Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

FELIX BINKOWSKI

On the convergence behavior of spectral deferred correction methods for convection-diffusion equations

Based on Master's thesis
Felix Binkowski, July 10, 2017

ZIB Report 17-53 (November 2017)

Zuse Institute Berlin
Takustr. 7
14195 Berlin
Germany

Telephone: +49 30-84185-0
Telefax: +49 30-84185-125

E-mail: bibliothek@zib.de
URL: http://www.zib.de

ZIB-Report (Print) ISSN 1438-0064
ZIB-Report (Internet) ISSN 2192-7782

On the convergence behavior of spectral deferred correction methods for convection-diffusion equations

Felix Binkowski

November 30, 2017

Zusammenfassung

Spectral deferred correction (SDC) Methoden, vorgestellt von Dutt, Greengard und Rokhlin in [17], sind iterative Verfahren zur numerischen Lösung von Anfangswertproblemen für gewöhnliche Differentialgleichungen. Wenn diese Methoden konvergieren, dann wird unter Verwendung von Zeitschrittverfahren niedriger Ordnung eine Kollokationslösung berechnet. Die Lösung von steifen Anfangswertproblemen ist eine relevante Problemstellung in der numerischen Mathematik. SDC-Methoden, speziell für steife Probleme, werden von Martin Weiser in [55] konstruiert. Die Theorie und die Experimente beziehen sich dabei auf Probleme, die aus räumlich semidiskretisierten Reaktions-Diffusions-Gleichungen entstehen.
In dieser Arbeit werden die Ansätze aus [55] auf Konvektions-Diffusions-Gleichungen angewendet und das resultierende Konvergenzverhalten von SDC-Methoden untersucht. Basierend auf einem einfachen Konvektions-Diffusions-Operator, dessen spektrale Eigenschaften umfassend studiert werden, wird ein Schema zur Verbesserung dieses Verhaltens entwickelt. Numerische Experimente zeigen, dass eine Verbesserung der in [17] eingeführten SDC-Methoden möglich ist. Die Untersuchungen ergeben weiterhin, dass das auch für komplexere Konvektions-Diffusions-Probleme gilt.

Contents

Acknowledgments

I am very grateful to Prof. Dr. Jörg Liesen for introducing me to the field of numerical linear algebra. His very exciting lectures are the reason for my fascination for this part of numerical mathematics. I am indebted to him for his constant support and all the constructive discussions on my Master's thesis.
I would like to express my sincere gratitude to Dr. Martin Weiser for answering and discussing all of my questions. Thanks to him for deep insights into the numerics of ordinary and partial differential equations and into many other mathematical fields. This will assuredly be the foundation of my future work. He always had an open door for me to share his expertise. I am very thankful for the opportunity to work with him on relevant topics of numerical mathematics at Zuse Institute Berlin (ZIB).
I acknowledge, with thanks, the Computational Medicine group at ZIB. The atmosphere in this group always helped to have a huge motivation and the countless useful discussions contributed to the results of this thesis.
Last, but definitely not least, I want to thank my parents for their everlasting support and for giving me the chance to find and develop my interests in applied mathematics.

1 Introduction

> ... Around 1960, things became completely different and everyone became aware that the world was full of stiff problems.
>
> (G. Dahlquist in Aiken 1985) [32, p. 2]

The numerical solution of initial value problems (IVPs) for systems of ordinary differential equations (ODEs) has been an essential task in numerical mathematics for a long time. This thesis concerns spectral deferred correction (SDC) methods, as introduced by Dutt, Greengard and Rokhlin in [17], for solving IVPs. These methods are iterative schemes which compute an approximate solution to a collocation solution by applying a low order time stepping scheme in each iteration step. Actual research areas involving SDC methods are the construction of efficient time-parallel solvers [4], [5], [8], [20], [22], [43] and solving IVPs with an inexact right hand side [57]. SDC methods function as preconditioners for collocation systems and Krylov subspace methods can be used to solve these systems efficiently. This is applied to construct appropriate methods for the solution of differential algebraic equation IVPs [6], [33], [34], [36]. A further numerical research field is the modeling of multiscale problems, where a spatio-temporal adaptivity can be a promising approach. An application for the cardiac electro-mechanical coupling model can be found in [58].
In all of these applications, we are confronted with stiff ODEs. A definition of a stiff problem is that for this type of differential equation, the implicit Euler method is more suitable than the explicit Euler method. By using implicit methods, in each step a linear system has to be solved and the Jacobian of the ODE has to be computed. The resulting increase of computational cost makes the numerical solution of stiff problems more challenging compared to the solution of non-stiff problems. The derivation of approaches and efficient algorithms for stiff problems is a crucial topic in the numerical treatment of ODEs.
The construction of SDC methods for stiff problems is addressed in, for example, [2], [17], [48] and [55]. With the aim of obtaining specific convergence properties, Martin Weiser [55] introduces new SDC methods using implicit Runge-Kutta (DIRK) sweeps. These methods are based on simple linear algebraic considerations. The study of [55] formulates convergence objectives and presents approaches on how to achieve them. SDC methods are interpreted as fixed point iterations and their construction is based on linear collocation systems resulting of Dahlquist's equation. The motivation there is the numerical solution of time dependent partial differential equations (PDEs) which are before semi-discretized in space. The considered class of PDEs are reaction-diffusion equations.
More challenging in the numerical treatment are convection-diffusion equations, which arise in numerous physical problems, for example, in models of flow, models of semiconductor devices or other transport phenomena of physical quantities. In this thesis, we

apply the approaches of [55] to convection-diffusion equations and introduce a framework for constructing specific SDC methods for this type of PDEs. In Chapter 2, SDC methods are derived as in [17] and considered as fixed point iterations for solving linear collocation systems. Chapter 3 summarizes the results of [55] to prepare the study of convection-diffusion equations, which is covered in Chapter 4. After an introduction to these equations, a treatment of their spectral properties is presented. The obtained insights are used for the following construction of SDC methods with special convergence properties. Several numerical experiments are carried out for different convergence objectives. Finally, Chapter 5 provides numerical experiments to SDC methods for systems of ODEs resulting from convection-diffusion problems.

2 SDC methods from the perspective of linear algebra

This chapter provides an introduction to SDC methods. In the first Section 2.1, these methods are derived with standard concepts of numerical mathematics. As a preparation for the following chapters and for a better understanding of the SDC framework, SDC methods are then regarded as Runge-Kutta methods. The Section 2.2 leads to a matrix formulation and in Section 2.3, SDC methods are considered as fixed point iterations and a condition for their convergence is derived. We assume some background knowledge on the numerical solution of ODEs. A detailed treatment to this specifically can be found in [31] and [32]. For numerical analysis in general, we refer to the book series [11], [12], [13] and to the theoretical treatise [1].

2.1 SDC methods for solving ODEs

SDC methods can be used for solving IVPs for systems of ODEs. They are iterative schemes and in each iteration step a correction of an approximate solution is computed. Dutt, Greengard and Rokhlin introduced SDC methods in [17]. This approach for deferred correction methods is based on the Picard integral equation.

2.1.1 Derivation of SDC methods with the Picard integral equation

This work concerns the numerical solution of the IVP

$$\begin{aligned} y'(t) &= f(y(t)), \qquad t \in [0,\tau], \\ y(0) &= y_0 \end{aligned} \tag{2.1}$$

where $y(t), y_0 \in \mathbb{C}^N$ and $f : \mathbb{C}^N \to \mathbb{C}^N$. It is well known that a unique solution $y(t)$ exists if the right hand side f satisfies some simple regularity conditions, see Theorem 110C in [7] or Theorem 2.7 in [11]. To find a numerical solution of (2.1), we consider the error

$$\delta^{[0]}(t) = y(t) - y^{[0]}(t) \tag{2.2}$$

with a given approximate solution $y^{[0]}(t)$ where $y^{[0]}(0) = y_0$. Differentiation of (2.2) yields

$$\begin{aligned} \delta^{[0]\prime}(t) &= y'(t) - y^{[0]\prime}(t) \\ &= f(y(t)) - y^{[0]\prime}(t) \\ &= f\left(y^{[0]}(t) + \delta^{[0]}(t)\right) - y^{[0]\prime}(t) \end{aligned}$$

and by that the new IVP for the error is given by

$$\begin{aligned} \delta^{[0]\prime}(t) &= f\left(y^{[0]}(t) + \delta^{[0]}(t)\right) - y^{[0]\prime}(t), \qquad t \in [0, \tau], \\ \delta^{[0]}(0)) &= 0, \end{aligned} \tag{2.3}$$

where the right hand side f is from the IVP (2.1). The solution of (2.3) can be written as a Picard integral equation

$$\delta^{[0]}(t) = \delta^{[0]}(0) + \int_{s=0}^{t} \left[f\left(y^{[0]}(s) + \delta^{[0]}(s)\right) - y^{[0]\prime}(s)\right] \mathrm{d}s.$$

To achieve a numerical solution, the given time interval $[0, \tau]$ is discretized by the points $0 = t_0 < t_1 < \ldots < t_n = \tau$, hence

$$\delta^{[0]}(t_i) = \underbrace{\int_{s=0}^{t_{i-1}} \left[f\left(y^{[0]}(s) + \delta^{[0]}(s)\right) - y^{[0]\prime}(s)\right] \mathrm{d}s}_{=\delta^{[0]}(t_{i-1})} + \int_{s=t_{i-1}}^{t_i} \left[f\left(y^{[0]}(s) + \delta^{[0]}(s)\right) - y^{[0]\prime}(s)\right] \mathrm{d}s$$

at these time points. Rearranging this equation yields the formula

$$\begin{aligned} \delta^{[0]}(t_i) = \delta^{[0]}(t_{i-1}) &+ \int_{s=t_{i-1}}^{t_i} \left[f\left(y^{[0]}(s) + \delta^{[0]}(s)\right) - f\left(y^{[0]}(s)\right)\right] \mathrm{d}s \\ &+ \int_{s=t_{i-1}}^{t_i} f\left(y^{[0]}(s)\right) \mathrm{d}s - \left(y^{[0]}(t_i) - y^{[0]}(t_{i-1})\right), \end{aligned}$$

which has two integrals. The first one is approximated with a simple numerical time stepping scheme of low order, see Definition 1.2 in [30]. For the approximation of the second integral, a spectral integration is applied [26], [27].

Definition 2.1. The matrix $S \in \mathbb{R}^{(n+1)\times(n+1)}$ which coefficients are defined by

$$S_{ik} = \begin{cases} 0, & \text{if } i = 0, \\ \frac{1}{t_i - t_{i-1}} \int_{t=t_{i-1}}^{t_i} L_k(t)\,\mathrm{d}t, & \text{if } i > 0, \end{cases} \qquad i, k = 0, \ldots, n,$$

where $L_k(t)$ is a Lagrange basis polynomial, will be called the *spectral integration matrix* for the time points $t_0, \ldots, t_n$. If all these time points are interpolation nodes, the Lagrange basis polynomial is defined by

$$L_k(t) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{t - t_j}{t_k - t_j}. \tag{2.4}$$

Spectral integration matrices with interpolation nodes from quadrature formulas were introduced in [17]. In the next subsection, such nodes are covered in detail. Further information on polynomial interpolation can be found in Section 7.1 in [12].
Possible low order time stepping schemes are, for example, the explicit and implicit Euler method. A detailed treatment of these Euler methods can be found in Section 20 and 21 in [7]. Using the explicit Euler method as time stepping scheme yields the approximation

$$\begin{aligned}\delta_i^{[0]} = \delta_{i-1}^{[0]} &+ \tau_i \left[f\left(y_{i-1}^{[0]} + \delta_{i-1}^{[0]}\right) - f\left(y_{i-1}^{[0]}\right)\right] \\ &+ \tau_i \sum_{k=0}^{n} S_{ik}\, f\left(y_k^{[0]}\right) - \left(y_i^{[0]} - y_{i-1}^{[0]}\right), \qquad i = 1, \ldots, n \end{aligned} \tag{2.5}$$

for the time step $\tau_i = t_i - t_{i-1}$, where $\delta_i^{[0]} \approx \delta^{[0]}(t_i)$ and $y_i^{[0]} = y^{[0]}(t_i)$, starting at $\delta_0^{[0]} = 0$. With this equation it is possible to compute the errors $\delta_1^{[0]}, \ldots, \delta_n^{[0]}$. Approximating the first integral with the implicit Euler method yields

$$\begin{aligned}\delta_i^{[0]} = \delta_{i-1}^{[0]} &+ \tau_i \left[f\left(y_i^{[0]} + \delta_i^{[0]}\right) - f\left(y_i^{[0]}\right)\right] \\ &+ \tau_i \sum_{k=0}^{n} S_{ik}\, f\left(y_k^{[0]}\right) - \left(y_i^{[0]} - y_{i-1}^{[0]}\right), \qquad i = 1, \ldots, n. \end{aligned} \tag{2.6}$$

To solve this equation for $\delta_i^{[0]}$, we consider the first order Taylor polynomial

$$f\left(y_i^{[0]}\right) + f'\left(y_i^{[0]}\right)\left(\delta_i^{[0]} - y_i^{[0]}\right) \approx f\left(\delta_i^{[0]}\right) \tag{2.7}$$

as a linear approximation of f at the point $y_i^{[0]}$. From this linearization follows

$$f'\left(y_i^{[0]}\right)\left(\delta_i^{[0]}\right) \approx f\left(y_i^{[0]} + \delta_i^{[0]}\right) - f\left(y_i^{[0]}\right)$$

and inserting this into equation (2.6) leads to the linearly implicit scheme

$$\left(I - \tau_i f'\left(y_i^{[0]}\right)\right)\delta_i^{[0]} = \delta_{i-1}^{[0]} + \tau_i \sum_{k=0}^{n} S_{ik}\, f\left(y_k^{[0]}\right) - \left(y_i^{[0]} - y_{i-1}^{[0]}\right), \quad i = 1, \ldots, n, \tag{2.8}$$

which can be solved for $\delta_i^{[0]}$. For this scheme, the derivative of the right hand side f, the so-colled Jacobian $J_f = \partial f / \partial y$, has to be available.
Adding the error $\delta_i^{[0]}$ resulting from the explicit scheme (2.5) or the linearly implicit scheme (2.8) and the given approximate solution $y_i^{[0]}$ yields a new discrete corrected solution

$$y_i^{[1]} = y_i^{[0]} + \delta_i^{[0]}, \qquad i = 0, \ldots, n. \tag{2.9}$$

This usually leads to a better approximate solution

$$y^{[1]}(t) = y^{[0]}(t) + \sum_{k=0}^{n} \delta_k^{[0]} L_k(t),$$

realized by a polynomial interpolation. Iteration of the scheme (2.9) yields

$$y_i^{[j+1]} = y_i^{[j]} + \delta_i^{[j]}, \qquad i = 0, \dots, n, \quad j = 0, \dots . \tag{2.10}$$

This correction scheme employing the explicit Euler method was introduced in [17] as one of several SDC methods. Besides explicit and linearly implicit Euler formulations as presented here, SDC methods with other low order time stepping schemes are feasible. The choice of an implicit approach can be a reasonable way to solve ODEs, while explicit methods are not satisfactory in some case. Such problems are called stiff ODEs, see Subsection 112 in [7] or Subsection 4.1.3 in [11] for additional explanations.
Furthermore, the choice of the nodes for the Lagrange interpolation plays an important role for the construction of SDC methods. The selected nodes can enable desirable qualities, for example, a high order of SDC methods and reasonable stability properties. This is covered in the next subsection.

Remark 2.2. Without loss of generality, in this work, we regard autonomous ODEs $y' = f(y)$ as from the IVP (2.1), where the right hand side f does not explicitly depend on the time t. Considering a non-autonomous ODE $\varphi' = g(t, \varphi)$, the autonomous ODE $y' = f(y) = [g(t, \varphi)^T, 1]^T$ where $y = [\varphi^T, t]^T$ can be introduced.

2.1.2 SDC methods as Runge-Kutta methods and the collocation solution

A common approach that was also applied in [17] is to use a time discretization for SDC methods arising from quadrature rules. In the following, this application of quadrature nodes is motivated and the view of SDC methods as Runge-Kutta methods is presented. A detailed analysis of Runge-Kutta methods can be found in [30], [31] and [32].

Definition 2.3. (Definition 1.1 in [30]) An *s-stage Runge-Kutta method* for the IVP (2.1) on the time interval $[0, \tau]$ is defined by the real numbers b_i, a_{ij} where $i, j = 1, \dots, s$. The numerical solution of this Runge-Kutta method at the time point $t = \tau$ is given by

$$y_\tau = y_0 + \tau \sum_{i=1}^{s} b_i k_i,$$
$$k_i = f\left(y_0 + \tau \sum_{j=1}^{s} a_{ij} k_j \right) \quad \text{and} \quad c_i = \sum_{j=1}^{s} a_{ij}, \quad i = 1, \dots, s.$$

A common way to display the coefficients b_i and a_{ij} is the Butcher tableau

$$\begin{array}{c|ccc} c_1 & a_{11} & \cdots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \cdots & b_s \end{array} .$$

Runge-Kutta methods are one-step methods for the numerical solution of IVPs. The explicit and the implicit Euler method are 1-stage Runge-Kutta methods of order one. Due to a certain choice of the coefficients of a Runge-Kutta method, it is possible to construct methods of higher order.

Definition 2.4. (Definition 1.2 in [30]) A Runge-Kutta method has the *order* p if

$$y_\tau - y(\tau) = \mathcal{O}(\tau^{p+1}) \quad \text{as } \tau \to 0$$

holds for all sufficiently regular IVPs (2.1), where y is the exact solution of an IVP and y_τ is the numerical solution of the Runge-Kutta method at the end point of the time interval $[0, \tau]$.

For SDC methods, the ODE's right hand side is interpolated with Lagrange polynomials of the degree $s - 1$. In Definition 2.1, for example, there holds $s = n + 1$, and the regarded ODE is solved with an integration of the resulting interpolation polynomial. Assuming that an SDC method converges, the corresponding numerical solution is also a polynomial which satisfies the ODE in certain time points. Thus, SDC methods are iterative methods for the solution of collocation systems.

Definition 2.5. (Definition 1.3 in [30]) A *collocation method* for the IVP (2.1) is a numerical method which gives a polynomial approximation. This *collocation polynomial* $y_c(t)$ satisfies (2.1) in the *collocation discretization*, i.e., in the s distinct *collocation points* $c_1\tau, \ldots, c_s\tau \in [0, \tau]$, where $c_1, \ldots, c_s$ are real numbers. Thus, there holds

$$\begin{aligned} y_c'(c_i\tau) &= f(y_c(c_i\tau)), \qquad i = 1, \ldots, s, \\ y_c(0) &= y_0 \end{aligned}$$

for the IVP (2.1).

The Definitions 2.3 and 2.5 are the basis for the next lemma.

Lemma 2.6. (Theorem 1.4 in [30]) A collocation method of Definition 2.5 is an s-stage Runge-Kutta method of Definition 2.3. The coefficients of this Runge-Kutta method are given by

$$a_{ij} = \int_0^{c_i} L_j(t)\mathrm{d}t, \qquad b_i = \int_0^1 L_i(t)\mathrm{d}t$$

where $c_1, \ldots, c_s$ define the collocation discretization of Definition 2.5. The polynomial $L_i(t)$ is the Lagrange polynomial with $L_i(t) = \prod_{l \neq i}(t - c_l)/(c_i - c_l)$.

Proof. See proof of Theorem 1.4 in [30]. □

This reveals the link between SDC and Runge-Kutta methods. If an SDC method with the collocation points $c_1\tau, \ldots, c_s\tau \in [0, \tau]$ converges, it converges to the solution of an s-stage Runge-Kutta with coefficients of Lemma 2.6. Additionally, an SDC method with a fixed number of iterations is a Runge-Kutta method.

Definition 2.7. We call an SDC method with j iterations an *SDC-j method*. This method is a Runge-Kutta method where the Butcher tableau depends on the collocation points, the low order time stepping scheme and the number of iterations.

Common choices for the collocation discretization of Runge-Kutta methods are quadrature nodes from the Gauss, Radau or Lobatto quadrature with order $2s$, $2s-1$ and $2s-2$, respectively, see Subsection II.1.3 in [30]. These quadrature rules represent methods for numerical integration and they are exact for polynomials up to a certain degree. The so-called order of such a quadrature rule is the maximum order of the polynomials being integrated exactly.

Lemma 2.8. (Theorem 1.5 in [30]) Consider an s-stage Runge-Kutta method of Definition 2.3 where the coefficients a_{ij} and b_i are from Lemma 2.6. Let the collocation discretization be given by the nodes of a quadrature rule. The order of this Runge-Kutta method is the same as from the underlying quadrature rule.

Proof. See proof of Theorem 1.5 in [30]. □

For this work, more details on quadrature rules are not necessary and we refer for further information to Chapter 9 in [12]. Radau-IIa quadrature rules with s nodes have the order $2s-1$ and a Runge-Kutta method with Radau-IIa nodes is L-stable, see Theorem 6.46 in [11]. An L-stable method is A-stable, which means that the method is stable on the entire left half-plane of the complex plane. For an L-stable method further holds $\lim_{z\to\infty} R(z) = 0$ for the stability function $R(z)$, see Definition 3.7 in [32]. Therefore, the collocation with Radau-IIa nodes can be a reasonable way for constructing SDC methods. For a general meaning of these concepts and, in particular, of stability and the stability function, the reader is referred to Subsection 350 in [7], Subsection 6.1.2 in [11] or Section IV.3 in [32]. Furthermore, a detailed discussion of the choice of quadrature nodes for SDC methods is presented in [40].
At this point, a condition for the convergence of SDC methods is still missing. In the remaining part of this chapter, SDC methods are introduced as fixed point iterations in a matrix form and based on this, a condition for their convergence is derived.

