. *Journal of Mathematical Analysis and Applications*, 139:110–122, 1989.

[119] A. Griewank. On Automatic Differentiation. Technical report, Argonne National Laboratory, Argonne, US-IL, 1988.

[120] A. Griewank and A. Walther. *Evaluating derivatives: principles and techniques of algorithmic differentiation*. SIAM, Philadelphia, 2008.

[121] T. H. Gronwall. Note on the Derivatives with Respect to a Parameter of the Solutions of a System of Differential Equations. *The Annals of Mathematics*, 20:292–296, 1919.

[122] N. Guglielmi and E. Hairer. Implementing Radau IIA Methods for Stiff Delay Differential Equations. *Computing*, 67:1–12, 2001.

[123] N. Guglielmi and E. Hairer. Users Guide for the Code RADAR5 – Version 2.1. Technical report, Università dellAquila, LAquila, I, 2005.

[124] N. Guglielmi and E. Hairer. Computing breaking points in implicit delay differential equations. *Advances in Computational Mathematics*, 29:229–247, 2008.

[125] B. Günterberg. Thereotische und Numerische Behandlung von Problemen der Optimalen Steuerung bei Retardierten Systemen. Diploma thesis, Universität Bonn, 1989.

[126] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I*. Springer-Verlag, Berlin, Heidelberg, 2008.

[127] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Springer, Berlin, Heidelberg, New York, 2002.

[128] E. Hairer, G. Wanner, and C. Lubich. *Geometric Numerical Integration*. Springer, Berlin, Heidelberg, 2010.

[129] O. Hájek. Discontinuous Differential Equations, I. *Journal of Differential Equations*, 32:149–170, 1979.

[130] J. Hale and S. V. Verduyn Lunel. *Introduction to Functional Differential Equations*. Springer, New York, Berlin, Heidelberg, 1993.

[131] J. K. Hale and L. A. C. Ladeira. Differentiability with Respect to Delays. *Journal of Differential Equations*, 92:14–26, 1991.

[132] P. Hartman. *Ordinary Differential Equations*. SIAM Classics in Applied Mathematics, Philadelphia, 2002.

[133] F. Hartung. On differentiability of solutions with respect to parameters in a class of functional differential equations. *Functional Differential Equations*, 4:65–79, 1997.

[134] F. Hartung. Parameter Estimation by Quasilinearization in Functional Differential Equations with State-Dependent Delays: a Numerical Study. *Nonlinear Analysis*, 47:4557–4566, 2001.

[135] F. Hartung. Differentiability of Solutions with Respect to the Initial Data in Differential Equations with State-dependent Delays. *Journal of Dynamics and Differential Equations*, 23:843–884, 2011.

[136] F. Hartung. Parameter estimation by quasilinearization in differential equations with state-dependent delays. *Discrete–Continuous Dynamical Systems Series B*, 18:1611–1631, 2013.

[137] F. Hartung, T. Krisztin, H. O. Walther, and J. Wu. Functional Differential Equations with State–Dependent Delays: Theory and Application. In A. Cañada, P. Drábek, and A. Fonda, editors, *Handbook of Differential Equations*, chapter 5, pages 435–545. Elsevier B.V., 2006.

[138] F. Hartung and J. Turi. Identification of parameters in delay equations with state-dependent delays. *Nonlinear Analysis: Theory, Methods, and Applications*, 29:1303–1318, 1997.

[139] F. Hartung and J. Turi. On Differentiability of Solutions with respect to Parameters in State–Dependent Delay Equations. *Journal of Differential Equations*, 135:192–237, 1997.

[140] L. Hascoët and V. Pascual. TAPENADE 2.1. User's Guide. Technical report, Institut National de Récherche en Informatique et en Automatique (INRIA), Sophia Antipolis, F, 2004.

[141] L. Hascoët and V. Pascual. The Tapenade Automatic Differentiation tool: principles, model, and specification. Technical report, Institut national de recherche en informatique et en automatique (INRIA), Sophia Antipolis, F, 2012.

[142] J. D. Hass. Automatische Regularisierung beim Gauß-Newton-Verfahren für Parameterschätzprobleme über eine Schätzung der Inkompatibilitätskonstanten Kappa. Diploma thesis, Ruprecht–Karls–Universität Heidelberg, 2012.

[143] J. L. Hay, R. E. Crosbie, and R. I. Chaplin. Integration routines for systems with discontinuities. *The Computer Journal*, 17:275–278, 1974.

[144] H. Hayashi. *Numerical Solution of Retarded and Neutral Delay Differential Equations using Continuous Runge–Kutta Methods*. PhD thesis, University of Toronto, 1996.

[145] H. W. Hethcote, M. A. Lewis, and P. van den Driessche. An epidemiological model with a delay and a nonlinear incidence rate. *Journal of Mathematical Biology*, 27:49–64, 1989.

[146] E. Hilb. Zur Theorie der linearen funktionalen Differentialgleichungen. *Mathematische Annalen*, 78:137–170, 1917.

[147] A. C. Hindmarsh and R. Serban. User Documentation for CVODES v2.7.0. Technical report, Lawrence Livermore National Laboratory, Livermore, US-CA, 2012.

[148] A. L. Hodgin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117:500–544, 1952.

[149] C. Hoffmann. *Numerical and Statistical Aspects of Uncertainty in the Design of Optimal Experiments for Model Discrimination and Parameter Estimation*. PhD thesis, Ruprecht–Karls–Universität Heidelberg, 2014. In preparation.

[150] J. N. Holt and R. Fletcher. An Algorithm for Constrained Non-Linear Least-Squares. *IMA Journal of Applied Mathematics*, 23:449–463, 1979.

[151] W. Horbelt. *Maximum likelihood estimation in dynamical systems*. PhD thesis, Albert-Ludwigs-Universität Freiburg, 2001.

[152] W. Horbelt, J. Timmer, and H. U. Voss. Parameter Estimation in nonlinear delayed feedback systems from noisy data. *Physics Letters A*, 299:513–521, 2002.

[153] J. Hwang, E. P. Dougherty, S. Rabitz, and H. Rabitz. The Greens function method of sensitivity analysis in chemical kinetics. *The Journal of Chemical Physics*, 69:5180–5191, 1978.

[154] Z. Jackiewicz. Existence and Uniqueness of Solutions of Neutral Delay-Differential Equations with State Dependent Delays. *Funkcialaj Ekvacioj*, 30:9–17, 1987.

[155] A. Karoui and R. Vaillancourt. Computer Solutions of State-Dependent Delay Differential Equations. *Computers and Mathematics with Applications*, 27:37–51, 1994.

[156] A. Karoui and R. Vaillancourt. A numerical method for vanishing-lag delay differential equations. *Applied Numerical Mathematics*, 17:383–395, 1995.

[157] G. Kedem. Automatic Differentiation of Computer Programs. *ACM Transactions on Mathematical Software*, 6:150–165, 1980.

[158] M. Kiehl. Sensitivity analysis of ODEs and DAEs - theory and implementation guide. *Optimization Methods and Software*, 10:803–821, 1999.

[159] S. Kim, S. A. Campbell, and X. Liu. Stability of a Class of Linear Switching Systems with Delay. *IEEE Transactions on Circuits and Systems*, 53:384–393, 2006.

[160] C. Kirches. A Numerical Method for Nonlinear Robust Optimal Control with Implicit Discontinuities and an Application to Powertrain Oscillations. Diploma thesis, Ruprecht–Karls–Universität Heidelberg, 2006.

[161] C. Kirches, L. Wirsching, H. G. Bock, and J. P. Schlöder. Efficient direct multiple shooting for nonlinear model predictive control on long horizons. *Journal of Process Control*, 22:540–550, 2012.

[162] M. Koda. Sensitivity Analysis of Time-Delay Systems. *International Journal of Systems Science*, 12:1389–1397, 1981.

[163] V. Kolmanovskii and A. D. Myshkis. *Introduction to the Theory and Applications of Functional Differential Equations*. Kluwer Academic Publishers, Dordrecht, Boston, London, 1999.

[164] S. Körkel. *Numerische Methoden für Optimale Versuchsplanungsprobleme bei nichtlinearen DAE-Modellen*. PhD thesis, Universität Heidelberg, 2002.