2.2 Matrix formulation

In the same manner as in [55], we now introduce the view of SDC methods from the perspective of linear algebra. A linearly implicit formulation (2.8) is chosen for the construction of the SDC methods. The time interval is discretized with Radau-IIa nodes, where the left interval end point t_0 is not included. Therefore, all entries of the first column of the spectral integration matrix S are zero. Reducing this matrix by the first row and the first column yields a new integration matrix $S^r \in \mathbb{R}^{n\times n}$. To get a matrix formulation, we start with the linearly implicit Euler scheme (2.8) and define the *bidiagonal approximate differentiation matrix*

$$D_{ik} = \frac{\tau}{\tau_i}\left(\delta_{i,k} - \delta_{i,k+1}\right), \qquad i,k = 1,\dots,n, \tag{2.11}$$

where $\delta_{i,j}$ is the Kronecker-δ. Using the matrices S^r and D and multiplying (2.8) with τ/τ_i yields the matrix form

$$\left(I_N \otimes D - Z\left(I_N \otimes \hat{S}\right)\right)\bar{\delta}^{[0]} = -\left(I_N \otimes D\right)\bar{y}^{[0]} + \left(I_N \otimes S^r\right)\bar{f}(\bar{y}^{[0]}) + \frac{\tau}{\tau_1} y_0 \otimes e_1, \tag{2.12}$$

where $D, \hat{S}, S^r \in \mathbb{R}^{n\times n}$, the matrix $I_N \in \mathbb{R}^{N\times N}$ is the identity matrix and $\otimes$ is the Kronecker product. The matrix $Z \in \mathbb{C}^{Nn\times Nn}$ is composed of the matrices $Z_{ik} \in \mathbb{C}^{n\times n}$, where

$$Z_{ik} = \tau \begin{bmatrix} (J_f)_{ik}\left(y_{t_1}^{[0]}\right) & & \\ & \ddots & \\ & & (J_f)_{ik}\left(y_{t_n}^{[0]}\right) \end{bmatrix}, \qquad i,k = 1,\dots,N$$

with the Jacobian $J_f \in \mathbb{C}^{N\times N}$ and the vector valued approximations $y_{t_1}^{[0]}, \dots, y_{t_n}^{[0]} \in \mathbb{C}^N$ at certain time points. The vector $y_0 \in \mathbb{C}^N$ is the initial value of the IVP (2.1) and $e_1 = [1, 0, \dots, 0]^T \in \mathbb{R}^n$. This leads to the vectors $\bar{y}^{[0]}, \bar{\delta}^{[0]}, \bar{f}(\bar{y}^{[0]}) \in \mathbb{C}^{Nn}$ and their components are the entries of the vectors $y^{[0]}, \delta^{[0]}, f(y^{[0]}) \in \mathbb{C}^N$, successively evaluated at the time points $t_1, \dots, t_n$. Using the linearly implicit Euler scheme leads to $\hat{S} = I_n$, where $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix. This matrix takes on the role of an approximate integration matrix and thus it is an approximation of the spectral integration matrix S^r. Due to the Lagrange interpolation with s collocation points, the spectral integration with S^r is exact for polynomials up to a degree of $s-1$. If another low order time stepping scheme is chosen, the matrix $\hat{S}$ has to be replaced by another matrix. Taking, for example, the explicit Euler scheme as low order time stepping method yields

$$\hat{S}_{ik} = \delta_{i,k+1} \qquad i,k = 1,\dots,n.$$

2.3 SDC methods as fixed point iterations for solving linear collocation systems

In the following, SDC methods are considered as fixed point iterations $y^{[j+1]} = F(y^{[j]})$ with the mapping $F : \mathbb{C}^n \to \mathbb{C}^n$ and their convergence towards the collocation solution is analyzed. The principles of iterative methods can be found in almost all books of numerical linear algebra. See, for example, [53] for fundamental ideas and [25] as an encyclopedic treatise with further details. The textbook [49] covers iterative methods for linear systems in monograph form. Furthermore, for advanced techniques of iterative methods, we refer to [42], where an extensive study of Krylov subspace methods is presented.

Considering an ODE system with constant coefficients, i.e., its Jacobian is constant, the system can be decoupled if the Jacobian is diagonalizable. Each equation of the decoupled system can be solved independently and after this, the solutions can be superposed so that a solution of the ODE system is obtained. Thus, the behavior of the ODE system can be described by the separated solutions of the decoupled system. For a non-diagonalizable Jacobian, there exists such an approach using a Jordan decomposition. We refer to Section I.12 in [31] for more information on these approaches. Based on this, the approach for the following chapters is to regard simple one-dimensional IVPs which offer an insight into many properties of SDC methods. Furthermore, in this work, semi-discretized linear PDEs are of interest and thus the considered ODEs are also linear.

Definition 2.9. The *IVP of Dahlquist's equation* is given by

$$\begin{aligned} y'(t) &= f(y(t)) = \lambda y(t), \qquad t \in [0, \tau], \\ y(0) &= \gamma_0 = 1, \end{aligned} \tag{2.13}$$

where $y(t), \gamma_0 \in \mathbb{C}$ and $f : \mathbb{C} \to \mathbb{C}$ is linear with $\lambda \in \mathbb{C}$.

There is a scalar Jacobian, where $J_f = \lambda$, thus the derived linearly implicit scheme (2.8) can be readily applied. With the linear right hand side f the matrix form (2.12) simplifies to

$$\left(D - z\hat{S}\right) \delta^{[0]} = -\left(D - zS^r\right) y^{[0]} + \frac{\tau}{\tau_1} \gamma_0 \, e_1, \tag{2.14}$$

where $z = \lambda\tau$. Combining this matrix form of one SDC iteration step, in the following referred to as SDC sweep, with the correction scheme (2.10) yields the fixed point iteration

$$\begin{aligned} y^{[j+1]} &= y^{[j]} + \left(D - z\hat{S}\right)^{-1} \left(-\left(D - zS^r\right) y^{[j]} + \frac{\tau}{\tau_1} \gamma_0 \, e_1 \right) \\ &= \left[I - \left(D - z\hat{S}\right)^{-1} \left(D - zS^r\right) \right] y^{[j]} + \left(D - z\hat{S}\right)^{-1} \frac{\tau}{\tau_1} \gamma_0 \, e_1, \end{aligned}$$

provided $D - z\hat{S}$ is invertible. In the following, only invertible matrices $D - z\hat{S}$ are constructed. The next definition is based on this fixed point iteration.

Definition 2.10. Let the matrix $D - z\hat{S}$ be non-singular. An *SDC fixed point iteration* for the linear IVP of Dahlquist's equation of Definition 2.9 on a collocation discretization where the left interval end point t_0 is not included is given by

$$\begin{aligned} y^{[j+1]} &= G \, y^{[j]} + g, \quad j = 0, 1, \ldots, \\ G = I - \left(D - z\hat{S}\right)^{-1} \left(D - zS^r\right), \quad g &= \left(D - z\hat{S}\right)^{-1} \frac{\tau}{\tau_1} \gamma_0 \, e_1, \end{aligned} \tag{2.15}$$

where $y^{[j]}, g \in \mathbb{C}^n$, $z = \lambda\tau \in \mathbb{C}$ and $y^{[0]}$ is an arbitrary starting vector. The matrix $G \in \mathbb{C}^{n \times n}$ is the iteration matrix of the fixed point iteration. The matrix $S^r \in \mathbb{R}^{n \times n}$ is a spectral integration matrix of Definition 2.1 which is reduced by the left interval end point t_0. The matrix $D \in \mathbb{R}^{n \times n}$ (2.11) is an approximate bidiagonal differentiation matrix and $\hat{S} \in \mathbb{R}^{n \times n}$ takes on the role of an approximate integration matrix and is determined by the low order time stepping scheme of the SDC method.

As described in Subsection 2.1.2, SDC methods give a collocation solution of the corresponding IVP if they converge. The same insight is obtained by considering the equation (2.14) with an error $\delta^{[0]} = 0$. This leads to the linear system of equations

$$(D - zS^r) y^{[0]} = \frac{\tau}{\tau_1} \gamma_0 \, e_1.$$

Considering a $y^{[0]}$ satisfying this equation leads to a collocation solution of the IVP of Dahlquist's equation of Definition 2.9 which is given by

$$y_c(t) = \gamma_0 L_0(t) + \sum_{k=1}^{n} y_k^{[0]} L_k(t).$$

The polynomials $L_k(t)$, where $k = 0 \ldots, n$, are the Lagrange basis polynomials for the nodes $t_0, \ldots, t_n$. This leads to the next lemma.

Lemma 2.11. If the matrix $D - zS^r$ is non-singular, then the unique fixed point of a corresponding SDC fixed point iteration of Definition 2.10 is given by

$$y_c = (D - zS^r)^{-1} \frac{\tau}{\tau_1} \gamma_0 \, e_1.$$

Proof. Let y_c be a fixed point of an SDC fixed point iteration of Definition 2.10 and let $D - zS^r$ be non-singular. This leads to

$$\begin{aligned} y_c = Gy_c + g &= \left[I - \left(D - z\hat{S} \right)^{-1} (D - zS^r) \right] y_c + \left(D - z\hat{S} \right)^{-1} \frac{\tau}{\tau_1} \gamma_0 \, e_1 \\ \Leftrightarrow \left(D - z\hat{S} \right) y_c &= \left[\left(D - z\hat{S} \right) - (D - zS^r) \right] y_c + \frac{\tau}{\tau_1} \gamma_0 \, e_1 \\ \Leftrightarrow (D - zS^r) \, y_c &= \frac{\tau}{\tau_1} \gamma_0 \, e_1 \\ \Leftrightarrow \qquad y_c &= (D - zS^r)^{-1} \frac{\tau}{\tau_1} \gamma_0 \, e_1. \end{aligned} \tag{2.16}$$

□

For a certain start approximation $y^{[0]}$, the convergence behavior of such an iteration depends only on the properties of its iteration matrix G. In the following part of this section, two important properties are covered, based on which we present conditions for the convergence of SDC fixed point iterations of Definition 2.10. Furthermore, quantities for the convergence speed and the error reduction of these SDC methods are introduced. For the beginning, consider the spectral radius $\rho(G)$ of an iteration matrix G.

Definition 2.12. The *spectral radius* ρ of a matrix $G \in \mathbb{C}^{n \times n}$ is defined by

$$\rho(G) := \max_{1 \leq i \leq n} |\lambda_i(G)| \, ,$$

where λ_i are the eigenvalues of G.

This leads to the following theorem, which gives a necessary and sufficient condition for the convergence of SDC fixed point iterations of Definition 2.10.

Theorem 2.13. An SDC fixed point iteration of Definition 2.10 converges to its unique fixed point $y_c = (D - zS^r)^{-1} \frac{\tau}{\tau_1} \gamma_0 \, e_1$ for any start point $y^{[0]}$ if and only if there holds $\rho(G) < 1$, where G is the iteration matrix of the SDC method.

Proof. Consider an SDC fixed point iteration of Definition 2.10. Then, we know from Lemma 2.11 that the collocation solution $y_c = (D - zS^r)^{-1} \frac{\tau}{\tau_1} \gamma_0 \, e_1$ is the unique fixed point of this iteration. For the j-th iteration, we obtain

$$\begin{aligned} y^{[j]} - y_c &= G \left(y^{[j-1]} - y_c \right) \\ &= G \left(Gy^{[j-2]} - g - (Gy_c - g) \right) = \ldots = G^j \left(y^{[0]} - y_c \right), \end{aligned} \tag{2.17}$$

where $y^{[0]}$ is an arbitrary starting point.
From Theorem 7.1.9 in [25] or Section 16.2 in [41] we know that a Jordan decomposition $G = XJX^{-1}$ exists for the matrix $G \in \mathbb{C}^{n\times n}$, where $X \in \mathbb{C}^{n\times n}$ is non-singular and $J = \text{diag}(J_{l_1}, \ldots, J_{l_m}) \in \mathbb{C}^{n\times n}$. The Jordan blocks are defined as

$$J_{l_i} = \begin{bmatrix} \lambda_i & 1 & & \\ & \ddots & \ddots & \\ & & & 1 \\ & & & \lambda_i \end{bmatrix} \in \mathbb{C}^{l_i \times l_i},$$

where $l_1 + \ldots + l_m = n$. The complex numbers $\lambda_1, ..., \lambda_m$ are eigenvalues of G, which do not need to be pairwise distinct. The dimensions and the number of the Jordan blocks are unique. Forming the product of G yields $G^j = (XJX^{-1})^j = XJ^jX^{-1}$. If the iteration (2.17) converges to y_c, it has to satisfy

$$\lim_{j\to\infty} y^{[j]} - y_c = \lim_{j\to\infty} G^j \left(y^{[0]} - y_c\right) = \lim_{j\to\infty} XJ^jX^{-1} \left(y^{[0]} - y_c\right) = 0. \tag{2.18}$$

We obtain the block diagonal matrix $J^j = \text{diag}\left(J_{l_1}^j, \ldots, J_{l_m}^j\right)$ and the j-th power of the Jordan blocks can be derived with induction. From the result of this induction follows immediately that (2.18) holds for any $y^{[0]}$ if and only if $\rho(G) < 1$. □

A sufficient condition for the convergence of SDC fixed point iterations of Definition 2.10 is given in the next lemma.

Lemma 2.14. If $\|G\| < 1$ for some matrix norm $\|\cdot\|$, then the corresponding SDC fixed point iteration of Definition 2.10 converges for any starting point $y^{[0]}$ to its unique fixed point $y_c = (D - zS^r)^{-1} \frac{\tau}{\tau_1}\gamma_0\, e_1$.

Proof. Consider an SDC fixed point iteration of Definition 2.10. Then, we know from Lemma 2.11 that the unique fixed point is given by $y_c = (D - zS^r)^{-1} \frac{\tau}{\tau_1}\gamma_0\, e_1$. Considering the error $y^{[j]} - y_c$ in some consistent norm $\|\cdot\|$ for an arbitrary starting vector $y^{[0]}$ leads to

$$\left\|y^{[j]} - y_c\right\| = \left\|G^j \left(y^{[0]} - y_c\right)\right\| \leq \|G\|^j \left\|y^{[0]} - y_c\right\|$$

and using $\|G\| < 1$ yields

$$\begin{aligned} &\lim_{j\to\infty} \left\|y^{[j]} - y_c\right\| \leq \lim_{j\to\infty} \|G\|^j \left\|y^{[0]} - y_c\right\| = 0 \\ \Rightarrow &\lim_{j\to\infty} y^{[j]} = y_c. \end{aligned}$$

□

Equally important as the stated results about the existence of convergence is knowledge on the convergence speed of an SDC method.

Definition 2.15. (Section 4.2 in [49]) The *asymptotic contraction factor* of an SDC fixed point iteration of Definition 2.10 is defined by

$$\Phi_\infty = \lim_{j\to\infty} \left(\max_{y^{[0]}\in\mathbb{C}^n} \frac{\left\| y^{[j]} - y_c \right\|}{\left\| y^{[0]} - y_c \right\|} \right)^{1/j}, \tag{2.19}$$

where y_c is the unique fixed point of the SDC method and $y^{[j]} - y_c$ is the error in the j-th iteration.

Lemma 2.16. The asymptotic contraction factor Φ_∞ of an SDC fixed point iteration of Definition 2.10 is equal to the spectral radius $\rho(G)$, where G is the iteration matrix of the SDC method.

Proof. Inserting $y^{[j]} - y_c = G^j \left(y^{[0]} - y_c \right)$, where y_c is the unique fixed point of the SDC method, into (2.19) yields

$$\begin{aligned} \Phi_\infty &= \lim_{j\to\infty} \left(\max_{y^{[0]}\in\mathbb{C}^n} \frac{\left\| G^j \left(y^{[0]} - y_c \right) \right\|}{\left\| y^{[0]} - y_c \right\|} \right)^{1/j} \\ &= \lim_{j\to\infty} \left\| G^j \right\|^{1/j}, \end{aligned}$$

where the definition of an induced matrix norm $\|\cdot\|$ is used, see Subsection 4.2.1 in [49]. From Theorem 1.12 in [49] we know that $\rho(G) = \lim_{j\to\infty} \left\| G^j \right\|^{1/j}$ and therefore the equality $\Phi_\infty = \rho(G)$ holds. □

The asymptotic contraction factor determines the behavior of SDC methods in the limit case of $j \to \infty$ SDC sweeps. A further important objective is to have a high error reduction in the first few SDC sweeps. Performing just a few iterations and hereby obtaining a good approximation is one motivation for using iterative methods. Considering a certain starting point $y^{[0]}$ and the bound

$$\left\| y^{[j]} - y_c \right\| \leq \|G\|^j \left\| y^{[0]} - y_c \right\|$$

of the proof of Lemma 2.14 offers the insight that the norm $\|G\|$ gives information on the error reduction in each SDC sweep, which includes the first few SDC sweeps. Considering the error at all time points $t_1, \ldots, t_n$ leads to the next definition.

Definition 2.17. The *local pre-asymptotic contraction factor* of an SDC fixed point iteration of Definition 2.10 with respect to a matrix norm $\|\cdot\|$ is defined by

$$\Phi_l = \|G\|,$$

where G is the iteration matrix of the SDC method.

We can also regard the error at the end point t_n of the time interval, which determines the overall quality of the computed solution $y^{[j]}$.

Definition 2.18. The *global pre-asymptotic contraction factor* of an SDC fixed point iteration of Definition 2.10 with respect to a matrix norm $\|\cdot\|$ is defined by

$$\Phi_g = \left\| e_n^T G \right\|,$$

where G is the iteration matrix of the SDC method and $e_n = [0, \ldots, 0, 1]^T \in \mathbb{R}^n$ is the n-th cartesian unit vector.

Finally, we have conditions for the convergence and a quantity for the convergence speed of SDC fixed point iterations of Definition 2.10. There are different meanings for "convergence speed", "fast convergence" and "accelerating the convergence" in the literature. If these terms are used in the next chapters, the asymptotic contraction factor Φ_∞ is considered. Furthermore, the local and global pre-asymptotic contraction factor are quantities for the error reduction in the first few SDC sweeps. In the next chapter, all these factors are used for the study of the asymptotic and pre-asymptotic behavior of SDC methods for different IVPs of Dahlquist's equation (2.13).

Remark 2.19. For the derivation of the linearly implicit Euler scheme (2.8), we consider a linearization of the ODE's right hand side, see the Taylor polynomial (2.7). Dahlquist's equation itself is linear and therefore this linearization is not necessary in this special case. Evaluating the right hand side leads to the same implicit scheme.

3 Faster SDC convergence for reaction-diffusion equations

The main focus of this thesis is in the convergence behavior of SDC methods for the problem class of convection-diffusion equations. Thus, we are interested in the numerical solution of these PDEs with respect to the time variable. After spatial discretization of a PDE, the resulting IVP can be solved with the SDC framework presented in the previous chapter. The Jacobian of the right hand side of the corresponding ODE is equal to the discretized partial differential operator. An application-oriented introduction to the numerical solution of PDEs can be found in [28] and [35] gives an overview of the numerical treatment of time dependent PDEs. The reader is further referred to [21] for important topics in the theory of PDEs.
To study the SDC convergence behavior, SDC fixed point iterations of Definition 2.10 are considered. These iterations are constructed for Dahlquist's equation (2.13) on the interval $[0, \tau]$. The Jacobian of this one-dimensional IVP has a spectrum which consists of one eigenvalue λ. Furthermore, an SDC fixed point iteration of Definition 2.10 only depends on $z = \tau\lambda$. In Chapter 4, the approach is to choose such specific z so that τ is a common time interval length and λ is usually in the spectrum of discretized partial differential operators of convection-diffusion equations. This leads to different iteration matrices $G(z)$ of SDC fixed point iterations of Definition 2.10 and we will study their properties based on the previous chapter.
This approach is presented in [55], where the convergence behavior of SDC methods for reaction-diffusion equations is covered. The following chapter contains numerical experiments of [55] in order to prepare the study of SDC methods for convection-diffusion equations in Chapter 4. The steady-state case of reaction-diffusion equations has a self adjoint partial differential operator with a real spectrum, see [10] and [15] for a theoretically analysis of such operators. A treatment of the discrete Laplace operator determining discrete reaction-diffusion systems can be found in [28]. Applying this problem class with a negative spectrum to an SDC fixed point iteration of Definition 2.10 and thereby restricting to real Jacobians $J_f = \lambda < 0$ yields $0 > \tau\lambda = z \in \mathbb{R}$. For the SDC methods, Radau-IIa nodes are chosen as the collocation discretization. These nodes are a reasonable choice because the restriction $0 > z \in \mathbb{R}$ includes $z \to -\infty$, which is the limit case of stiff problems. The L-stable Radau-IIa collocation methods have a damping property in the case of $z \to -\infty$ so that oscillations can be prevented. For further information on corresponding stability properties, the reader is referred to Section IV.3 in [32].
The first Section 3.1 provides a treatment of the asymptotic contraction factor of Definition 2.15 in the limit cases of $z \to 0$ and $z \to -\infty$. This factor is quantified by the spectral radius $\rho(G(z))$, see Lemma 2.16. The Subsection 3.1.2 covers a modification of the iteration matrix $G(z)$ for an acceleration of the convergence in the case of $z \to -\infty$.