[165] M. A. Kramer and J. M. Calo. An improved computational method for sensitivity analysis: Green's function method with 'AIM'. *Applied Mathematical Modelling*, 5:432–441, 1981.

[166] S. V. Krishna and A. V. Anokhin. Delay Differential Systems with Discontinuous Initial Data and Existence and Uniqueness Theorems for Systems with Impulse and Delay. *Journal of Applied Mathematics and Stochastic Analysis*, 7:49–67, 1994.

[167] Y. Kuang. *Delay Differential Equations with Application in Population Dynamics*. Academic Press, London, 1993.

[168] M. Kuhn. Solution of Nonlinear Optimization Problems constrained by Delay Differential Equations with an Application to a Gene Regulatory Network. Master's thesis, Ruprecht–Karls–Universität Heidelberg, 2013.

[169] V. Lakshmikantham, D. D Bainov, and P. S. Simeonov. *Theory of Impulsive Differential Equations*. World Scientific, Singapore, Teaneck, London, 1989.

[170] F. Lehn, B. Tibken, and E. P. Hofer. Biomathematical Models with State–Dependent Delays for Granulocytopoesis. In M. Grötschel, S. O. Krumke, and J. Rambau, editors, *Online Optimization of Large Scale Systems*, pages 433–453. Springer, 2001.

[171] S. M. Lenz. Solution of Parameter Estimation Problems with the Direct Multiple Shooting Method in Discontinuous Models. Diploma thesis, Ruprecht–Karls–Universität Heidelberg, 2008.

[172] S. M. Lenz, H. G. Bock, J. P. Schlöder, E. A. Kostina, G. Gienger, and G. Ziegler. Multiple Shooting Method for Initial Satellite Orbit Determination. *AIAA Journal of Guidance, Control, and Dynamics*, 33:1334–1346, 2010.

[173] S. M. Lenz, J. P. Schlöder, and H. G. Bock. Numerical Computation of Derivatives in Systems of Delay Differential Equations. *Mathematics and Computers in Simulation*, 96:124–156, 2014.

[174] K. Levenberg. A method for the solution of certain problems in least squares. *Quaterly of Applied Mathematics*, 2:164–168, 1944.

[175] C. Li, F. Ma, and G. Feng. Hybrid impulsive and switching time-delay systems. *IET Control Theory and Applications*, 3:1487–1498, 2009.

[176] S. Li and L. R. Petzold. Design of New DASPK for Sensitivity Analysis. Technical report, University of California, Santa Barbara, US-CA, 1999.

[177] S. Li and L. R. Petzold. Description of DASPKADJOINT: An Adjoint Sensitivity Solver for Differential-Algebraic Equations. Technical report, University of California, Santa Barbara, US-CA, 2002.

[178] E. Lindelöf. Sur l'application des méthodes d'approximations successives à l'étude des intégrales réelles des équations différentielles ordinaires. *Journal de Mathématiques Pures et Appliquées*, 10:117–128, 1894.

[179] J. Liu, X. Liu, and W.-C. Xie. Input-to-state stability of impulsive and switching hybrid systems with time-delay. *Automatica*, 47:499–908, 2011.

[180] X. Liu, X. Shen, and Y. Zhang. Stability Analysis of a Class of Hybrid Dynamic Systems. *Dynamics of Continuous, Discrete and Impulsive Systems Series B: Applications & Algorithms*, 8:359–373, 2001.

[181] T. Luzyanina, K. Engelborghs, and D. Roose. Numerical Bifurcation Analysis of Differential Equations with State-Dependent Delay. *International Journal of Bifurcation and Chaos*, 11:737–753, 2001.

[182] M. C. Mackey and L. Glass. Oscillation and Chaos in Physiological Control Systems. *Science*, 197:287–289, 1977.

[183] T. Maly and L. R. Petzold. Numerical methods and software for sensitivity analysis of differential-algebraic systems. *Applied Numerical Mathematics*, 20:57–79, 1996.

[184] G. Mao and L. R. Petzold. Efficient Integration over Discontinuities for Differential–Algebraic Systems. *Computers and Mathematics with Applications*, 43:65–79, 2002.

[185] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the Society for Industrial and Applied Mathematics*, 11:431—441, 1963.

[186] MATLAB. *version 2012b*. The MathWorks Inc., Natick, MA, USA, 2012.

[187] S. Mehrkanoon, S. Mehrkanoon, and J. A. K. Suykens. Parameter estimation of delay differential equations: An integration-free LS-SVM approach. *Communications in Nonlinear Science and Numerical Simulation*, 19:830–841, 2014.

[188] J. P. Meijaard. Efficient Numerical Integration of the Equations of Motion of Non-Smooth Mechanical Systems. *Zeitschrift für Angewandte Mathematik und Mechanik*, 77:419–427, 1997.

[189] E. Meißner. Resonanz bei konstanter Dämpfung. *Zeitschrift für Angewandte Mathematik und Mechanik*, 15:62–70, 1935.

[190] B. K. Mishra and D. K. Saini. SEIRS epidemic model with delay for transmission of malicious objects in computer network. *Applied Mathematics and Computation*, 188:1476–1482, 2007.

[191] K. A. Murphy. Estimation of Time- and State-Dependent Delays And Other Parameters in Functional Differential Equations. *SIAM Journal on Applied Mathematics*, 50:972–1000, 1990.

[192] J. A. Nelder and R. Mead. A simplex method for function minimization. *The Computer Journal*, 7:308–313, 1965.

[193] K. W. Neves. Automatic Integration of Functional Differential Equations: An Approach. *ACM Transactions on Mathematical Software*, 1:375–368, 1975.

[194] T. Nguyen-Ba, H. Yagoub, Y. Li, and R. Vaillancourt. Variable-step variable-order 3-stage Hermite–Birkhoff ODE solver of order 5 to 15. *Canadian Applied Mathematics Quaterly*, 14:413–437, 2006.

[195] J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer, New York, 2006.

[196] H. J. Oberle and H. J. Pesch. Numerical Treatment of Delay Differential Equations by Hermite Interpolation. *Numerische Mathematik*, 37:235–255, 1981.

[197] M. S. Olufsen and J. T. Ottesen. A practical approach to parameter estimation applied to model predicting heart rate regulation. *Journal of Mathematical Biology*, 67:39–68, 2013.

[198] J. Oppelstrup. The RKFHB4 Method for Delay Differential Equations. In R. Bulirsch, R. Grigorieff, and J. Schröder, editors, *Numerical Treatment of Differential Equations*, volume 631 of *Lecture Notes in Mathematics*, pages 133–146. Springer, 1978.

[199] G. Orosz, J. Moehlis, and R. M. Murray. Controlling biological networks by time-delayed signals. *Philosophical Transactions of the Royal Society A*, 368:439–454, 2010.

[200] T. Park and P. I. Barton. State Event Location in Differential-Algebraic Models. *ACM Transactions on Modeling and Computer Simulation*, 6:137–165, 1996.

[201] C. A. H. Paul. Developing a delay differential equation solver. *Applied Numerical Mathematics*, 9:403–414, 1992.

[202] C. A. H. Paul. A Test Set of Functional Differential Equations. Technical report, University of Manchester, Manchester, UK, 1994.

[203] C. A. H. Paul. *A User Guide to Archi - an Explicit Runge–Kutta Code for Solving Delay and Neutral Differential Equations and Parameter Estimation Problems*. Manchester Centre for Computational Mathematics, Department of Mathematics, University of Manchester, Manchester, England, 1997.

[204] C. A. H. Paul. Designing efficient software for solving delay differential equations. *Journal of Computational and Applied Mathematics*, 125:287–295, 2000.

[205] T. Pavlidis. Stability of Systems Described by Differential Equations Containing Impulses. *IEEE Transactions on Automatic Control*, 12:43–45, 1967.

[206] T. Pavlidis and E. I. Jury. Analysis of a New Class of Pulse-Frequency Modulated Feedback Systems. *IEEE Transactions on Automatic Control*, 10:35–43, 1965.

[207] G. Peano. Démonstration de l'intégrabilité des équations différentielles ordinaires. *Mathematische Annalen*, 37:182–228, 1890.