In Section 3.2, a direct optimization approach is applied to get a faster convergence and a better local and global pre-asymptotic contraction factor of Definition 2.17 and 2.18, respectively, for other ranges of z.

3.1 SDC convergence of the limit cases

At first, the focus is on the two limit cases $z \to 0$ and $z \to -\infty$, which lead to non-stiff and stiff problems, respectively.

3.1.1 Non-stiff problems

The SDC convergence behavior of problems with $\tau \to 0$ and a fixed λ is studied, i.e., $z \to 0$ as the limit case of non-stiff problems is considered. The following result concerns the order of SDC-j methods.

Theorem 3.1. Consider an SDC fixed point iterations of Definition 2.10. Performing j iterations leads to an SDC-j method of Definition 2.7. The order p of this SDC-j method is at least $p = \min\{j, q\}$, where q is the order of the underlying quadrature rule.

Proof. Let $y^{[0]}$ be an arbitrary starting point and y_c the unique fixed point of an SDC fixed point iteration of Definition 2.10. This fixed point is the collocation solution of the considered IVP from Dahlquist's equation of Definition 2.9, see Lemma 2.11. For the solution $y^{[j]}$ of the corresponding SDC-j method, there holds

$$y^{[j]} - y_c = G\left(y^{[j-1]} - y_c\right) = \ldots = G^j\left(y^{[0]} - y_c\right)$$

and this leads to

$$\begin{aligned}
y^{[j]} - y_c = G^j\left(y^{[0]} - y_c\right) &= \left(I - \left(D - z\hat{S}\right)^{-1}(D - zS^r)\right)^j \left(y^{[0]} - y_c\right) \\
&= \left(\left(D - z\hat{S}\right)^{-1}\left(\left(D - z\hat{S}\right) - (D - zS^r)\right)\right)^j \left(y^{[0]} - y_c\right) \\
&= \left(\left(D - z\hat{S}\right)^{-1} z\left(S^r - \hat{S}\right)\right)^j \left(y^{[0]} - y_c\right) \\
&= \tau^j \left(\left(D - z\hat{S}\right)^{-1} \lambda\left(S^r - \hat{S}\right)\right)^j \left(y^{[0]} - y_c\right).
\end{aligned}$$

Thus, we get the result

$$y^{[j]} - y_c = \tau^j \left(D^{-1}\lambda\left(S^r - \hat{S}\right)\right)^j \left(y^{[0]} - y_c\right) \quad \text{as } \tau \to 0.$$

For the order of a method, see Definition 2.4, we consider the error

$$e_n^T\left(y^{[j]} - y\right) = \tau^j e_n^T \left(D^{-1}\lambda\left(S^r - \hat{S}\right)\right)^j \left(y^{[0]} - y_c\right) + e_n^T(y_c - y) \quad \text{as } \tau \to 0$$

at the end point t_n of the time interval, where $e_n^T = [0, \ldots, 0, 1] \in \mathbb{R}^n$ and y is the exact solution of the IVP we want to solve. Considering a consistent norm of this error with the triangle inequality leads to the bound

$$\left\| e_n^T \left(y^{[j]} - y \right) \right\| \leq \tau^j \left\| e_n^T \left(D^{-1} \lambda \left(S^r - \hat{S} \right) \right)^j \right\| \left\| \left(y^{[0]} - y_c \right) \right\| + \left\| e_n^T \left(y_c - y \right) \right\| \quad \text{as } \tau \to 0.$$

Lemma 2.8 yields $\left\| e_n^T \left(y_c - y \right) \right\| = \mathcal{O}(\tau^{q+1})$ if $\tau \to 0$, where q is the order of the underlying quadrature rule. Furthermore, we have $y_0^{[0]} = \gamma_0$ with γ_0 the initial value of the IVP of Dahlquist's equation and $y^{[0]}$ the starting point of the fixed point iteration. As a consequence, $\left\| y^{[0]} - y_c \right\| = \mathcal{O}(\tau)$ if we let $\tau \to 0$ and hence

$$\left\| e_n^T \left(y^{[j]} - y \right) \right\| \leq C_1 \tau^{j+1} + C_2 \tau^{q+1} \quad \text{as } \tau \to 0,$$

where C_1 and C_2 are some constants. By Definition 2.4 regarding the order of a Runge-Kutta methods, we obtain the result that the SDC-j method has at least the order $p = \min\{j, q\}$. □

This theorem gives a result for an SDC-j methods and $\tau \to 0$. Furthermore, considering SDC fixed point iteration of Definition 2.10, where $j = 1, 2, \ldots$, there holds $G(z \to 0) \to 0$ and therefore $\rho(G(z \to 0)) \to 0$, i.e., the iteration converge to the collocation solution for sufficiently small τ. Figure 3.1 provides some numerical examples for the asymptotic contraction factor for SDC methods on equidistant and Radau-IIa grids with different numbers of collocation points. These first examples are for SDC methods for Dahlquist's equation which apply the linearly implicit Euler method as low order time stepping scheme.

Definition 3.2. SDC fixed point iterations of Definition 2.10 with an approximate integration matrix $\hat{S} = I$ will be called *Eul-SDC methods*. These methods have the linearly implicit Euler method as low order time stepping scheme.

Remark 3.3. Using equidistant nodes yields a collocation discretization including the first interval point t_0. In this case, the spectral integration matrix is not reduced by the first row and column, as described in Section 2.2. Thus, for the computation of $G(z)$, a matrix $S \in \mathbb{R}^{(n+1)\times(n+1)}$ of Definition 2.1 is obtained. This leads to a new formulation of the fixed point iteration.

Remark 3.4. Similar results as in Theorem 3.1, but with other proofs, can be found in [29] and [33]. The main theorem of the first paper gives a detailed proof and information on the increasing order of SDC methods.

3.1.2 Stiff problems and the LU decomposition approach

In the following, problems with $|\lambda| \gg \tau^{-1}$ are regarded. For $\lambda \to -\infty$ and usual step sizes τ, this leads to $z \to -\infty$ the limit case of stiff problems. Using Eul-SDC methods, we obtain $G(z \to -\infty) \to I - S^r$ and therefore, in general, a non-vanishing spectral radius $\rho(I - S^r) > 0$. Figure 3.1 illustrates this result for equidistant and Radau-IIa nodes. In these experiments the SDC methods have a faster convergence on equidistant

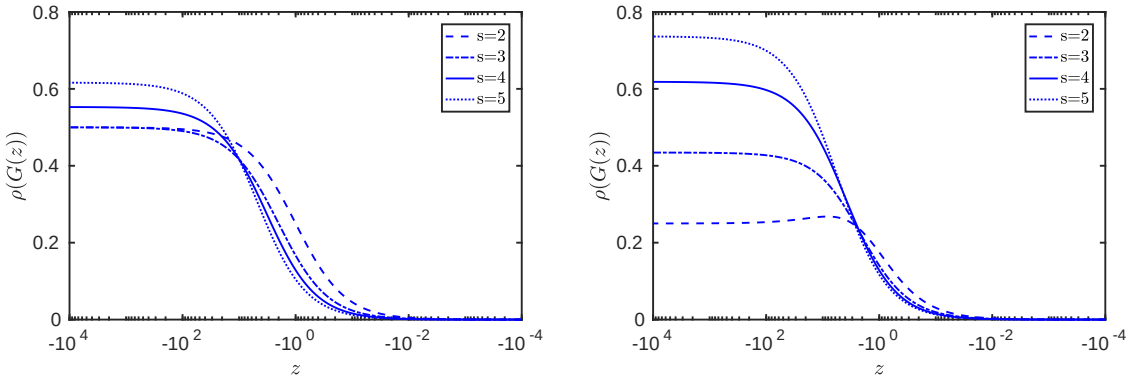

Fig. 3.1: Spectral radius $\rho(G(z))$ of linearly implicit Euler based SDC methods for Dahlquist's equation on equidistant (left) and Radau-IIa grids (right) with s collocation points.

than on Radau-IIa grids if the problem becomes very stiff and the number of collocation points is $s > 2$. Thus, in this subsection, the acceleration of the SDC convergence on Radau-IIa grids for stiff problems is covered.
As derived in Chapter 2, different low order time stepping schemes for SDC methods can be used by simply replacing the approximate integration matrix $\hat{S}$ in (2.14). The idea in [55] is to choose $\hat{S}$ first and then interpret the SDC sweeps

$$\left(D - z\hat{S}\right)\delta^{[k]} = -\left(D - zS^r\right)y^{[k]} + \frac{\tau}{\tau_1}\gamma_0\, e_1, \qquad k = 0, \ldots, j-1$$

of an SDC-j method as the steps of a Runge-Kutta method, see Subsection 2.1.2 and Definition 2.7. If the lower triangular shape of $\hat{S}$ with non-vanishing diagonal entries is retained, the matrix $D - z\hat{S}$ is lower triangular as well. Thus, SDC-j methods are obtained which are diagonally implicit Runge-Kutta methods, see Section IV.5 and IV.6 in [32] and Section 6.2 in [11]. This allows an efficient implementation where the system of equations in each SDC sweep is easier to solve. The evaluation of the iteration matrix $G(z)$ needs less computational effort compared to the use of a dense matrix $\hat{S}$, i.e., using fully implicit Runge-Kutta methods. The choice of $\hat{S}$ has no influence on the limit behavior of the fixed point iteration for $z \to 0$ and thus for the asymptotic contraction factor, there still holds $\rho(G(z \to 0)) \to 0$.
The aim is now to get a faster SDC convergence for the other limit case $z \to -\infty$. For such problems, $G(z \to -\infty) \to I - \hat{S}^{-1}S^r$ holds. As in [55], the matrix $\hat{S}$ is selected based on an LU decomposition $(S^r)^T = LU$, where L is a unit lower triangular matrix, i.e., in particular, all diagonal entries are equal to one, and U is an upper triangular matrix. Choosing $\hat{S} = U^T$ yields the limit iteration matrix

$$G(z \to -\infty) \to I - U^{-T}U^T L^T = I - L^T,$$

see Lemma 3.1 in [55]. Due to the unit diagonal of the lower triangular matrix L, the spectral radius satisfies $\rho(I - L^T) = 0$. With this choice of $\hat{S}$ it is possible to enforce a faster convergence for the limit case of stiff problems, as demonstrated in Figure 3.2. The numerical experiments further show that the asymptotic contraction factor improves for all values of z compared to the Eul-SDC methods.

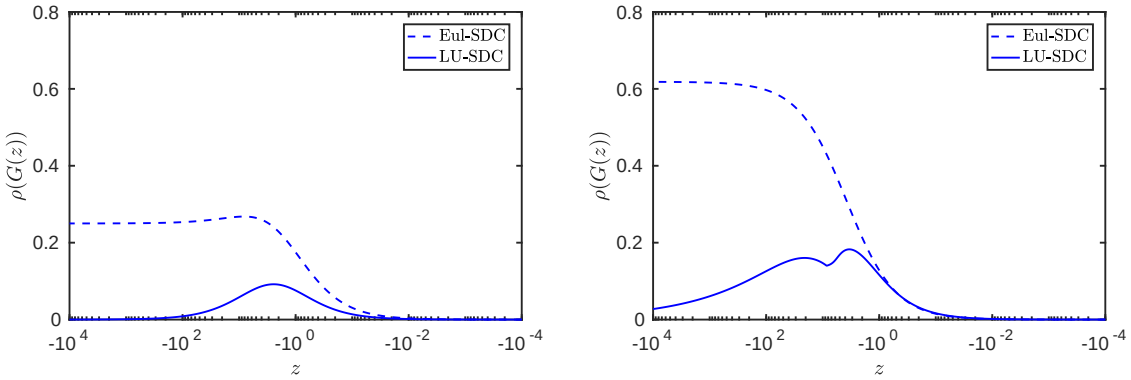

Fig. 3.2: Spectral radius $\rho(G(z))$ of Eul-SDC and LU-SDC methods on Radau-IIa grids with $s = 2$ (left) and $s = 4$ (right) collocation points.

Definition 3.5. SDC fixed point iterations of Definition 2.10 with an approximate integration matrix $\hat{S} = U^T$ where the matrix U is from the LU decomposition $(S^r)^T = LU$ will be called *LU-SDC methods*.

This special LU decomposition, introduced by Martin Weiser in [55] and colloquially known as LU trick or St. Martin's trick [4],[5],[48], has become a known approach in scientific research on SDC methods for stiff problems. Numerical experiments for the resulting LU-SDC methods demonstrate a superior convergence behavior for $z \to -\infty$. This motivates the proof of the next theorem, where we show that there always exists a unique LU decomposition of S^{rT}. This means that pivoting is never necessary for this decomposition and this is of relevance because pivoting would lead to a modification of the spectral integration matrix. This can be essentially done, but it would also lead to a modification of the whole SDC sweep structure and the resulting LU-SDC methods would not longer run forward in time, see Remark 3.1 in [55].

Theorem 3.6. Let the spectral integration matrix of Definition 2.1 be reduced by the left interval end point t_0 and let the time points $t_1, \ldots, t_n$ be collocation points. This leads to the spectral integration matrix $S^r \in \mathbb{R}^{n\times n}$, as introduced in Section 2.2. For S^{rT} exists a unique LU decomposition such that $S^{rT} = LU$, where $L \in \mathbb{R}^{n\times n}$ is a unit lower triangular matrix and $U \in \mathbb{R}^{n\times n}$ is an upper triangular matrix.

Proof. From Theorem 3.2.1 in [25] we know that there exists a unique LU decomposition for the matrix S^r if the determinants $\det(S^r(1\colon k, 1\colon k)) \neq 0$ for $k = 1, \ldots, n$. We simply write in the following S^r_k for $S^r(1\colon k, 1\colon k)$. The matrices S^r_k, where $k = 1, \ldots, n-1$, are the leading principal submatrices of S^r and $S^r_n = S^r$. The stated condition holds if and only if $S^r_k x = 0 \Leftrightarrow x = 0$ for $k = 1, \ldots, n$, where $x \in \mathbb{R}^k$, see Theorem 1.3 in [53]. The

corresponding k-th system of equations is given by

$$S_k^r x = \begin{bmatrix} \frac{1}{\tau_1}\int_{t_0}^{t_1} L_1(t)\,\mathrm{d}t & \frac{1}{\tau_1}\int_{t_0}^{t_1} L_2(t)\,\mathrm{d}t & \dots & \frac{1}{\tau_1}\int_{t_0}^{t_1} L_k(t)\,\mathrm{d}t \\ \frac{1}{\tau_2}\int_{t_1}^{t_2} L_1(t)\,\mathrm{d}t & \frac{1}{\tau_2}\int_{t_1}^{t_2} L_2(t)\,\mathrm{d}t & \dots & \frac{1}{\tau_2}\int_{t_1}^{t_2} L_k(t)\,\mathrm{d}t \\ \vdots & \vdots & & \vdots \\ \frac{1}{\tau_k}\int_{t_{k-1}}^{t_k} L_1(t)\,\mathrm{d}t & \frac{1}{\tau_k}\int_{t_{k-1}}^{t_k} L_2(t)\,\mathrm{d}t & \dots & \frac{1}{\tau_k}\int_{t_{k-1}}^{t_k} L_k(t)\,\mathrm{d}t \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

and we show with a proof by contradiction that

$$\frac{1}{\tau_i}\int_{t_{i-1}}^{t_i} \left(\sum_{j=1}^{k} x_j L_j(t)\right) \mathrm{d}t = 0, \qquad i = 1, \dots, k \tag{3.1}$$

holds if and only if $x_1 = \ldots = x_k = 0$ for all $k = 1, \dots, n$.
Let $x \neq 0$ and $k \in \{1, \dots, n\}$ where $x \in \mathbb{R}^k$. Then, the interpolation polynomial $p(t) = \sum_{j=1}^{k} x_j L_j(t)$ is of degree $n-1$, which follows immediately from the definition of the Lagrange basis polynomials (2.4). From the fundamental theorem of algebra follows that this polynomial has $n-1$ roots. The Lagrange basis polynomials $L_1 \dots, L_k$ are by construction zero at $t_{k+1}, \dots, t_n$ and thus the interpolation polynomial $p(t)$ has also roots at $t_{k+1}, \dots, t_n$. From this follows that $p(t)$ has at most $k-1$ roots in the interval (t_0, t_k). Furthermore, if the i-th equation of (3.1) holds, then the interpolation polynomial $p(t)$ has a root in the interval (t_{i-1}, t_i). Assuming without loss of generality that the equations (3.1) hold for $i = 1, \dots, k-1$, then $p(t)$ has $k-1$ roots in (t_0, t_{k-1}). If this is the case, then $p(t)$ has no root in (t_{k-1}, t_k) and thus the k-th equation of (3.1) can not be satisfied by an $x \neq 0$.
This result and that $x = 0 \Rightarrow S_k^r x = 0$ with $x \in \mathbb{R}^k$, which is trivial, lead to the consequence that $S_k^r x = 0 \Leftrightarrow x = 0$ for $k = 1, \dots, n$, where $x \in \mathbb{R}^k$. Thus, S^r has a unique LU decomposition. From Subsection 2.1.6 in [25] we know that $\det(S_k^r) = \det\left(S_k^{rT}\right)$. Therefore, $\det\left(S_k^{rT}\right) \neq 0$ for $k = 1, \dots, n$ and it follows that the matrix S^{rT} has also a unique LU decomposition. □

3.2 Direct optimization for a faster contraction

In the previous subsection, an approximate integration matrix $\hat{S}$ is selected to accelerate the convergence of SDC methods for the stiff problems. The next reasonable step is now to choose different matrices $\hat{S}$ to achieve another desirable behavior of SDC fixed point iterations. We can even go a step further away from the classical SDC framework. The equations (2.16) show the possibility to replace not only the matrix $\hat{S}$ as before, but also take a matrix $\hat{D} - z\hat{S}$ with free selectable matrices $\hat{D}$ and $\hat{S}$. Regarding SDC methods as Runge-Kutta methods gives the same insight because replacing $\hat{D}$ and $\hat{S}$ yields a different Butcher tableau, but the same collocation solution. Thus, we obtain a

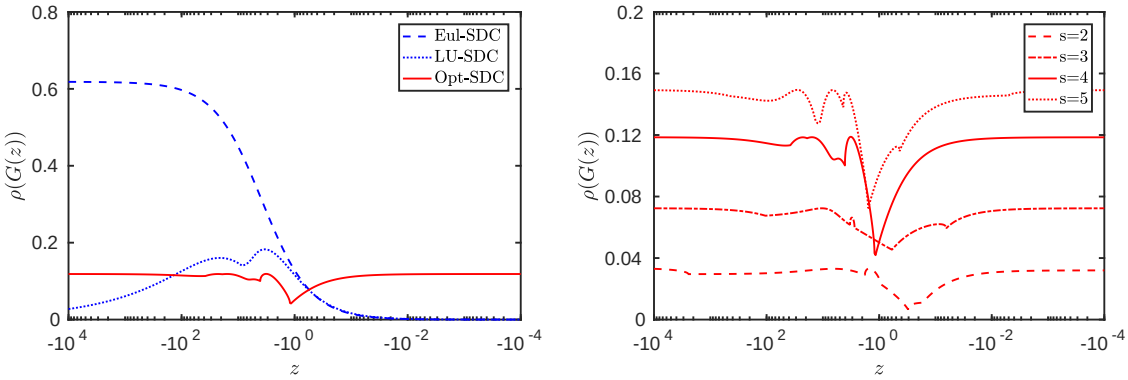

Fig. 3.3: Spectral radius $\rho(G(z))$ of Eul-SDC, LU-SDC and Opt-SDC methods on Radau-IIa grids with $s = 4$ collocation points (left) and Opt-SDC methods on Radau-IIa grids with different numbers s of collocation points (right). Opt-SDC methods based on the optimization problem (3.3).

more flexible formulation

$$\begin{aligned} y^{[j+1]} &= G(z)\, y^{[j]} + g, \\ G(z) &= \left[I - \left(\hat{D} - z\hat{S} \right)^{-1} (D - zS^r) \right], \quad g = \left(\hat{D} - z\hat{S} \right)^{-1} \frac{\tau}{\tau_1} \gamma_0\, e_1 \end{aligned} \tag{3.2}$$

for SDC fixed point iteration of Definition 2.10. As in the previous subsection, we want to interpret the SDC-j methods as diagonally implicit Runge-Kutta methods and therefore $\hat{D} - z\hat{S}$ is restricted to be a lower triangular matrix. For a certain starting point $y^{[0]}$, the choice of $\hat{D} - z\hat{S}$ determines the behavior of the fixed point iterations.
One possibility to select $\hat{D}$ and $\hat{S}$ is their optimization with respect to convergence objectives. The reduction of the maximum of $\rho(G(z))$ for $z < 0$ is an example for such an objective. The following subsections present results of numerical experiments of such optimizations for the parameters $\hat{D}$ and $\hat{S}$, as in [55], thereby demonstrating possible improvement. Due to the ease of use, we optimize $\hat{D}$ and $\hat{S}$ with the nonlinear programming solver *fminsearch* from MATLAB®. This minimizer works with a derivative-free method. The objective functions are computed on a logarithmic grid for $z \in [-10^4, -10^{-4}]$ with 100 points and with initial matrices from the LU decomposition approach, as described in Subsection 3.1.2. The resulting minima are not necessarily global minima.

Definition 3.7. SDC fixed point iterations of Definition 2.10 which are modified as in (3.2) and where $\hat{D}$ and $\hat{S}$ are the result of a direct optimization approach will be called *Opt-SDC methods*.