[208] L. R. Petzold and U. M. Ascher. *Computer methods for ordinary differential equations and differential-algebraic equations*. SIAM, Philadelphia, 1998.

[209] C. Picard. Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives. *Journal de Mathématiques Pures et Appliquées*, 6:145–210, 1890.

[210] P. T. Piiroinen and Y. A. Kuznetsov. An Event-Driven Method to Simulate Filippov Systems with Accurate Computing of Sliding Motions. *ACM Transactions on Mathematical Software*, 34:13:1–13:24, 2008.

[211] H. Rabitz, M. A. Kramer, and D. Dacol. Sensitivity analysis in chemical kinetics. *Annual Review of Physical Chemistry*, 34:419–461, 1983.

[212] I. Reinecke. *Mathematical Modeling and Simulation of the Female Menstrual Cycle*. PhD thesis, Freie Universität Berlin, 2008.

[213] E. C. Rhodes. Population Mathematics. *Journal of the Royal Statistical Society*, 103:61–89, 1940.

[214] F. A. Rihan. Adjoint Sensitivity Analysis of Neutral Delay Differential Models. *Journal of Numerical Analysis, Industrial and Applied Mathematics*, 5:95–101, 2010.

[215] L. M. Rios and N. V. Sahinidis. Derivative-free optimization: a review of algorithms and comparison of software implementations. *Journal on Global Optimization*, 56:1247–1293, 2013.

[216] D. Robbins. *Sensitivity Functions for Delay Differential Equation Models*. PhD thesis, North Carolina State University, 2011.

[217] J. J. Robertson and J. Williams. Some Properties of Algorithms for Stiff Differential Equations. *IMA Journal of Applied Mathematics*, 16:23–34, 1975.

[218] G. Röst, S. Y. Huang, and L. Székely. On a SEIR epidemic model with delay. *Dynamic Systems and Applications*, 21:33–48, 2012.

[219] A. Sandu, D. N. Daescu, and G. R. Carmichael. Direct and adjoint sensitivity analysis of chemical kinetic systems with KPP: Part I – theory and software tools. *Atmospheric Environment*, 37:5083–5096, 2003.

[220] A. Sandu and P. Miehe. Forward, tangent linear, and adjoint Runge–Kutta methods for stiff chemical kinetic simulations. *International Journal of Computer Mathematics*, 87:2458–2479, 2010.

[221] T. Schlegl, M. Buss, and G. Schmidt. Development of Numerical Integration Methods for Hybrid (Discrete–Continuous) Dynamical Systems. In *Proceedings of Advanced Intelligent Mechatronics*, 1997.

[222] J.P. Schlöder. *Numerische Methoden zur Behandlung hochdimensionaler Aufgaben der Parameteridentifizierung*. PhD thesis, Universität Bonn, 1987.

[223] J.P. Schlöder and H.G. Bock. Identification of Rate Constants in Bistable Chemical Reactions. In P. Deuflhard and E. Hairer, editors, *Numerical Treatment of Inverse Problems in Differential and Integral Equations, Progress in Scientific Computing*, pages 27–47. Birkhäuser, 1983.

[224] W. W. Schmaedeke. Optimal control theory for nonlinear vector differential equations containing measures. *SIAM Journal on Control*, 3:231–280, 1965.

[225] E. Schmidt. Über eine Klasse linearer funktionaler Differentialgleichungen. *Mathematische Annalen*, 70:499–524, 1911.

[226] J. T. Schnute, A. Couture-Beil, and R. Haigh. A Users Guide to the R Package PBSddesolve. Technical report, 2013. Version 1.10.

[227] J. Schumann-Bischoff, S. Luther, and U. Parlitz. Nonlinear system identification employing automatic differentiation. *Communications in Nonlinear Science and Numerical Simulation*, 18:2733–2742, 2013.

[228] F. Schürer. Zur Theorie des Balancierens. *Mathematische Nachrichten*, 1:295–331, 1948.

[229] L. F. Shampine. Some practical Runge–Kutta Formulas. *Mathematics of Computation*, 46:135–150, 1986.