3.2.1 Asymptotic contraction factor

For LU-SDC methods, a small spectral radius can be observed for non-stiff and stiff problems, see Figure 3.2. The new objective is to get a smaller $\rho(G(z))$ for intermediate values of z and this leads to the optimization problem

$$\min J(\hat{D}, \hat{S}) := \max_{z \leq 0} \rho(G(z)). \tag{3.3}$$

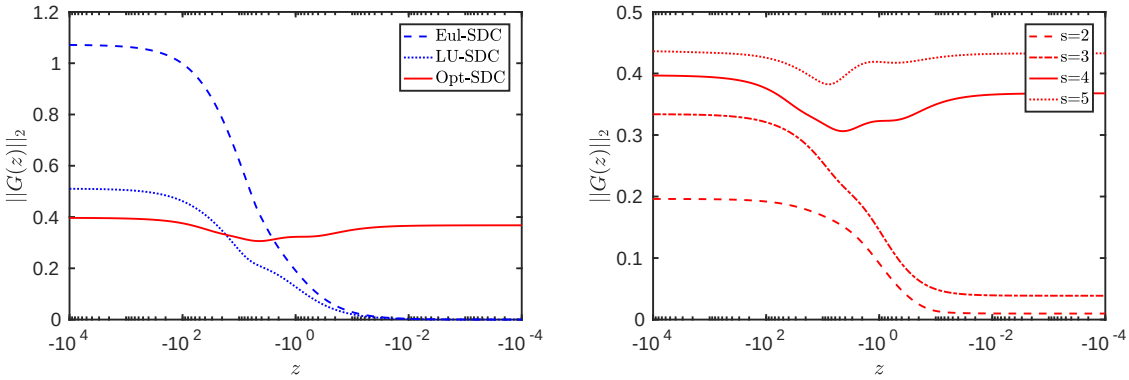

Fig. 3.4: Norm $\|G(z)\|_2$ of Eul-SDC, LU-SDC and Opt-SDC methods on Radau-IIa grids with $s = 4$ collocation points (left) and Opt-SDC methods on Radau-IIa grids with different numbers s of collocation points (right). Opt-SDC methods based on the optimization problem (3.4).

Considering $s = 4$ collocation points, this local optimization already yields a visible reduction of the maximum of $\rho(G(z))$ compared to LU-SDC methods, see Figure 3.3 (left). The trade-off for this improvement is a worse convergence speed in the limit cases $z \to 0$ and $z \to -\infty$. The results for different numbers of collocation points are presented in Figure 3.3 (right), showing that more collocations points lead to a higher asymptotic contraction factor. The same behavior can be found in Figure 3.1 for Eul-SDC methods and in Figure 3.2 for LU-SDC methods.

3.2.2 Local pre-asymptotic contraction factor

In this subsection, the local pre-asymptotic contraction factor of Definition 2.17 is covered and the objective is again to reduce its maximum. To this end, we consider the optimization problem

$$\min J(\hat{D}, \hat{S}) := \max_{z \leq 0} \|G(z)\| \tag{3.4}$$

with a matrix norm $\|\cdot\|$. Numerical experiments for the 2-norm are presented in Figure 3.4. The norm $\|G(z)\|_2$ for the iteration matrix of Opt-SDC, LU-SDC and Eul-SDC methods is shown in the left plot. Compared to the original implicit Euler approach, a significant improvement by the LU decomposition is observed. The optimization approach leads to an additional reduction of the maximum of $\|G(z)\|_2$, but $z \to 0$ and intermediate values of z yield a worse local pre-asymptotic contraction factor than for the LU-SDC or Eul-SDC methods. For two or three collocation points, the optimized matrices $\hat{D}$ and $\hat{S}$ of the Opt-SDC methods are nearly equal to that from the LU-SDC methods, see Figure 3.4 (right). In that case, the used optimizer yields no remarkable reduction of the maximum of $||G(z)||_2$. Note that another setting of the optimizer or an entirely different programming solver could give greater improvement at this point.

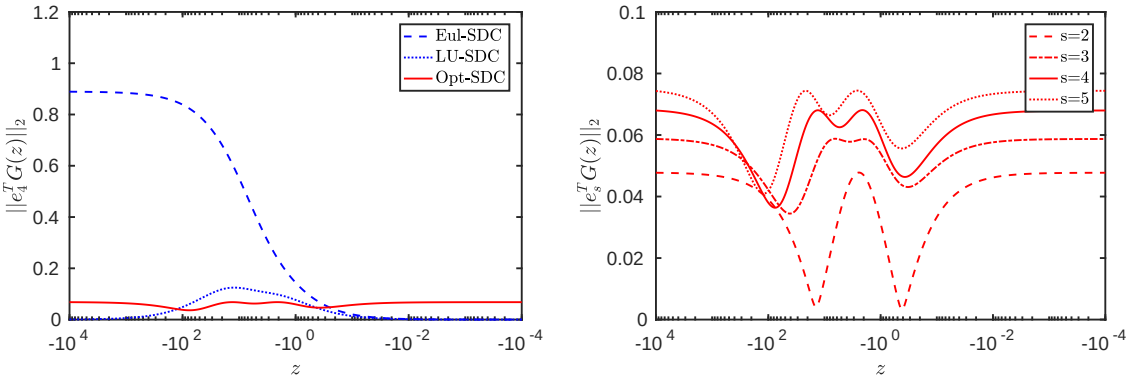

Fig. 3.5: Norm $\left\|e_s^T G(z)\right\|_2$ of Eul-SDC, LU-SDC and Opt-SDC methods on Radau-IIa grids with $s = 4$ collocation points (left) and Opt-SDC methods on Radau-IIa grids with different numbers s of collocation points (right). Opt-SDC methods based on the optimization problem (3.5).

3.2.3 Global pre-asymptotic contraction factor

The next objective is to get a smaller maximum of the global pre-asymptotic contraction factor of Definition 2.18. This leads to the optimization problem

$$\min J(\hat{D}, \hat{S}) := \max_{z \leq 0} \left\|e_n^T G(z)\right\| \tag{3.5}$$

with a matrix norm $\|\cdot\|$ and the n-th cartesian unit vector $e_n = [0, \ldots, 0, 1]^T \in \mathbb{R}^n$. Due to the collocation with s Radau-IIa nodes, there holds $n = s$. Experiments for the 2-norm are presented in Figure 3.5. In the left plot, Opt-SDC, LU-SDC and Eul-SDC methods for $s = 4$ collocation points can be compared. Several Opt-SDC methods with different numbers of collocation points are shown in the right plot. Compared to the Eul-SDC methods, the LU decomposition approach leads to a vastly improved global pre-asymptotic contraction factor for very stiff problems and for intermediate values of z with no disadvantage for $z \to 0$. The Opt-SDC methods have a lower maximum of $\left\|e_n^T G(z)\right\|_2$, but again the beneficial properties in the limit cases $z \to 0$ and $z \to -\infty$ are sacrificed. As observed before for the other contraction factors, a higher number of collocation points leads to a higher global pre-asymptotic contraction factor.

3.2.4 Sweep blocks

The aim is now to obtain matrices $\hat{D}$ and $\hat{S}$ which are optimal with respect to the resulting approximation after a certain number m of successively performed SDC sweeps, in the following referred to as *SDC sweep blocks*. For this, we formulate the optimization problem

$$\min J(\hat{D}, \hat{S}) := \max_{z \leq 0} \left\|e_n^T G(z)^m\right\|^{1/m}. \tag{3.6}$$

The objective is the same as for the problem (3.5) and the new relevant error reduction for a certain number of SDC sweeps leads to other optimal matrices $\hat{D}$ and $\hat{S}$. In Figure

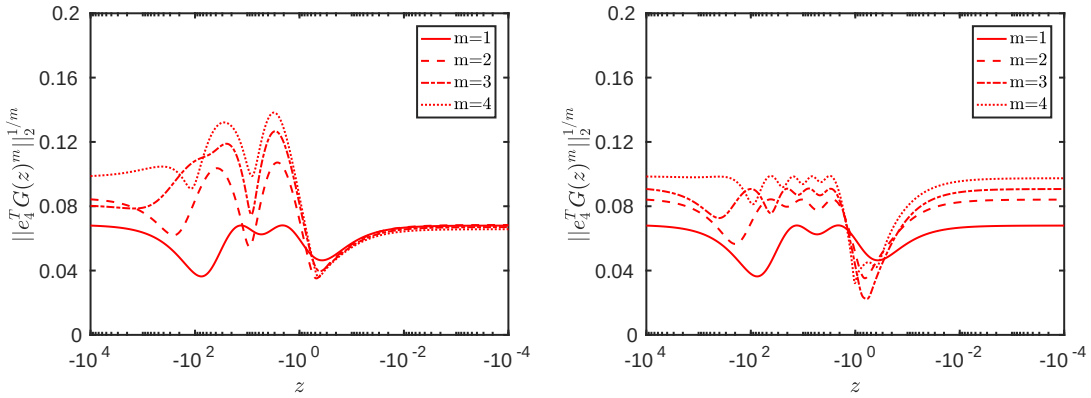

Fig. 3.6: Norm $\left\|e_4^T G(z)^m\right\|_2^{1/m}$ for Opt-SDC methods on Radau-IIa grids with $s = 4$ collocation points. Opt-SDC methods based on the optimization problem (3.5) (left) and (3.6) (right).

3.6 (left), we first optimize the approximate matrices $\hat{D}$ and $\hat{S}$ for one SDC sweep as in problem (3.5) and after this form the product of the iteration matrix $G(z)$. This represents the execution of several SDC sweeps with optimization with respect to just one SDC sweep. In Figure 3.6 (right), the sweep block size m is first determined and then the product $\left\|e_n^T G(z)^m\right\|^{1/m}$ is optimized.

It can be observed that there is better global error reduction per SDC sweep if the number of sweeps is given and optimization is done with respect to an approximation after these sweeps. Furthermore, increasing m leads to a worse global error reduction per sweep. The reason for this observation is that the consideration of just few SDC sweeps give deceptive results. By considering just one sweep, the values in the rows $1, \ldots, n-1$ of $G(z)$ have no influence on the global error at t_n and thus they can be chosen arbitrarily. But for increasing SDC sweeps, a worse error reduction for $t_0, \ldots, t_{n-1}$ leads to increasingly larger errors at t_n because of the iteration process with a product of iteration matrices. For additional numerical experiments to SDC sweep blocks for the local and global pre-asymptotic contraction factor, we refer to [55].

This chapter can be summarized as follows: The modification of the iteration matrix $G(z)$ for $0 > z \in \mathbb{R}$ is promising. With the LU decomposition approach it is possible to accelerate the SDC convergence significantly and to improve the local and global pre-asymptotic contraction factors compared to the original linearly implicit Euler approach. Furthermore, numerical experiments show that the direct optimization approach can lead to an improvement for a specific range of z.

4 Convection-diffusion equations

In the previous chapter, the strategy to obtain special convergence properties of SDC fixed point iterations of Definition 2.10 was carried out for a special class of PDEs, namely reaction-diffusion equations. In Section 4.4, the approaches are studied for convection-diffusion equations. Precedingly, in Section 4.1, a physical background is presented, mainly based on [19] and [44]. Two examples for common spatial discretization techniques are regarded in Section 4.2. Hereafter, in Section 4.3, the spectral properties of a simple one-dimensional convection-diffusion problem are covered.

4.1 Physical background and typical problems

Convection-diffusion equations are PDEs that arise in numerous physical problems, including the large area of modeling flows. In this work, we cover the convection-diffusion equation

$$\frac{\partial u}{\partial t} = \alpha \Delta u - w \cdot \nabla u + r \tag{4.1}$$

with the common simplifications of an incompressible vector field $w(x,t) \in \mathbb{R}^N$, i.e., it is divergence free with $\nabla \cdot w = 0$, and a constant diffusion coefficient $\alpha \in \mathbb{R}^+$. The function $u(x,t) \in \mathbb{C}$ is the solution at position $x \in \Omega \subseteq \mathbb{R}^N$ and time $t \in [0,T]$. The vector field w, also called the wind, describes a velocity that the solution is moving with and the function $r(u,x,t) \in \mathbb{C}$ can be interpreted as sources or sinks of the variable of interest. The term $\alpha \Delta u$, with the Laplace operator $\Delta = \nabla^2$, describes the diffusion and $w \cdot \nabla u$ the convection or advection of the problem. A standard example for a function that satisfies equation (4.1) is a concentration of a pollutant which is moving within an incompressible stream and diffusing into its environment. Considering the heat equation with the temperature of a fluid as the variable of interest and assuming that this fluid is moving yields another convection-diffusion problem. If we are interested in the physics of a vector-valued flow velocity of a moving fluid where the fluid is a Newtonian fluid, i.e., it has a linearly viscous behavior, then we have to deal with the Navier-Stokes equations. Convection and diffusion terms are important parts of them. The diffusive effects arise from the viscosity and the diffusion coefficient α becomes the kinematic viscosity of the fluid. Considering the Navier-Stokes equations for incompressible fluids, the incompressible vector field w from equation (4.1) becomes the flow velocity of the fluid itself. Due to the ability of Navier-Stokes equations to cover the important phenomena of boundary layers and turbulence, they have large areas of applications, for example, the modeling of the flow around an airplane, the weather or the flow of water in a pipeline. The reader is referred to [3] for an introduction to fluid dynamics and to its Section 5.7 for a physical treatment of boundary layers. Chapter 12 of the textbook [50] gives an application-oriented introduction to boundary layers.

In many physical problems, the effect of convection is more significant than the effect of diffusion, thus implying $\alpha \ll \|w\|$. However, for the following applications, we want to introduce, as in Chapter 6 from [19], a more meaningful measure of the quantitative relationship between these two mechanisms. For this, the steady-state version of (4.1), i.e., $\partial u/\partial t = 0$, is considered and normalized. If $x \in \Omega \subseteq \mathbb{R}^N$ are elements of the domain where the solution $u(x)$ lives in, then let $\xi = x/L$ be elements of a normalized domain. For this new coordinate ξ, the new functions $u_*(\xi) = u(x), w_*(\xi) = (1/W)w(x)$ and $r_*(\xi) = (L^2/\alpha)r(x)$ are defined. The measure L denotes a characterizing length scale for the domain Ω and $W \in \mathbb{R}^+$ is a constant, such that $\|w_*\|$ has the value unity in some norm $\|\cdot\|$. This leads to the convection-diffusion equation

$$\begin{aligned} & 0 = \alpha\Delta_x u_*(\xi) - W w_*(\xi) \cdot \nabla_x u_*(\xi) + \frac{\alpha}{L^2} r_*(\xi) \\ \Rightarrow \quad & 0 = \frac{\alpha}{L^2}\Delta_\xi u_*(\xi) - \frac{W}{L} w_*(\xi) \cdot \nabla_\xi u_*(\xi) + \frac{\alpha}{L^2} r_*(\xi) \\ \Rightarrow \quad & 0 = \Delta_\xi u_*(\xi) - \frac{WL}{\alpha} w_*(\xi) \cdot \nabla_\xi u_*(\xi) + r_*(\xi) \end{aligned}$$

for a normalized domain, see equation (6.5) in [19]. We now have all information about convection and diffusion in a dimensionless number

$$\mathrm{Pe} := \frac{WL}{\alpha}, \tag{4.2}$$

which is called the *Peclet number*. It provides a simple measure of distinction between the diffusion dominated case of $\mathrm{Pe} \leq 1$ and the convection dominated case of $\mathrm{Pe} \gg 1$ and allows to select appropriate methods accordingly.
Convection-diffusion equations combine both parabolic and hyperbolic PDEs. We refer to Chapter 1 in [28] for an introduction to the classification of PDEs and to [21] for a theoretical treatment of PDEs. When diffusive effects are dominating as $\mathrm{Pe} \to 0$, the convection term of equation (4.1) can be neglected and thus the second order parabolic PDE

$$\frac{\partial u}{\partial t} = \alpha\Delta u + r$$

is obtained. On the other hand, if $\mathrm{Pe} \to \infty$, i.e., convection is much more significant, the diffusion term can be neglected and this yields the first order hyperbolic PDE

$$\frac{\partial u}{\partial t} = -w \cdot \nabla u + r,$$

which is the linear transport equation. This equation leads to the simplest example for the meaning of convection. For this, consider a corresponding homogeneous IVP

$$\begin{aligned} \frac{\partial u}{\partial t} + w \cdot \nabla u = 0 \qquad & \text{in } \mathbb{R} \times (0, \infty), \\ u(x, 0) = g(x) \qquad & \text{at } \mathbb{R} \times 0 \end{aligned} \tag{4.3}$$

where $w \in \mathbb{R}$ is constant. It is easy to see that the function $u : \mathbb{R} \times [0, \infty) \to \mathbb{R}$, $u(x, t) = g(x - tw)$ satisfies the problem (4.3) for an at least once differentiable function $g : \mathbb{R} \to \mathbb{R}$. This solution can be interpreted as follows: Some initial data g will

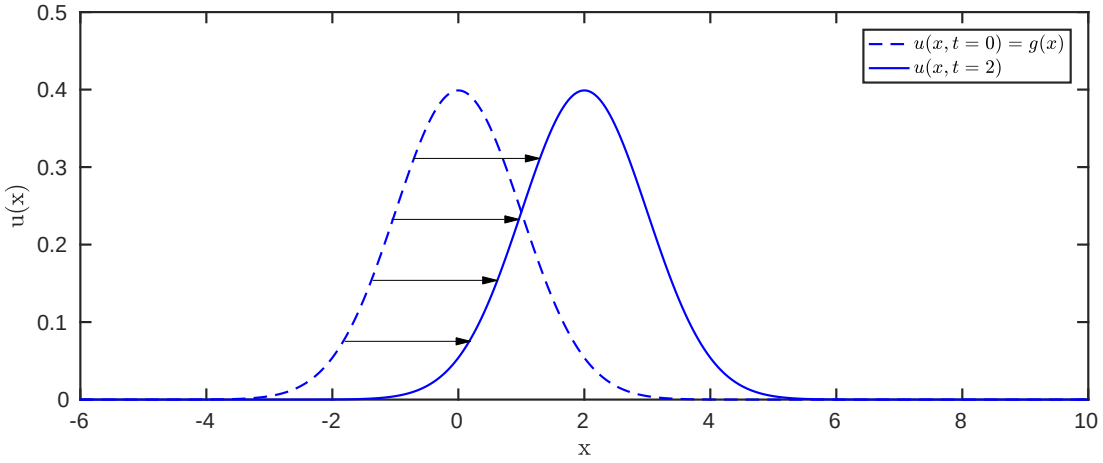

Fig. 4.1: Solution of the linear transport equation for a probability density function of a pollutant in a stream at the time $t = 0$ and $t = 2$.

be transported in time in the positive x-direction with the velocity w. Now, the example from above is picked up and a pollutant without diffusion in a one-dimensional stream which flows in the positive x-direction is considered. The variable of interest could be a relative quantity of this pollutant. For this, a probability density function $g(x) = \left(1/\left(\sigma\sqrt{2\pi}\right)\right) \exp\left(-(x-\mu)^2/(2\sigma^2)\right)$ of a Gaussian distribution is considered as the initial value. A standard deviation $\sigma = 1$ and a mean $\mu = 0$ lead to the solution $u(x,t) = \left(1/\sqrt{2\pi}\right) \exp\left(-(x-tw)^2/2\right)$. Measuring the time in seconds and the distance of the moving pollutant in meter yields a movement of two meter downstream within two seconds with a velocity of $w = 1\,\mathrm{m/s}$. Figure 4.1 illustrates this convection effect.

4.2 Common spatial discretizations

This section presents a short overview of finite difference methods and the finite element method, which can be used for the numerical solution of boundary value problems for PDEs. Later, their discretization techniques are used to semi-discretize the considered convection-diffusion operators in space.