[230] L. F. Shampine. Solving ODEs and DDEs with residual control. *Applied Numerical Mathematics*, 52:113–127, 2005.

[231] L. F. Shampine, I. Gladwell, and R. W. Brankin. Reliable Solution of Special Event Location Problems for ODEs. *ACM Transactions on Mathematical Software*, 17:11–25, 1991.

[232] L. F. Shampine, I. Gladwell, and S. Thompson. *Solving ODEs with Matlab*. Cambridge University Press, Cambridge, 2003.

[233] L. F. Shampine and S. Thompson. Solving DDEs in MATLAB. *Applied Numerical Mathematics*, 37:441–458, 2001.

[234] D. F. Shanno. Conditioning of Quasi-Newton Methods for Function Minimization. *Mathematics of Computation*, 24:647–656, 1970.

[235] J. Sieber, P. Kowalczyk, S. J. Hogan, and M. Di Bernado. Dynamics of Symmetric Dynamical Systems with Delayed Switching. *Journal of Vibration and Control*, 16:1111–1140, 2010.

[236] J. Sieber and B. Krauskopf. Complex balancing motions of an inverted pendulum subject to delayed feedback control. *Physica D*, 197:332–345, 2004.

[237] R. M. Sievert. Zur theoretisch-mathematischen Behandlung des Problems der biologischen Strahlenwirkung. *Acta Radiologica (old series)*, 22:237–251, 1941.

[238] D. J. W. Simpson, R. Kuske, and Y.-X. Li. Dynamics of Simple Balancing Models with Time-Delayed Switching Feedback Control. *Journal of Nonlinear Science*, 22:135–167, 2012.

[239] H. Smith. *An Introduction to Delay Differential Equations with Applications to the Life Sciences*. Springer, New York, Heidelberg, Dordrecht, 2010.

[240] A. Sommer, S. Depner, M. S. Mommer, H. G. Bock, J. P. Schlöder, and M. M. Mueller. Crosstalk effects in cytokine signaling: how important are they? The case of IL–6 and GM–CSF. *International Journal of Biomathematics and Biostatistics*, 2014. accepted.

[241] J. Stoer and R. Bulirsch. *Numerische Mathematik 2*. Springer, Berlin, Heidelberg, New York, 2005.

[242] S. Støren and T. Hertzberg. Obtaining sensitivity information in dynamic optimization problems solved by the sequential approach. *Computers and Chemical Engineering*, 23:807–819, 1999.

[243] K. Strehmel and R. Weiner. Behandlung steifer Anfangswertprobleme gewöhnlicher Differentialgleichungen mit adaptiven Runge–Kutta–Methoden. *Computing*, 29:153–165, 1982.

[244] Y. Takeuchi, W. Ma, and E. Beretta. Global asymptotic properties of a delay SIR epidemic model with finite incubation times. *Nonlinear Analysis*, 42:931–947, 2000.

[245] M. L. Taylor and T. W. Carr. An SIR epidemic model with partial temporary immunity modeled with delay. *Journal of Mathematical Biology*, 59:841–880, 2009.

[246] S. Thompson and L. F. Shampine. A friendly Fortran DDE solver. *Applied Numerical Mathematics*, 56:503–516, 2006.

[247] I.-B. Tjoa and L. T. Biegler. Simultaneous Solution and Optimization Strategies for Parameter Estimation of Differential-Algebraic Equation Systems. *Industrial & Engineering Chemistry Research*, 30:376–385, 1991.

[248] M. Tolan. *Manchmal gewinnt der Bessere: Die Physik des Fußballspiels*. Piper Taschenbuch, München, Zürich, 2011.

[249] J. E. Tolsma and P. I. Barton. Hidden Discontinuities and Parametric Sensitivity Calculations. *SIAM Journal on Scientific Computing*, 23:1861–1874, 2002.

[250] T. Turányi. Sensitivity analysis of complex kinetic systems. Tools and applications. *Journal of Mathematical Chemistry*, 5:203–248, 1990.

[251] M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung*. Birkhäuser, Basel, 2012.

[252] R. von Schwerin, M. Winckler, and V. Schulz. Parameter estimation in discontinuous descriptor models. In D. Bestle and W. Schiehlen, editors, *IUTAM Symposium on Optimization of Mechanical Systems*, pages 269–276. Kluwer, 1996.