4.2.1 Finite difference methods

The following considerations are mainly based on [28]. The basic idea of finite difference methods is to approximate partial derivatives with finite differences. To explain this idea, the equation (4.1) in a one-dimensional and steady-state case is considered. This leads to the convection-diffusion boundary value problem

$$-\alpha u''(x) + w u'(x) = r(x) \quad \text{in } (0, \delta), \qquad u(0) = u(\delta) = 0 \tag{4.4}$$

with homogeneous Dirichlet boundary conditions and applying Taylor's theorem around the point x yields

$$u(x+h) = u(x) + h\,u'(x) + \frac{h^2}{2}u''(x) + \frac{h^3}{6}u'''(x) + \mathcal{O}(h^4),$$
$$u(x-h) = u(x) - h\,u'(x) + \frac{h^2}{2}u''(x) - \frac{h^3}{6}u'''(x) + \mathcal{O}(h^4).$$

The addition of these two equations leads to

$$u''(x) = \frac{1}{h^2}\left(u(x-h) - 2u(x) + u(x+h)\right) + \mathcal{O}(h^2).$$

Thus, we get an approximation of the second derivative by

$$u''(x) \approx \frac{1}{h^2}\left(u(x-h) - 2u(x) + u(x+h)\right).$$

Furthermore, this approach can be used to derive three different possibilities to approximate the first derivative, which are given by

$$\begin{aligned}
&\textit{central differences} && u'(x) \approx \frac{1}{2h}\left(u(x+h) - u(x-h)\right),\\
&\textit{forward differences} && u'(x) \approx \frac{1}{h}\left(u(x+h) - u(x)\right),\\
&\textit{backward differences} && u'(x) \approx \frac{1}{h}\left(u(x) - u(x-h)\right).
\end{aligned}$$

Based on these considerations, the domain $(0, \delta)$ is discretized so that the derivatives can be approximated at a finite number of $N-2$ points. The set of these points $x_1 < x_2 < \ldots < x_{N-2}$ with the boundary $x_0 = 0$ and $x_{N-1} = \delta$ is called the grid or mesh and $h_i = x_i - x_{i-1}$ is a so-called grid size. Such a discretization and the approximation of the derivatives result in a linear system. By using a backward difference approximation and an equidistant grid, i.e., a constant grid size h, we obtain, for example,

$$\left(-\frac{\alpha}{h^2}\begin{bmatrix} -1 & 0 & 0 & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & 0 & 0 & -1 \end{bmatrix} + \frac{w}{h}\begin{bmatrix} 0 & 0 & 0 & & \\ -1 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 1 & 0 \\ & & 0 & 0 & 0 \end{bmatrix}\right)\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{bmatrix} = \begin{bmatrix} (\alpha/h^2)\,u_0 \\ r_1 \\ \vdots \\ r_{N-2} \\ (\alpha/h^2)\,u_{N-1} \end{bmatrix}.$$

With such a linear system the discretization is complete and solving it with an appropriate solver leads to a numerical solution of (4.4), given by $u_i \approx u(x_i)$, where $i = 1, \ldots, N-2$. For this, the common compact notation is

$$L_h u_h = r_h,$$

where $L_h \in \mathbb{R}^{N\times N}$ and $r_i = r(x_i)$ for $i = 1, \ldots, N-2$.
It is straightforward to apply this framework to other types of PDEs or boundary conditions, which is subject of almost all introductory textbooks to the numerical treatment of PDEs. There is a common approach for the numerical solution of steady-state

convection-diffusion equations with finite difference methods. Depending on the sign of the wind w, i.e., its direction, reasonable choices for the approximation of the first derivative are either forward or backward differences. If the sign is positive in (4.4), the point $x - h$ is at the upstream side of x and thus the approximation of $u'(x)$ with backward differences provides a benefit. The transport of information has the direction of the wind and backward differences consider then the information of the convection effect. If the sign of the wind is negative, forward differences are a reasonable choice. This approach is called *upwind scheme* in the literature. In Section 5.1, it is applied for numerical experiments.
Finite difference methods are a popular choice for the discretization of PDEs. Above all, their implementation is straightforward and their theoretical foundation is easy to understand. The idea of a one-dimensional discretization can often be extended to a higher dimension. Consistency estimates and stability bounds are easily derived with Taylor series, although this leads to high requirements on the smoothness of the solution. Furthermore, general domains can reduce the order of consistency and convergence and adding a new point for the discretization can cause a larger effort than for other discretization techniques. The reader is referred to [28] for further information on finite difference methods.

4.2.2 Finite element method

The following introduction to the finite element method is based on [13], [19] and [56] and we refer to [59] for a detailed treatment of the finite element fundamentals. One disadvantage of the formulation (4.4) with a second order PDE is the consideration of an at least twice differentiable function $u(x)$.
To handle this problem, we introduce the concept of weak solutions. For this, consider the two-dimensional steady-state case of (4.1) with the boundary value problem

$$\begin{aligned} -\alpha\Delta u + w\cdot\nabla u &= r \quad \text{in } \Omega, \\ u = g_D \quad \text{on } \partial\Omega_D \quad \text{and} \quad \frac{\partial u}{\partial n} &= g_N \quad \text{on } \partial\Omega_N, \end{aligned} \tag{4.5}$$

where $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ is the boundary of the domain $\Omega \subset \mathbb{R}^2$. There are Dirichlet boundary conditions on $\partial\Omega_D$ and Neumann boundary conditions on $\partial\Omega_N$. The directional derivative $\partial u/\partial n$ is in the normal direction of the boundary $\partial\Omega_N$. A function u which satisfies this problem is called a classical solution. Multiplying the PDE with test functions v and integrating yields

$$\int_\Omega (\alpha\Delta u - w \cdot \nabla u + r)\, v \,\mathrm{d}x = 0. \tag{4.6}$$

Integration by parts and the divergence theorem lead to

$$\begin{aligned} -\alpha \int_\Omega v\Delta u \,\mathrm{d}x &= \alpha \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \alpha \int_\Omega \nabla \cdot (v\nabla u)\, \mathrm{d}x \\ &= \alpha \int_\Omega \nabla u \cdot \nabla v \,\mathrm{d}x - \alpha \int_{\partial\Omega} v\frac{\partial u}{\partial n}\,\mathrm{d}x, \end{aligned} \tag{4.7}$$

see Section 1.2 in [19]. For the following considerations, the choice of the test functions v is essential. If the functions u and v live in the Sobolev space

$$\mathcal{H}^1(\Omega) := \left\{ u : \Omega \to \mathbb{R} : u, \frac{\partial u}{\partial x}, \frac{\partial u}{\partial y} \in L_2(\Omega) \right\},$$

where $L_2(\Omega)$ is the Lebesgue space defined by

$$L_2(\Omega) := \left\{ u : \Omega \to \mathbb{R} : \int_\Omega u^2 \, \mathrm{d}x < \infty \right\}, \tag{4.8}$$

then the integral $\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x$ of equation (4.7) is well defined. Furthermore, if $r \in L_2(\Omega)$ and $g_N \in L_2(\partial\Omega_N)$, the integrals $\int_\Omega vr \, \mathrm{d}x$ and $\int_{\partial\Omega_N} vg_N \, \mathrm{d}x$ are also well defined. Let the solution and test spaces be defined by

$$\begin{aligned} \mathcal{H}^1_E &:= \left\{ u \in \mathcal{H}^1(\Omega) : u = g_D \quad \text{on } \partial\Omega_D \right\}, \\ \mathcal{H}^1_{E_0} &:= \left\{ v \in \mathcal{H}^1(\Omega) : v = 0 \quad \text{on } \partial\Omega_D \right\}, \end{aligned}$$

respectively. For correctness, the solution space $\mathcal{H}^1_E$ is no linear space because it is not closed under addition. These definitions and substituting (4.7) into (4.6) lead to the weak formulation as follows: Find a function $u \in \mathcal{H}^1_E$ such that

$$\alpha \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \int_\Omega (w \cdot \nabla u) \, v \, \mathrm{d}x = \int_\Omega vr \, \mathrm{d}x + \alpha \int_{\partial\Omega_N} vg_N \, \mathrm{d}x \tag{4.9}$$

for all $v \in \mathcal{H}^1_{E_0}$, see (6.11) in [19]. This function u is a weak solution of the problem (4.5). Compared to a classical solution, the required smoothness of a weak solution is reduced. A short notation of (4.9) is given by

$$a(u, v) = b(v) \quad \forall v \in \mathcal{H}^1_{E_0}(\Omega), \tag{4.10}$$

where

$$a : \mathcal{H}^1(\Omega) \times \mathcal{H}^1(\Omega) \to \mathbb{R}, \quad a(u, v) = \alpha \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x + \int_\Omega (w \cdot \nabla u) \, v \, \mathrm{d}x$$

is an unsymmetric bilinear form and

$$b : \mathcal{H}^1(\Omega) \to \mathbb{R}, \quad b(v) = \int_\Omega vr \, \mathrm{d}x + \alpha \int_{\partial\Omega_N} vg_N \, \mathrm{d}x$$

is a linear functional.
In this work, we are interested in the numerical solution of PDEs. To this end, a finite-dimensional trial space $V_{h,E} \subset \mathcal{H}^1_E$ and test space $V_{h,0} \subset \mathcal{H}^1_{E_0}$ is considered. This leads to the discrete weak formulation:

$$\text{Find a } u_h \in V_{h,E} \text{ such that} \quad a(u_h, v_h) = b(v_h) \quad \forall v_h \in V_{h,0}. \tag{4.11}$$

The restriction of (4.10) to finite dimensional spaces is called Galerkin discretization, see Section 1.3 in [19] and Section 1.4 in [56].

Now, let $V_{h,0}$ be an N-dimensional vector space of test functions and the functions $(\varphi_i)_{1\leq i\leq N}$ the basis of this space. As in Section 1.3 in [19], this basis is extended by the functions $(\hat{\varphi}_i)_{1\leq i\leq N_\delta}$ so that g_D is interpolated by $\sum_{j=1}^{N_\partial}\delta_j\hat{\varphi}_j$ on the boundary $\partial\Omega_D$, where $\delta_{1\leq i\leq N_\partial}$ are suitable coefficients. With the coefficient vector $u_h^\varphi\in\mathbb{R}^N$ we obtain

$$u_h=\sum_{j=1}^{N}(u_h^\varphi)_j\varphi_j+\sum_{j=1}^{N_\partial}\delta_j\hat{\varphi}_j. \tag{4.12}$$

This representation of $u_h\in V_{h,E}$ and the discrete weak formulation (4.11) lead to the system of linear equations

$$L_hu_h^\varphi=b_h^\varphi \tag{4.13}$$

with

$$\begin{aligned}(L_h)_{ij}&=a(\varphi_j,\varphi_i),\qquad i,j=1,\dots,N,\\ (b_h^\varphi)_i&=b(\varphi_i)-\alpha\sum_{j=1}^{N_\partial}\delta_j\int_\Omega\nabla\hat{\varphi}_j\cdot\nabla\varphi_i\,\mathrm{d}x,\qquad i=1,\dots,N.\end{aligned}$$

The linear system (4.13) is called Galerkin system und the solution u_h, defined by (4.12) with u_h^φ resulting from (4.13), is called Galerkin solution.
The last step for the derivation of the finite element method, which is also the reason for its name, is the construction of the finite-dimensional trial space $V_{h,E}$ and test space $V_{h,0}$ with the corresponding basis $(\varphi_i)_{1\leq i\leq N}$. Different spaces and bases could be used, for example, polynomial and trigonometric functions, radial basis functions or finite elements, see Chapter 1 in [56]. Finite elements are piecewise polynomial functions which are continuous in the domain Ω. One of the main ideas of the finite element method is that for the linear system (4.13), the matrix L_h should be extremely sparse. This reduces the computational costs in the step of solving the linear system. To achieve this, the domain Ω is decomposed in, for example, triangles or rectangles and the basis functions are constructed such that they have an as small as possible support in the domain Ω. For more information on the mesh generation for the finite element method, we refer to [23].
One specific approach for solving convection-diffusion problems with the finite element method is the *streamline diffusion method*. The solution resulting from (4.13) can be unsatisfactory if the mesh does not resolve boundary layers. The streamline diffusion method can be seen as a parameterized weak formulation for convection-diffusion equations, see (6.40) in [19]. A further term is added to the weak formulation (4.9) and depending on the element, this weighted term introduces additional diffusion in the direction of the wind. The streamline diffusion method can be interpreted as an upwind scheme for the finite element method, see Section 9.2 in [38]. The Peclet number is used to choose the weights and thus, this method takes into account the fact that the solution u depends mainly on its behavior along the streamlines if the Peclet number is large. We refer to the Subsection 6.3.2 in [19] or Section 9.2 in [38] for a detailed derivation of the streamline diffusion method.

The *mesh Peclet number*

$$\mathrm{Pe}_h := \frac{Wh}{\alpha} \tag{4.14}$$

plays an important role for this derivation and in general for the numerical treatment of convection-diffusion equations. As before, α is the diffusion coefficient, $W \in \mathbb{R}^+$ describes the wind in some sense, see 4.2, and h is the grid size. As an example, consider the one-dimensional problem (4.4) with a constant $W = |w|$. Then, the mesh Peclet number describes the relation between the numerical diffusion, which depends on h, and the real physical diffusion depending on α. To achieve reasonable accurate solutions, the numerical diffusion has to be small compared to the real diffusion and thus Pe_h has to be small.
In Section 6.4 in [19], error bounds can be found which use this mesh Peclet number. Without applying the streamline diffusion method, Theorem 6.4 in [19] implies that $\|\nabla(u - u_h)\| \lesssim (1/2)\mathrm{Pe}_h\, h$, where h is the length of the longest element edge, see Remark 6.8 in [19]. On the other hand, using the streamline diffusion method yields $\left\|w^\perp \cdot \nabla(u - u_h)\right\| \lesssim ((1/2)\mathrm{Pe}_h)^{1/2}\, h$, where $w^\perp$ is the crosswind, see Remark 6.8 in [19]. This remark contains the conclusion that if $\mathrm{Pe}_h > 2$, the application of the streamline diffusion method is more reliable than the consideration of the introduced weak formulation (4.9).
To summarize, the finite element method is widely used for the discretization of PDEs and has a solid theoretical background [56]. The implementation is not as straightforward than for finite difference methods, but the consideration of more irregular domains is possible without a loss of the order of consistency and convergence.

4.3 Spectral properties

After introducing spatial discretization techniques, the next aim is to gain an insight into the spectral properties of convection-diffusion problems. We consider the steady-state case of the equation (4.1) and study the spectrum of the resulting partial differential operator $\mathcal{L} = \alpha\Delta - w \cdot \nabla$. For simplification, a one-dimensional operator with homogeneous Dirichlet boundary conditions and constant coefficients α, w is considered. This convection-diffusion operator is a linear operator and we refer to the three volume treatise [14], [15], [16] of Dunford and Schwartz for an extensive discourse on linear operators and for the mathematical tools which are needed in the following. Based on this study, we want to choose eigenvalues λ which are usually in the spectrum of such convection-diffusion operators. These specific λ are the basis for the work on the convergence behavior of SDC fixed point iterations of Definition 2.10 in the next Section 4.4. At first, a continuous convection-diffusion operator $\mathcal{L}$ is considered. Afterwards, discrete convection-diffusion operators are covered which result from a discretization techniques for PDEs, see the previous Section 4.2.
The convection term $w \cdot \nabla u$ from equation (4.1) leads to crucial differences in the numerical treatment of convection-diffusion problems compared to pure parabolic PDEs. In such problems with just diffusion, for example, reaction-diffusion equations as considered before in Chapter 3, we deal with the Laplace operator Δ, which in general is unbounded, linear and elliptic. This partial differential operator has special properties which enable the application of a powerful spectral theory.

Definition 4.1. Let X, Y be complex Hilbert spaces and $\mathcal{L} : D_1 \to Y$ a linear operator where $D_1 \subset X$. Furthermore, consider the linear operator $\mathcal{L}^* : D_2 \to X$ where $D_2 \subset Y$ is the subspace which consists of y for which $f(x) = \langle \mathcal{L}x, y\rangle$ is continuous on D_1 and $\langle \mathcal{L}x, y\rangle_Y = \langle x, \mathcal{L}^* y\rangle_X$ for all $x \in D_1$. This operator $\mathcal{L}^*$ is the *adjoint operator* of $\mathcal{L}$ and it fulfills

$$\langle \mathcal{L}x, y\rangle_Y = \langle x, \mathcal{L}^* y\rangle_X \quad \forall x \in D_1 \quad \text{and} \quad \forall y \in D_2,$$

where $\langle \cdot, \cdot\rangle_X$ and $\langle \cdot, \cdot\rangle_Y$ are the inner products on X and Y, respectively. If $X = Y$ and $\mathcal{L} = \mathcal{L}^*$, then $\mathcal{L}$ is called *self-adjoint*, see the first definition in Section 1.2 in [10] and Definition II.7.1 and Remarks II.7.2 in [24].

Definition 4.2. Let $\mathcal{L} : D \to H$ be a linear operator on the complex Hilbert space H where $D \subset H$ is a subset. The operator $\mathcal{L}$ is called *symmetric* with respect to the inner product $\langle \cdot, \cdot\rangle_H$ if

$$\langle \mathcal{L}x, y\rangle_H = \langle x, \mathcal{L}y\rangle_H \quad \forall x, y \in D.$$

Lemma 4.3. Let $D = \mathcal{C}_c^\infty(\Omega)$ be the space of smooth functions with compact support on the open and non-empty subset $\Omega \subset \mathbb{R}^n$ and let $H = L_2(\Omega)$ be the Hilbert space of quadratically integrable functions on Ω, see (4.8). Consider cartesian coordinates and the corresponding Laplace operator $\Delta : D \to H$,

$$\Delta = \sum_{j=1}^{n} \frac{\partial^2}{\partial x_j^2}.$$

This partial differential operator is symmetric with respect to the inner product defined by $\langle \varphi, \psi\rangle_{L_2(\Omega)} = \int_\Omega \varphi(x)\overline{\psi(x)}\mathrm{d}x$, where $\varphi, \psi \in L_2(\Omega)$.

Proof. Let $\varphi, \psi \in \mathcal{C}_c^\infty(\Omega)$. Due to the compact support of the considered functions on the open set Ω, the functions are zero on the corresponding boundary $\partial\Omega$. Thereby and with integration by parts we obtain the result

$$\begin{aligned}
\langle \Delta\varphi, \psi\rangle_{L_2(\Omega)} &= \int_\Omega \Delta\varphi(x)\overline{\psi(x)}\mathrm{d}x = \int_\Omega \sum_{j=1}^{n} \frac{\partial^2\varphi(x)}{\partial x_j^2}\overline{\psi(x)}\mathrm{d}x = \sum_{j=1}^{n} \int_\Omega \frac{\partial^2\varphi(x)}{\partial x_j^2}\overline{\psi(x)}\mathrm{d}x \\
&= \sum_{j=1}^{n} \left(\underbrace{\left[\frac{\partial\varphi(x)}{\partial x_j}\overline{\psi(x)}\right]_{\partial\Omega}}_{=0} - \int_\Omega \frac{\partial\varphi(x)}{\partial x_j}\frac{\partial\overline{\psi(x)}}{\partial x_j}\mathrm{d}x \right) \\
&= \sum_{j=1}^{n} \left(\underbrace{-\left[\varphi(x)\frac{\partial\overline{\psi(x)}}{\partial x_j}+\right]_{\partial\Omega}}_{=0} + \int_\Omega \varphi(x)\frac{\partial^2\overline{\psi(x)}}{\partial x_j^2}\mathrm{d}x \right) \\
&= \sum_{j=1}^{n} \int_\Omega \varphi(x)\frac{\partial^2\psi(x)}{\partial x_j^2}\mathrm{d}x = \int_\Omega \varphi(x)\overline{\Delta\psi(x)}\mathrm{d}x = \langle \varphi, \Delta\psi\rangle_{L_2(\Omega)}.
\end{aligned}$$

□

The ellipticity and symmetry of the Laplace operator and Corollary 3.5.4 of [10] leads to the result that its spectrum is real. This was used in Chapter 3 to choose λ for Dahlquist's equation in the reaction-diffusion case. It can be proven that the Laplace operator is also essentially self-adjoint and there is a well developed spectral theory about self-adjoint operators in Hilbert space, see, for example, [10] and [15]. This theory is based on the important result that the spectrum of any self-adjoint operator is real and non-empty, see Theorem 1.2.10 in [10].
But we are interested in convection-diffusion operators, as, for example, $\mathcal{L} = \alpha\Delta - w \cdot \nabla$ from (4.1), and these operators are generally not self-adjoint and more complex. Thus, the theory of them is less developed. Additionally, we have to deal with the concept of pseudospectra by considering non-self-adjoint operators.

Definition 4.4. Let $\Lambda(\mathcal{L})$ be the spectrum of a closed linear operator $\mathcal{L}$ in Hilbert space. A value λ is in the spectrum of $\mathcal{L}$ if the resolvent $(\lambda I - \mathcal{L})^{-1}$ is unbounded or nonexistent. In this case, λ is called eigenvalue of $\mathcal{L}$. The *ϵ-pseudospectrum of $\mathcal{L}$*, where $\epsilon \geq 0$, is a subset of the complex plane and it is defined as

$$\Lambda_\epsilon(\mathcal{L}) = \{\lambda \in \mathbb{C} : \lambda \in \Lambda(\mathcal{L} + \mathcal{E}) \ \mid \ \|\mathcal{E}\| \leq \epsilon\},$$

where $\mathcal{E}$ is a perturbation of the operator $\mathcal{L}$. The norm $\|\cdot\|$ is induced by the inner product on the Hilbert space.

We refer to the book [54] for a detailed treatment of pseudospectra or to the paper [52] for a shorter discussion with some interesting examples, including a convection-diffusion problem. The Section 7.9 in [25] also provides a short overview of the topic of pseudospectra.

Definition 4.5. A *normal operator* $\mathcal{L}$ is one that satisfies $\mathcal{L}\mathcal{L}^* = \mathcal{L}^*\mathcal{L}$, where $\mathcal{L}^*$ is the adjoint operator of $\mathcal{L}$.

Equivalently, a normal operator has a complete set of orthogonal eigenvectors. In particular, self-adjoint operators are normal. If convection-diffusion operators with constant coefficients are defined on an unbounded domain, then they are also normal. But by applying boundary conditions or variable coefficients, the convection-diffusion operators become non-normal, see Section 12 in [54].
If the interest is in the general behavior of a PDE, the spectrum of the corresponding operator can lead to useful information, see Chapter I in [54] for some illustrative examples. However, the spectrum of a non-normal operator can also give misleading results. To explain this, we consider perturbations of an operator, which arise, for example, due to inexact arithmetic in the numerical solution of PDEs. The spectrum of the perturbed operator $\mathcal{L} + \mathcal{E}$ is a subset of the ϵ-pseudospectrum of the unperturbed operator $\mathcal{L}$ where $\epsilon = \|\mathcal{E}\|$. The ϵ-pseudospectrum provides an information about the distance from the spectrum of the unperturbed to the spectrum of the perturbed operator. The ϵ-pseudospectrum of a normal operator is composed of the corresponding ϵ-balls around its eigenvalues, see Theorem 2.2 in [54]. Thus, the spectrum of a perturbed operator is not far away from the spectrum of the unperturbed operator if the perturbation is small. This means that the results from the spectral theory of normal operators can be applied in the case of perturbed normal operators as well. On the other hand, by considering

convection-diffusion problems, a non-normal partial differential operator can be obtained and the eigenvalues of a non-normal operator are more sensitive to perturbations, see [37]. The ϵ-pseudospectrum of such an operator can cover a large region despite a small perturbation and thus the eigenvalues of a perturbed non-normal operator can be far away from the eigenvalues of the corresponding unperturbed operator. Therefore, the additional study of the pseudospectra of non-normal operators can be more meaningful than considering just the spectrum.