[253] R. Weiner and K. Strehmel. A Type Insensitive Code for Delay Differential Equations Basing on Adaptive and Explicit Runge–Kutta Interpolation Methods. *Computing*, 40:255–265, 1988.

[254] R. Weiss. Transportation Lag – An annotated Bibliography. *IRE Transactions on Automatic Control*, 4:56–64, 1959.

[255] D. R. Willé. DELSOL – a numerical code for the solution of delay differential equations. *Applied Numerical Mathematics*, 9:223–234, 1992.

[256] D. R. Willé and C. T. H. Baker. The tracking of derivative discontinuities in systems of delay-differential equations. *Applied Numerical Mathematics*, 9:209–222, 1992.

[257] R. B. Wilson. *A simplicial algorithm for concave programming.* PhD thesis, Harvard University, 1963.

[258] L. Wirsching. An SQP Algorithm with Inexact Derivatives for a Direct Multiple Shooting Method for Optimal Control Problems. Diploma thesis, Ruprecht–Karls–Universität Heidelberg, 2006.

[259] S. N. Wood. *Solv95: a numerical solver for systems of delay differential equations with switches.* University of St. Andrews, United Kingdom, 1999.

[260] W. H. Wu, F. S. Wang, and M. S. Chang. Sensitivity analysis of dynamic biological systems with time-delays. *BMC Bioinformatics*, 11:S12, 2010.

[261] L. Wunderlich. *Analysis and Numerical Solution of Structured and Switched Differential-Algebraic Systems.* PhD thesis, Technische Universität Berlin, 2008.

[262] D. Xu and Z. Yang. Impulsive delay differential inequality and stability of neural networks. *Journal of Mathematical Analysis and Applications*, 305:107–120, 2005.

[263] V. T. Xuyen and W. Y. Svrcek. On Equivalence of the Gauss-Newton Technique, the Parameter Influence Coefficient Technique, and the Quasilinearization Technique in Dynamic System Identification by Least Squares. *Journal of Optimization Theory and Applications*, 22:117–123, 1977.

[264] H. Yagoub, T. Nguyen-Ba, and R. Vaillancourt. Variable-step variable-order 3-stage Hermite–Birkhoff–Obrechkoff DDE solver of order 4 to 14. *Applied Mathematics and Computation*, 217:10247–10255, 2011.

[265] J. Yan, A. Zhao, and J. J. Nieto. Existence and Global Attractivity of Positive Periodic Solution of Periodic Single-Species Impulsive Lotka-Volterra Systems. *Mathematical and Computer Modelling*, 40:509–518, 2004.

[266] C. Yang and W. Zhu. Stability analysis of impulsive switched systems with time delays. *Mathematical and Computer Modelling*, 50:1188–1194, 2009.

[267] Y. Yang and J. Cao. Stability and Periodicity in Delayed Cellular Neural Networks with Impulsive Effects. *Nonlinear Analysis and Real World Applications*, 8:362–374, 2007.

[268] M. Zennaro. One–Step Collocation: Uniform Superconvergence, Predictor–Corrector Method, Local Error Estimate. *SIAM Journal on Numerical Analysis*, 22:1135–1152, 1985.

[269] M. Zennaro. Natural Continuous Extensions of Runge–Kutta Methods. *Numerische Mathematik*, 53:423–438, 1986.

[270] H. Ziegler. Resonanz bei konstanter Dämpfung. *Ingenieur–Archiv*, 9:50–76, 1938.

[271] H. Zivari-Piran. *Efficient Simulation, Accurate Sensitivity Analysis and Reliable Parameter Estimation For Delay Differential Equations.* PhD thesis, University of Toronto, 2009.

[272] H. Zivari-Piran and W. H. Enright. An efficient unified approach for the numerical solution of delay differential equations. *Numerical Algorithms*, 53:397–417, 2010.

[273] H. Zivari-Piran and W. H. Enright. Accurate first-order sensitivity analysis for delay differential equations. *SIAM Journal on Scientific Computing*, 34:A2704–A2717, 2012.