Remark 4.6. An unbounded linear operator $\mathcal{L}$ is self-adjoint if $\mathcal{L}$ is symmetric and the domains of $\mathcal{L}$ and $\mathcal{L}^*$ are equal. There is no difference between symmetry and self-adjointness for bounded linear operators, see Section 1.2 in [10].

4.3.1 Pseudospectra of the simplest one-dimensional convection-diffusion operator

At first, a continuous convection-diffusion operator is regarded. Based on [45], we consider $\mathcal{N} : D \to H$,

$$\mathcal{N}u_* = \Delta u_* + \nabla u_* = u_*'' + u_*', \qquad u_*(0) = u_*(d) = 0, \tag{4.15}$$

acting in the Hilbert space $H = L^2([0, d]))$. The domain $D = \mathcal{C}^2((0, d))$ is the space of twice differentiable functions which fulfill homogeneous Dirichlet boundary conditions. The constant diffusion and convection coefficients are equal to one in this case. Thus, the Peclet number $\mathrm{Pe} = WL/\alpha$ (4.2) only depends on the characteristic length scale L, which is equal to the length d of the domain. An increase of d leads to stronger convective effects.
Theorem 1 from [45] leads to the spectrum $\Lambda(\mathcal{N}) = \cup_{n>0}\{\lambda_n\}$ of $\mathcal{N}$, with

$$\lambda_n = -\frac{1}{4} - \frac{\pi^2 n^2}{d^2}, \quad n = 1, 2, 3, \ldots . \tag{4.16}$$

This discrete spectrum is real and negative. With the considerations above regarding perturbations of non-normal operators, we investigate the pseudospectra of $\mathcal{L}$ in the following. This is also the main topic of [45] and its leading Figure is presented here in Figure 4.2. The dots are the first 27 eigenvalues of $\mathcal{N}$ and the parabolas are the contours of its pseudospectra due to more or less perturbations of the operator. For example, the values inside the contour nearest to the eigenvalues of $\mathcal{N}$ are the possible eigenvalues of a perturbed $\mathcal{N}$ due to perturbations $\mathcal{E}$ where $\|\mathcal{E}\| \leq 10^{-7}$.
The dashed line is the *critical parabola*

$$\mathrm{Re}(\lambda) = -(\mathrm{Im}(\lambda))^2$$

and with the theoretic case of $d = \infty$ we obtain the spectrum $\Lambda(\mathcal{N}) = \Pi$ which is the region inside the critical parabola. Furthermore, the ϵ-pseudospectrum is then given by $\Lambda_\epsilon(\mathcal{N}) = \Pi + \Delta_\epsilon$ for all $\epsilon \geq 0$, where $\Delta_\epsilon = \{\lambda \in \mathbb{C} : |\lambda| \leq \epsilon\}$, see Theorem 3 in [45]. We refer to [45] for the meaning of the applied sum of two sets. If we consider a perturbation $\mathcal{E}$ with a possible eigenvalue of the corresponding perturbed operator outside the critical parabola, then the distance from this eigenvalue to the critical parabola grows linearly if $\|\mathcal{E}\|$ increases linearly with a constant equal to one, see Theorem 4 in [45]. On the other

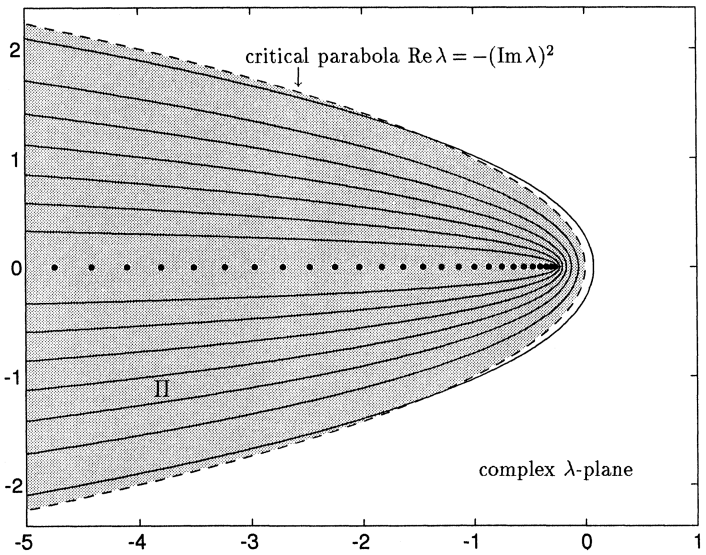

FIG. 1. *Contour plot of the resolvent norm surface* $\|(\lambda I - \mathcal{L})^{-1}\|$ *in the complex* λ*-plane* ($d = 40$). *The contours represent levels* $10^1, 10^2, \ldots, 10^7$, *and the dots are the eigenvalues. At each point* λ *in the interior of the region* Π (*shaded*), $\|(\lambda I - \mathcal{L})^{-1}\|$ *grows exponentially as* $d \to \infty$. *Equivalently, the figure can be interpreted as a depiction of* ϵ*-pseudospectra of* $\mathcal{L}$ *for* $\epsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-7}$.

Fig. 4.2: From *Pseudospectra of the Convection-Diffusion Operator* by Satish C. Reddy and Lloyd N. Trefethen [45].

hand, considering a certain region of an ϵ-pseudospectrum which is inside this parabola, and then letting $d \to \infty$, then ϵ decreases exponentially for this fixed region, see Theorem 5 in [45]. Furthermore, we have a continuous change for $\Lambda_\epsilon(\mathcal{N}) \to \Pi + \Delta_\epsilon$ as $d \to \infty$ for all $\epsilon \geq 0$, see Theorem 6 in [45]. Thus, as a first heuristically idea of the critical parabola, the perturbation has to be quite large so that the spectrum of a perturbed operator is far out of the critical parabola. For small perturbations, the spectrum of a perturbed operator stays inside the critical parabola. It further requires no large perturbations to get the whole region inside the critical parabola as the ϵ-pseudospectrum if we consider large interval lengths d. From Theorem 7 in [45] follows a more general bound

$$\Lambda_\epsilon(\mathcal{N}) \subseteq \left\{ \lambda \in \mathbb{C} : |\mathrm{Im}(\lambda)| \leq \epsilon e^{d/2} \right\}$$

for the ϵ-pseudospectrum than obtained by the critical parabola. The ϵ-pseudospectrum from the operator (4.15) is contained in a strip of finite width equal to $\epsilon\, e^{d/2}$, i.e., the eigenvalues of a perturbed operator $\mathcal{N} + \mathcal{E}$ change by at most $e^{d/2}\, \|\mathcal{E}\|$. This bound is known as the *Bauer-Fike Theorem*, compare with Theorem 7.2.2 in [25].
The results from [45] are stated for the simplest one-dimensional convection-diffusion operator (4.15). All these results carry over to the more general operator

$$\mathcal{L}u = \alpha \Delta u + w \nabla u = \alpha u'' + w u', \qquad u(0) = u(\delta) = 0 \tag{4.17}$$

where α and w are constant coefficients. By following the Appendix in [46], we start

with the dimensional equation

$$\mathcal{L}u = \alpha\frac{\partial^2 u}{\partial x^2} + w\frac{\partial u}{\partial x}, \qquad u(0) = u(\delta) = 0$$

and by substituting $u(x) = u_*(\xi)$ on a kind of normalized domain with $\xi = wx/\alpha$, we obtain

$$\mathcal{L}u = \frac{w^2}{\alpha}\frac{\partial^2 u_*}{\partial \xi^2} + \frac{w^2}{\alpha}\frac{\partial u_*}{\partial \xi}, \qquad u_*(0) = u_*(\frac{w}{\alpha}\delta) = 0.$$

With $\mathcal{L} = w^2\mathcal{N}/\alpha$ and the further substitution $d = w\delta/\alpha$ we get the dimensionless equation

$$\mathcal{N}u = \frac{\partial^2 u_*}{\partial \xi^2} + \frac{\partial u_*}{\partial \xi}, \qquad u_*(0) = u_*(d) = 0,$$

compare with equation (4.15). This transformation shows that all convection-diffusion problems with $\mathcal{L} = \alpha\Delta + w\nabla$ on the interval $[0, \delta]$ are equivalent to the problem with the dimensionless operator $\mathcal{L} = \Delta + \nabla$ on the interval $[0, w\delta/\alpha]$. By comparing the parameter $d = w\delta/\alpha$ with the definition of the Peclet number, we see that they are equal. Thus, for the consideration of different convection-diffusion problems we need just the simplest operator (4.15) and a dimensionless number. In this way, the results of [45] can be simply applied to a more general operator. For example, the critical parabola $\mathrm{Re}(\lambda) = -(\mathrm{Im}(\lambda))^2$ becomes the parabola

$$\mathrm{Re}(\lambda) = -\frac{\alpha}{w^2}\left(\mathrm{Im}(\lambda)\right)^2. \tag{4.18}$$

If the ϵ-pseudospectrum is now inside this parabola, then ϵ decreases exponentially as $w\delta/\alpha \to \infty$. Meaning that Theorem 5 of [45] now concerns the limit case of $w\delta/\alpha \to \infty$ by considering the more general operator (4.17), see the Appendix in [46].

4.3.2 Pseudospectra of the discrete operators

Until now, a continuous linear operator $\mathcal{L}$ is considered. For the numerical solution of PDEs, such an operator has to be discretized and this yields a matrix $L_h \in \mathbb{R}^{N\times N}$ as a discrete convection-diffusion operator. Section 4.2 provides an introduction to spatial discretization techniques, which are now used for the study of the pseudospectra of different matrices L_h. Such matrices are approximations of $\mathcal{L}$ and thus the pseudospectra of a matrix L_h approximate the pseudospectra of the continuous operator $\mathcal{L}$. The deviation between these pseudospectra depends on the discretization.
We want to illustrate the connection between the critical parabola and the pseudospectra of a discrete convection-diffusion operator L_h with a simple numerical example. Consider the one-dimensional convection-diffusion boundary value problem

$$\alpha u'' + u' = 1 \quad \text{ in } (0, \delta), \qquad u(0) = u(\delta) = 0 \tag{4.19}$$

with the convection-diffusion operator $\mathcal{L} = \alpha\Delta + \nabla$. By decreasing α or increasing δ, it is possible to construct a convection dominated boundary value problem. The continuous

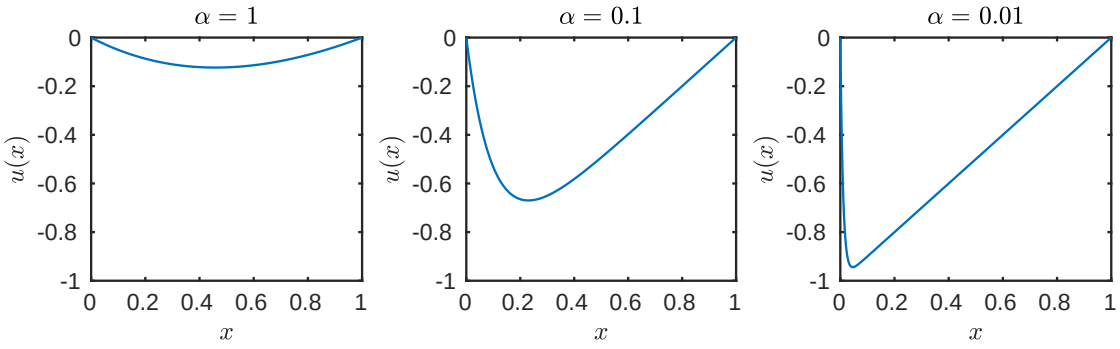

Fig. 4.3: Exact solution of a one-dimensional convection-diffusion boundary value problem with different diffusion coefficients α and a convection coefficient $w = 1$. Increasing convection domination from the left to the right.

operator $\mathcal{L}$ is discretized by a finite difference method where central difference approximations are applied for the first derivative and an equidistant grid $\{x_i\}_{i=0}^{N-1}$ is used on the interval $[0, \delta]$. Thus, we obtain the approximations $u''(x_i) \approx (u_{i+1} - 2u_i + u_{i-1})/h^2$ and $u'(x_i) \approx (u_{i+1} - u_{i-1})/(2h)$ with the grid size $h = x_i - x_{i-1}$ and $u_i \approx u(x_i)$, see Section 6 in [45]. The exact solution of the convection-diffusion problem (4.19) is given by $u(x) = x - \delta \left(e^{(\delta - x)/\alpha} - e^{\delta/\alpha}\right) / \left(1 - e^{\delta/\alpha}\right)$. For $\alpha \to 0$, at $x = 0$ we get the simplest case of a singularly perturbed differential equation, see Figure 4.3. The pseudospectra of L_h are obtained by the computation of the minimal singular value of the matrix $zI - L_h$ for complex numbers z, see Section 4 in [52] for an explanation. After this, a contour plotter is used for the singular values. The lines in the resulting plots correspond to the contours of the pseudospectra of L_h. For the singular value decomposition, we use the MATLAB® function *svd*. Figure 4.4 contains pseudospectra of different matrices L_h that arise from problem (4.19). There, we select combinations from $\alpha = \{0.1, 0.2\}$ and $\delta = \{10, 20\}$ and set the mesh Peclet number (4.14) to $\mathrm{Pe}_h = 1$. For $\alpha = 0.2$ and $\delta = 10$, the ϵ-pseudospectrum where $\epsilon = 10^{-16}$ is so small that it is not shown. The pseudospectra of the matrices L_h are bounded by the critical parabola $\mathrm{Re}(\lambda) = -\alpha\,(\mathrm{Im}(\lambda))^2$ if a domination of convection is forced by an increase of δ or a decrease of α. If the interval length δ is increased, the pseudospectra move closer together and tend towards the critical parabola.

The pseudospectra of the continuous operator $\mathcal{L}$, which have a parabolic contour, are shown in Figure 4.2. The discrete operators L_h with homogeneous Dirichlet boundary conditions are tridiagonal Toeplitz matrices and a property of such matrices is that they map circles about the origin onto ellipses, see [45] or [47]. Thus, the pseudospectra of the discrete operators L_h have an elliptical contour.

For the next numerical example, consider the convection-diffusion operator

$$\mathcal{L}u = \Delta u - \nabla u = u'' - u', \qquad u(0) = u(\delta) = 0. \tag{4.20}$$

The change of sign of the convection term leads to another spectrum and other pseudospectra of $\mathcal{L}$ than before, but the theory of Section 2 in [45] can be easily applied for this case, too. The operator is discretized with a uniform grid and central difference approximations again. For this example, we use the MATLAB® function *eig* to compute the spectra of the discrete operators L_h. These computed spectra are not the exact

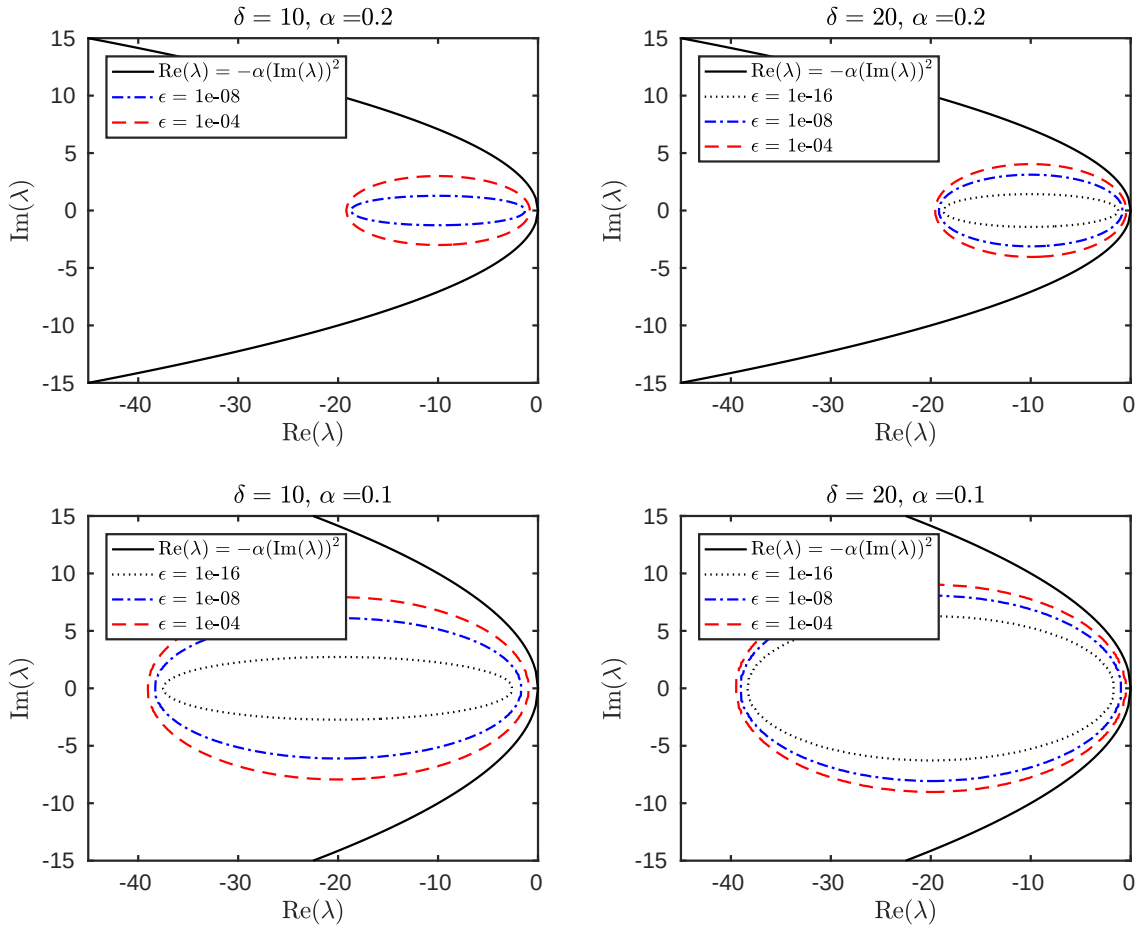

Fig. 4.4: ϵ-pseudospectrum of discrete convection-diffusion operators L_h based on $\mathcal{L} = \alpha\Delta + \nabla$ with homogeneous Dirichlet boundary conditions and different interval lengths δ and diffusion coefficients α. The convection coefficient is $w = 1$ and there holds $\mathrm{Pe}_h = 1$.

spectra of the matrices L_h, they can be seen as the spectra of some perturbed matrices with perturbations depending on the accuracy of the MATLAB® function *eig*. The computed spectra $\Lambda_{eig}(L_h)$ on different intervals $[0, \delta]$ with $\mathrm{Pe}_h = 1$ are shown in the plots of Figure 4.5. With increasing δ, i.e., the Peclet number Pe increases, the computed spectra of L_h approach the critical parabola. The reason for this is that in the case of a larger Peclet number the matrices L_h are more sensitive to perturbations and despite the high accuracy of the MATLAB® function *eig* the errors in the computation of the eigenvalues can be quite large.

4.3.3 The β-parabola-region

As motivated in the previous subsections, we follow the idea of considering the pseudospectra instead of the spectrum of a convection-diffusion operator $\mathcal{L}$. The ultimate aim of this work is the numerical solution of unsteady convection-diffusion problems with SDC methods. Thus, the pseudospectra of the matrices L_h, which arise due to a spatial

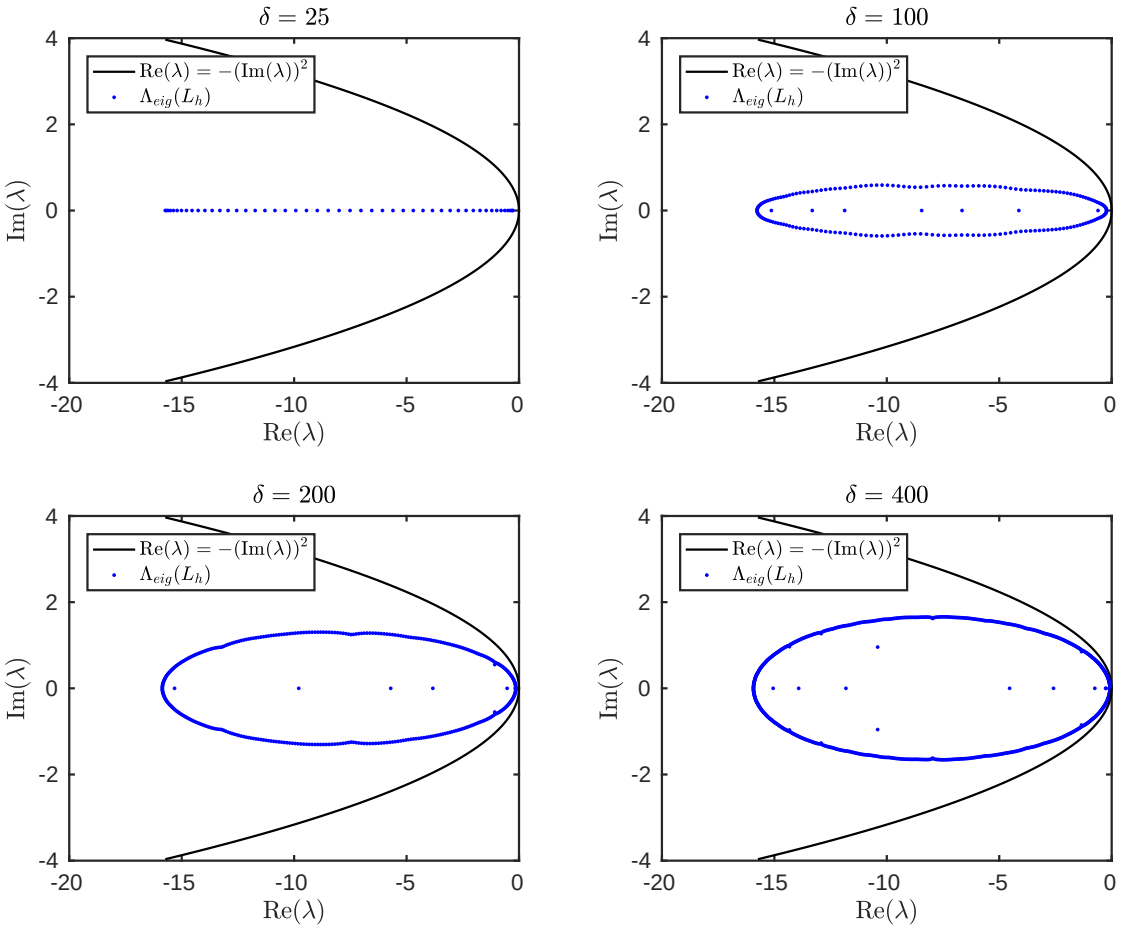

Fig. 4.5: Computed spectra $\Lambda_{eig}(L_h)$ of discrete convection-diffusion operators L_h resulting of the problem (4.20) with $\alpha = 1$, $w = 1$ and different interal lengths δ. The grid size is fixed with $h = 1$ and thus $\mathrm{Pe}_h = 1$.

discretization, are of interest. These pseudospectra depend on the discretization, but we want to derive an independent framework. The idea is that the existence of highly accurate spatial discretizations is assumed so that the pseudospectra of the discrete operators approximate the pseudospectra of the continuous operators in such a way that the results of the continuous case can be applied to the discrete case, at least approximately. Thus, in the remaining part of this work is assumed that the considerations regarding the critical parabola (4.18) carry over to the matrices L_h. The numerical experiments in the last Subsection 4.3.2 support this assumption.
Furthermore, for simplification, the following study is based on the one-dimensional convection-diffusion operator $\mathcal{L} = \alpha\Delta + w\nabla$ defined on $(0, \delta)$ with homogeneous Dirichlet boundary conditions and constant coefficients α, w, see Remark 4.8. From Subsection 4.3.1 we know that for small perturbations the possible eigenvalues of a corresponding perturbed convection-diffusion operator are somewhere inside the critical parabola $\mathrm{Re}(\lambda) = -(\alpha/w^2)(\mathrm{Im}(\lambda))^2$. If $\mathrm{Pe} \to \infty$, the pseudospectra of the operator $\mathcal{L}$ become the

whole region Π inside this parabola, also with very small perturbations. For smaller Pe, the pseudospectra are subsets of Π which are also bounded by parabolas and for $\mathrm{Pe} \to 0$ these parabolas become narrower with the limit case of the negative real axis in the complex plane. However, in the following, it is assumed that there are such large Peclet numbers and possible perturbations of the operators so that the whole region inside the critical parabola is considered as their ϵ-pseudospectrum.
After a spatial discretization of the operator $\mathcal{L}$ with some grid size h, an IVP for the convection-diffusion equation (4.1) is obtained. Thus, we need a temporal discretization with a time step size for solving this IVP in time. In our SDC framework, this time step size is equal to the time interval length τ of Definition 2.9. We want to investigate the behavior of SDC fixed point iterations of Definition 2.10 for $z = \tau\lambda$ where $\lambda \in \Pi \subset \mathbb{C}$, as described above, and $\tau > 0$. This and the critical parabola $\mathrm{Re}(\lambda) = -(\alpha/w^2)(\mathrm{Im}(\lambda))^2$ lead to the next definition.

Definition 4.7. We call the set

$$\Pi_z(\beta) = \left\{ z \in \mathbb{C} : \mathrm{Re}\,(z) \leq -\beta(\mathrm{Im}(z))^2 \right\}$$

the *β-parabola-region* for the convection-diffusion problem (4.1), where $\beta = \alpha/(\tau W^2)$. The value $W \in \mathbb{R}^+$ is a parameter for the wind $w(x)$ where $\|w(x)/W\| = \mathcal{O}(1)$ in some norm $\|\cdot\|$. The scalar β is a kind of parameter for unsteady convection-diffusion problems.

To get a more general framework in this work, it is assumed that the idea of the β-parabola-region can be also applied to more general convection-diffusion operators. For the numerical experiments in Section 5.2, we consider, for example, a two-dimensional problem, inhomogeneous Dirichlet boundary conditions and a non-constant wind $w(x)$.

Remark 4.8. At this point, we have to mention that by changing the type of boundary conditions or the spatial dimension of the problem, the results of the critical parabola (4.18) will probably change. For non-constant coefficients α and w, this is also the case, compare Theorem 10.2 in [54] with [9].

4.4 Faster SDC convergence for convection-diffusion equations

With the β-parabola-region of Definition 4.7 we can make an appropriate choice of the parameter $z = \tau\lambda$ for SDC fixed point iterations of Definition 2.10 so that λ is usually in the spectrum of a convection-diffusion operator. As in the previous chapter concerning reaction-diffusion problems, the SDC methods are constructed with L-stable Radau-IIa collocation discretizations. Highly accurate spatial discretizations of convection-diffusion problems, as assumed in this work, can lead to stiff components in the discretized partial differential operators. These are the eigenvalues with a large negative real part. With an L-stable Radau-IIa method, these stiff components can be damped out and oscillations in the numerical solution can be prevented. On the other hand, the damping property of Radau-IIa methods can also lead to disadvantages if components with a large imaginary part are damped out. Gauss or Lobatto collocation discretizations, which are A-stable,

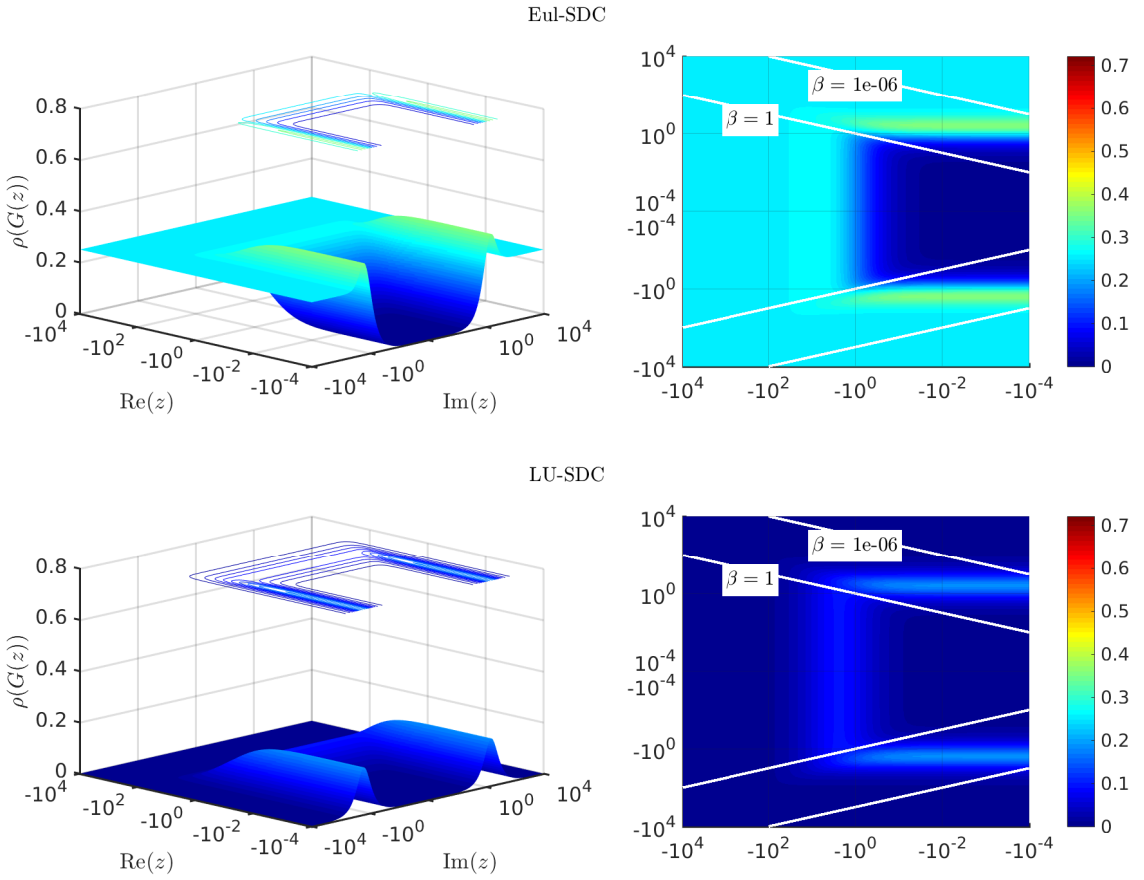

Fig. 4.6: Spectral radius $\rho(G(z))$ of Eul-SDC and LU-SDC methods on Radau-IIa grids with $s = 2$ collocation points. Bounds of β-parabola-regions are marked with white lines. Re(z) and Im(z) are scaled logarithmically.

but not L-stable, can be reasonable choices if the β-parabola-region is determined by a very large β and the operators have no very stiff components. For a detailed treatment of the stability properties of collocation discretizations, the reader is referred to [31] and [32].

The following section covers SDC fixed point iterations of Definition 2.10 for $z \in \Pi_z(\beta)$ and different convergence objectives as in the last chapter. There, the function *fminsearch* from MATLAB® is applied for the direct optimization approach. This derivative-free programming solver does not lead to satisfying results for $z \in \Pi_z(\beta)$. Thus, in the following, the programming solver *fminunc* from MATLAB® is used, which works with the approximation of derivatives. The objective functions are computed on a logarithmic grid for values of z with $-10^4 \leq \mathrm{Re}(z) \leq -10^{-4}$ and $-10^4 \leq \mathrm{Im}(z) \leq 10^4$. We use 50 points for the real part and 100 points for the imaginary part. The initial matrices arise from the LU decomposition approach, see Subsection 3.1.2, and the objective functions are approximated with the l^p norm $\|\cdot\|_{64}$ as in [55].

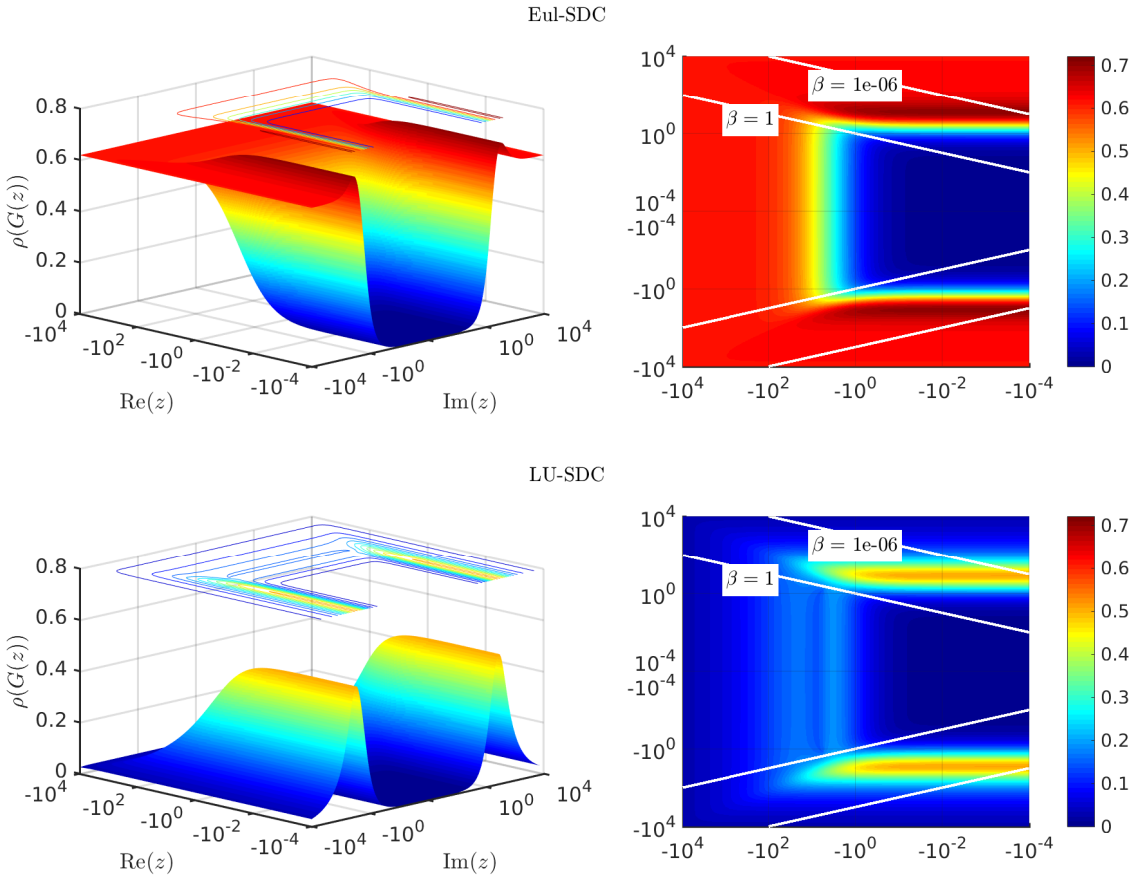

Fig. 4.7: Spectral radius $\rho(G(z))$ of Eul-SDC and LU-SDC methods on Radau-IIa grids with $s = 4$ collocation points. Bounds of β-parabola-regions are marked with white lines. Re(z) and Im(z) are scaled logarithmically.

4.4.1 Asymptotic contraction factor

At first, we study the asymptotic contraction factor Φ_∞ of Definition 2.15, i.e., the spectral radius $\rho(G(z))$ of SDC fixed point iterations of Definition 2.10 is considered for values of z in the complex plane.
In the Figures 4.6 and 4.7, Eul-SDC and LU-SDC methods with $s = 2$ and $s = 4$ collocation points can be compared. The first discovery for the Eul-SDC methods is that $\rho(G(|z| \to 0)) = 0$ and $\rho(G(|z| \to \infty)) \neq 0$, see the top plots in the Figures 4.6 and 4.7. Furthermore, $\rho(G(z))$ grows for larger $|z|$ as the number of collocation points s increases. In the previous chapter, we observed the same behavior for reaction-diffusion problems, where $0 > z \in \mathbb{R}$. The LU-SDC methods for reaction-diffusion problems have a superior convergence behavior in the limit case $z \to -\infty$. The experiments for LU-SDC methods with $z \in \mathbb{C}$ where $\text{Re}(z) < 0$ shows this behavior as well, see the bottom plots of Figure 4.6 and 4.7. The maximum of the asymptotic contraction factor $\rho(G(z))$ reduces and we obtain further the property of a vanishing $\rho(G(z))$ in the limit case $|z| \to \infty$. This includes $\text{Re}(z) \to -\infty$ the limit case of stiff problems. The next theorem summarizes this result.

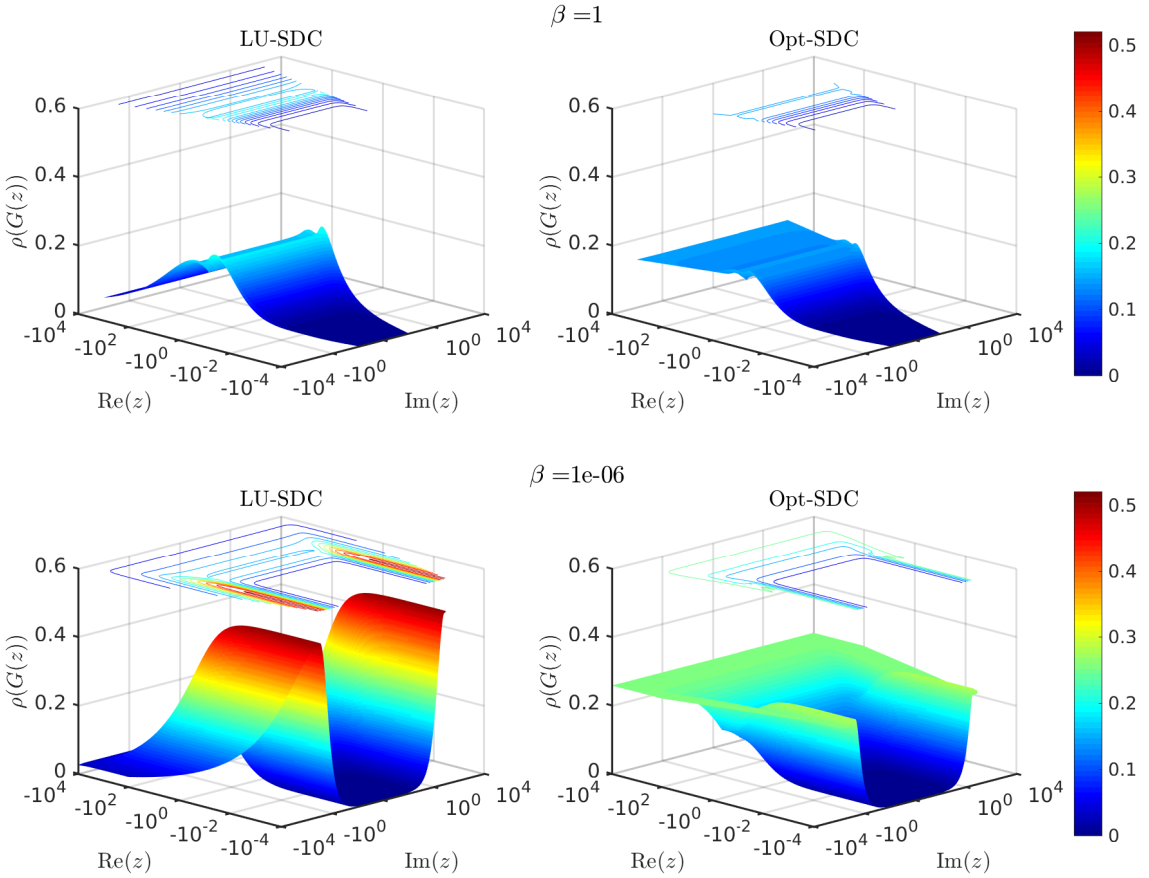

Fig. 4.8: Spectral radius $\rho(G(z))$ of LU-SDC and Opt-SDC methods on Radau-IIa grids with $s = 4$ collocation points. The optimizations (4.21) are performed for different β-parabola-regions. $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ are scaled logarithmically.

Lemma 4.9. Considering an LU-SDC method of Definition 3.5 and complex-valued $z \in \mathbb{C}$, then there holds $\rho(G(|z| \to \infty)) = 0$, where $G(z)$ is the iteration matrix of the LU-SDC method.

Proof. By the LU decomposition $S^{rT} = LU$, see Theorem 3.6, we obtain a unit lower triangular matrix L and an upper triangular matrix U. We take an approximate integration matrix $\hat{S} = U^T$ and this leads to the iteration matrix $G = I - \left(D - zU^T\right)^{-1}\left(D - zU^T L^T\right)$. In the limit case, there holds $G(|z| \to \infty) = I - U^{-T}U^T L^T = I - L^T$. This matrix is an upper triangular matrix where all diagonal entries are equal to zero and therefore $\rho(G(|z| \to \infty)) = 0$, see Lemma 7.1.1 in [25]. □

For LU-SDC methods with $s = 4$ collocation points, see the bottom plots of Figure 4.7, the spectral radius $\rho(G(z))$ reaches maximum values in the regions of approximately $|\mathrm{Im}(z)| = 10$ and $-10 < Re(z) < 0$. We observe that these regions are not in the β-parabola-region $\Pi_z(\beta = 1)$, but they increasingly become part of $\Pi_z(\beta)$ as $\beta \to 0$. This leads to the presumption that the benefit of using LU-SDC methods is much higher for convection-diffusion problems with a larger β. Nevertheless, the numerical experiments

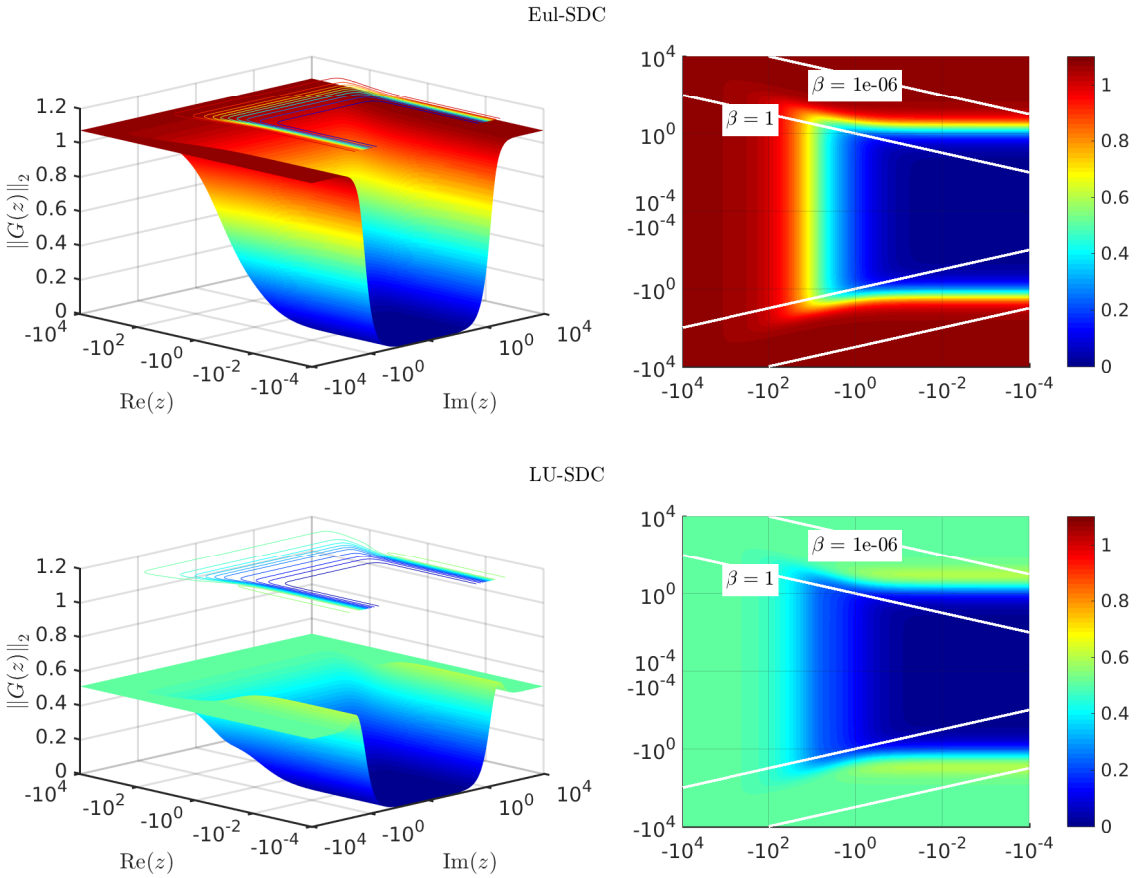

Fig. 4.9: Norm $\|G(z)\|_2$ of Eul-SDC and LU-SDC methods on Radau-IIa grids with $s = 4$ collocation points. Bounds of β-parabola-regions are marked with white lines. Re(z) and Im(z) are scaled logarithmically.

show that the LU-SDC methods have an asymptotic contraction factor which is at least as good as the one from the Eul-SDC methods.
For the next considerations, we regard β-parabola-regions $\Pi_z(\beta)$ for different β. The numerical experiments demonstrate that the direct optimization approach

$$\min J(\hat{D}, \hat{S}) := \max_{z \in \Pi_z(\beta)} \rho(G(z)) \tag{4.21}$$

can lead to a reduction of the maximum of $\rho(G(z))$. However, this comes at the expense of a worsened asymptotic contraction factor for $|z| \to \infty$ compared to the LU-SDC methods and this trade-off has the most impact for z with $\mathrm{Re}(z) \to -\infty$. If there is a convection-diffusion problem so that the observed regions of the highest $\rho(G(z))$ are somewhere in the β-parabola-region $\Pi_z(\beta)$, then the optimization can lead to a significant improvement of this maximum of $\rho(G(z))$. Figure 4.8 compares experiments for LU-SDC and Opt-SDC methods using $s = 4$ collocation points.

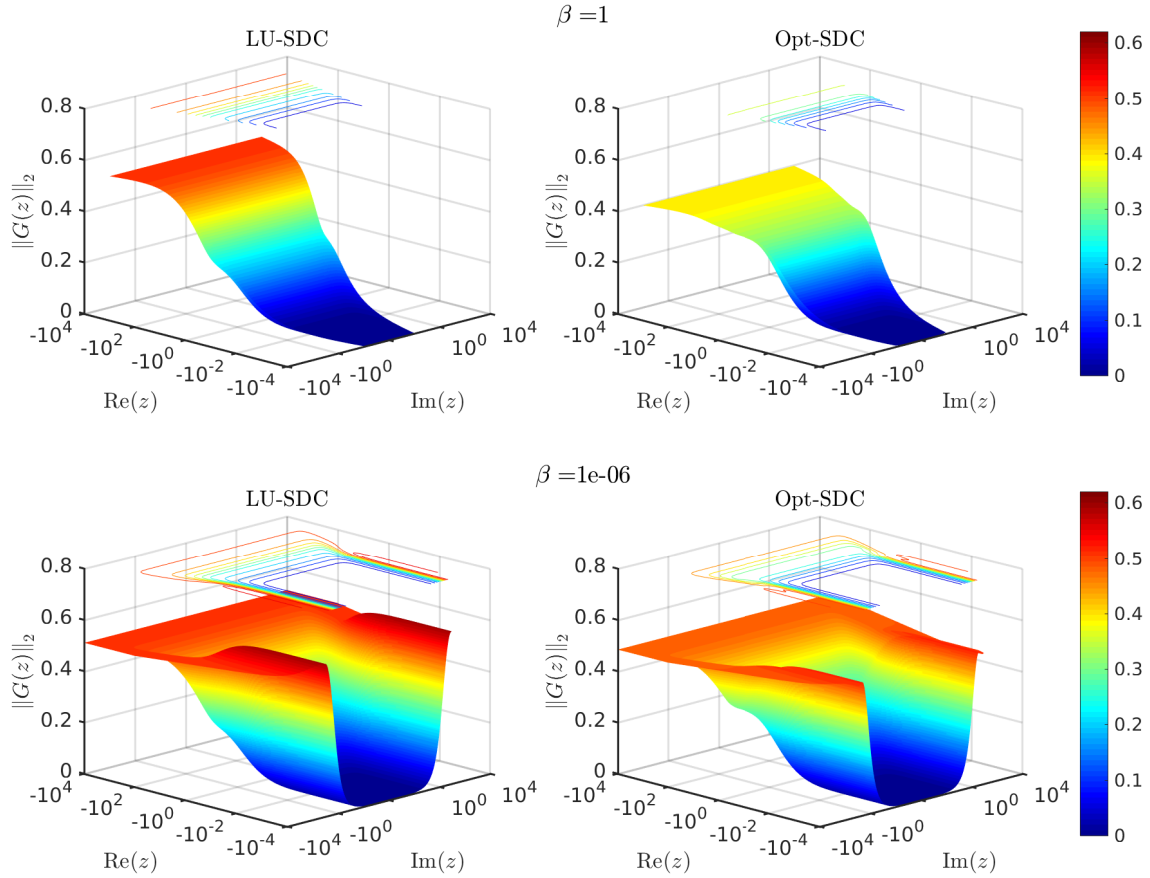

Fig. 4.10: Norm $\|G(z)\|_2$ of LU-SDC and Opt-SDC methods on Radau-IIa grids with $s = 4$ collocation points. The optimizations (4.22) are performed for different β-parabola-regions. $\mathrm{Re}(z)$ and $\mathrm{Im}(z)$ are scaled logarithmically.

4.4.2 Local pre-asymptotic contraction factor

In this subsection, we study the norm $\|G(z)\|$, which is the local pre-asymptotic contraction factor Φ_l of Definition 2.17. This quantity determines the error reduction at all collocation points of an SDC fixed point iteration of Definition 2.10. The following numerical experiments are for the 2-norm. For Eul-SDC methods with $s = 4$ collocation points, they demonstrate that the local pre-asymptotic contraction factor takes its maximum at $|z| \to \infty$, see the top plots of Figure 4.9. The aim is now to reduce Φ_l and, in particular, its maximum with the LU decomposition and optimization approach.
The first observation is that improvements by LU-SDC methods are possible, see Figure 4.9. In these experiments, the local pre-asymptotic contraction factor of the LU-SDC methods is at least as good as the one of the Eul-SDC methods and its maximum is much smaller than the one of the Eul-SDC methods. Compared to the asymptotic contraction factor, the local pre-asymptotic contraction factor does not vanish for $|z| \to \infty$ by considering LU-SDC methods. The direct optimization approach

$$\min J(\hat{D}, \hat{S}) := \max_{z \in \Pi_z(\beta)} \|G(z)\|, \tag{4.22}$$

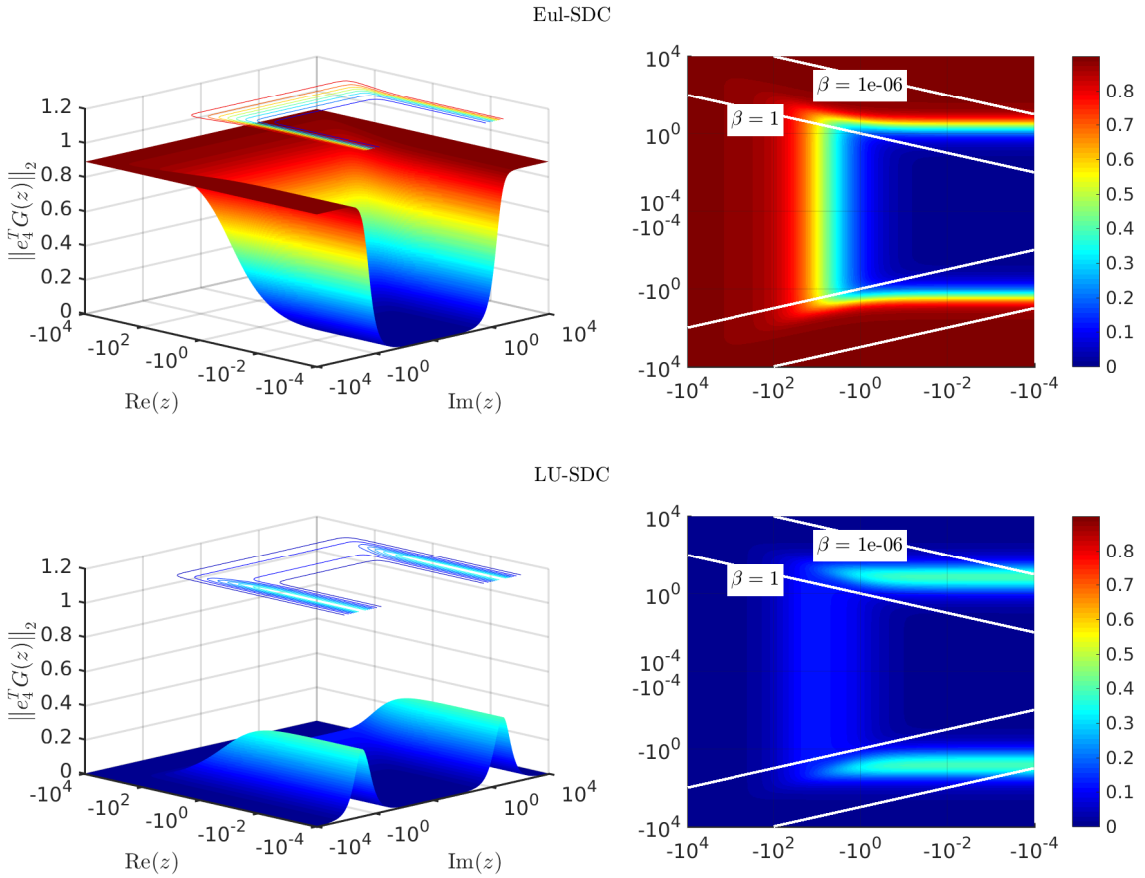

Fig. 4.11: Norm $\left\|e_4^T G(z)\right\|_2$ of Eul-SDC and LU-SDC methods on Radau-IIa grids with $s = 4$ collocation points. Bounds of β-parabola-regions are marked with white lines. Re(z) and Im(z) are scaled logarithmically.

see Figure 4.10, leads to an additional reduction of the maximum of Φ_l with no disadvantages. A vanishing Φ_l can be observed for $|z| \to 0$ because only the approximate integration matrix $\hat{S}$ changes due to the optimization. The approximate differentiation matrix $\hat{D}$ remains the same, which is reasoned by the selected settings of the programming solver *fminunc* from MATLAB®.

4.4.3 Global pre-asymptotic contraction factor

The covered objective of this subsection is to improve the error reduction at the right time interval end point t_n, i.e., to reduce the global pre-asymptotic contraction factor $\Phi_g = \left\|e_n^T G(z)\right\|$ of Definition 2.18. We consider $s = 4$ collocation points and the 2-norm for the numerical experiments. Figure 4.11 compares Eul-SDC methods and LU-SDC methods with $z \in \mathbb{C}$ where $\mathrm{Re}(z) < 0$. The LU decomposition approach leads to an improvement of Φ_g for all values of z and to a vanishing Φ_g for $|z| \to \infty$.

Lemma 4.10. Considering an LU-SDC method of Definition 3.5 with complex-valued $z \in \mathbb{C}$ and the global pre-asymptotic contraction factor Φ_g of Definition 2.18, then there holds $\Phi_g(G(|z| \to \infty)) = 0$, where $G(z)$ is the iteration matrix of the LU-SDC method.

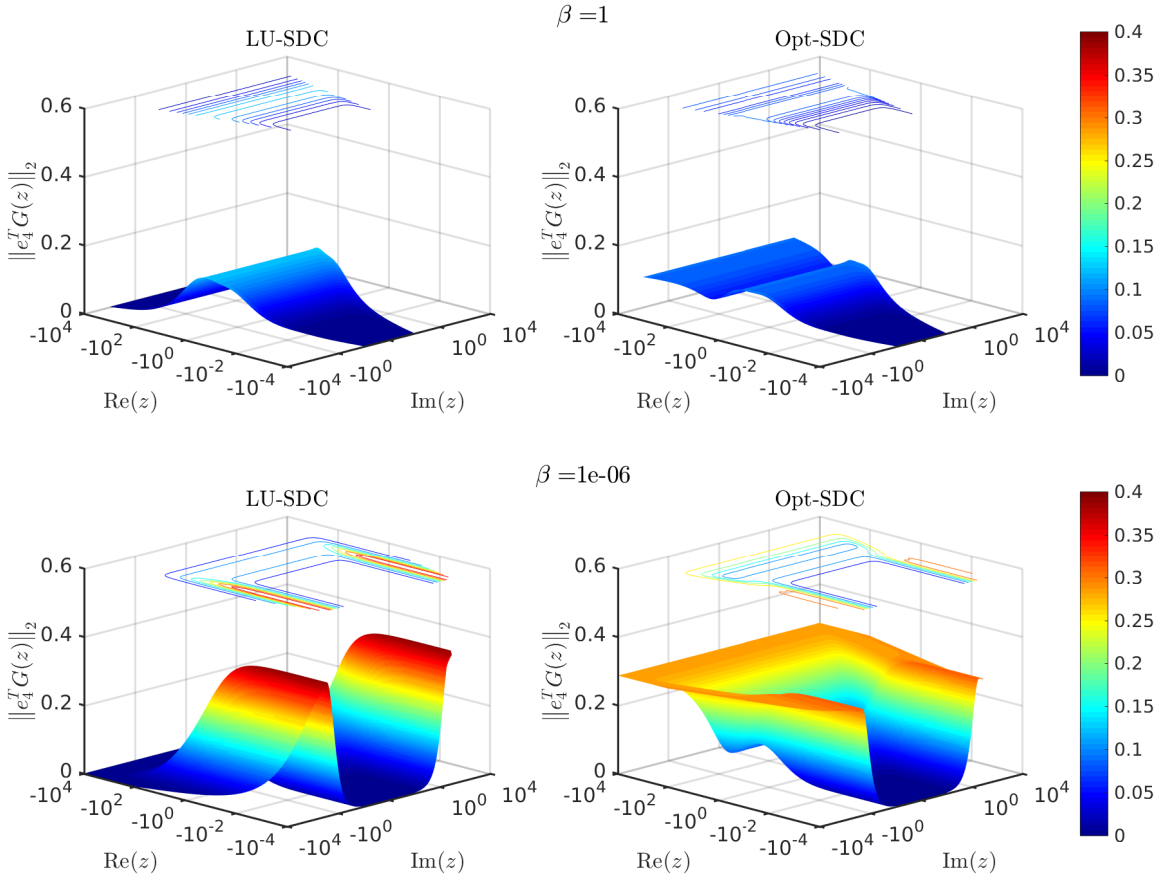

Fig. 4.12: Norm $\left\|e_4^T G(z)\right\|_2$ of LU-SDC and Opt-SDC methods on Radau-IIa grids with $s = 4$ collocation points. The optimizations (4.23) are performed for different β-parabola-regions. Re(z) and Im(z) are scaled logarithmically.

Proof. As in the proof of Lemma 4.9, we have the limit matrix $G(|z| \to \infty) = I - L^T$. The last row of this matrix is given by $e_n^T \left(I - L^T\right)$ and with a unit upper triangular matrix L^T we obtain $e_n^T \left(I - L^T\right) = 0$. This leads to the result $\left\|e_n^T G(|z| \to \infty)\right\| = 0$. □

As in Subsection 4.4.1, the factor Φ_g of LU-SDC methods with $s = 4$ collocation points takes its maximum at intermediate values of z. Thus, we also presume here that LU-SDC methods have a greater benefit for the global pre-asymptotic contraction factor if the convection-diffusion problems have a larger β. In the presented experiment with $s = 4$ collocation points, this is the case for approximately $\beta > 1$.
The direct optimization approach

$$\min J(\hat{D}, \hat{S}) := \max_{z \in \Pi_z(\beta)} \left\|e_4^T G(z)\right\| \tag{4.23}$$

leads to a reduction of this maximum of Φ_g, see Figure 4.12. The trade-off is given by a worse global pre-asymptotic contraction factor for $|z| \to \infty$. As in the previous subsection, the optimization does not affect the approximate differentiation matrix $\hat{D}$ and this leads to a vanishing Φ_g for $|z| \to 0$.

We summarize this section concerning Eul-SDC, LU-SDC and Opt-SDC methods of the Definitions 3.2, 3.5 and 3.7, respectively, as follows: The LU decomposition and the optimization approach are promising for SDC fixed point iterations of Definition 2.10 for the problem class of convection-diffusion equations. One of the main results is that the LU decomposition approach shows a superior convergence behavior for problems where $|z| \to \infty$. Considering LU-SDC methods yields regions in the complex plane of worse asymptotic and global pre-asymptotic contraction factors, see the bottom plots of the Figures 4.7 and 4.11, respectively. Each convection-diffusion problem has a certain β-parabola-region and depending on β, the regions of worse contraction factors lie more or less inside the β-parabola-region. With the optimization approach for different β-parabola-regions a further improvement can be achieved, see the bottom plots of the Figures 4.8 and 4.12.

5 Numerical experiments

The previous chapters dealt with SDC methods for the scalar IVP of Dahlquist's equation 2.9. The next step is to test SDC methods for ODE systems where the corresponding discrete convection-diffusion operators have pseudospectra in the complex plane. The present chapter covers numerical experiments for ODE systems resulting from convection-diffusion problems (4.1) where different spatial discretization techniques are applied. In Section 5.1, we first study the behavior of SDC methods for a one-dimensional problem with constant coefficients which is discretized by finite differences. This example is very similar to the problems in Subsection 4.3.2. Then, the example in Section 5.2 concerns a two-dimensional convection-diffusion problem with a variable wind. It is based on an example of [19] and discretized by the finite element method.

5.1 One-dimensional finite difference discretization

Consider the one-dimensional convection-diffusion initial boundary value problem

$$\begin{aligned} \frac{\partial u}{\partial t} &= \alpha \Delta u - w \cdot \nabla u + r \quad \text{in } (0,\delta) \times (0,T), \\ u(0,t) &= u(\delta,t) = 0, \quad u(x,0) = u_0(x) \end{aligned} \tag{5.1}$$

where $w = r = 1$, $\delta = 10$ and $u_0 = 0$. There are homogeneous Dirichlet boundary conditions and the initial solution u_0 satisfies them. Before this problem is solved in time, the convection-diffusion operator is semi-discretized in space with a finite difference method. We apply an upwind scheme with backward differences for the first derivative, see Subsection 4.2.1. This and an approximation for the second derivative lead to $u''(x_i) \approx (1/h_{i+1})\left((u_{i+1}-u_i)/h_{i+1} - (u_i - u_{i-1})/h_i\right)$ and $u'(x_i) \approx (u_i - u_{i-1})/h_i$, where $u_i \approx u(x_i)$ and $\{x_i\}_{i=0}^{N-1}$ is a general mesh with the grid size $h_i = x_i - x_{i-1}$. Equidistant grids are sometimes acceptable discretizations, but in the convection-diffusion case it is often a better choice to use an irregular grid. For example, consider the analytical solution $u(x) = x - \delta\left(e^{(x-\delta)/\alpha} - e^{-1/\alpha}\right) / \left(1 - e^{-1/\alpha}\right)$ of the steady-state version of (5.1), i.e., $\partial u/\partial t = 0$. For $\alpha \to 0$, equation (5.1) becomes the simplest case of a singularly perturbed differential equation in $x = \delta$, compare with Figure 4.3. A reasonable approach is to have finer grids in regions of rapid change of the solution. Following this idea, we discretize the interval $[0, \delta]$ with the piecewise-uniform *Shishkin grid*, where more points are placed near the upcoming singularity as $\alpha \to 0$, see [39]. The transition point of this piecewise-uniform mesh is set to $\sigma = 1 - \min\{(4\alpha/\delta)\ln(N), 1/2\}$, see Subsection 6.4.1 in [19] with the description of Table 6.3. In particular, this leads to a uniform mesh if $\sigma = 1/2$.

The resulting finite difference matrix $L_h \in \mathbb{R}^{N\times N}$ depends on α and N and once the

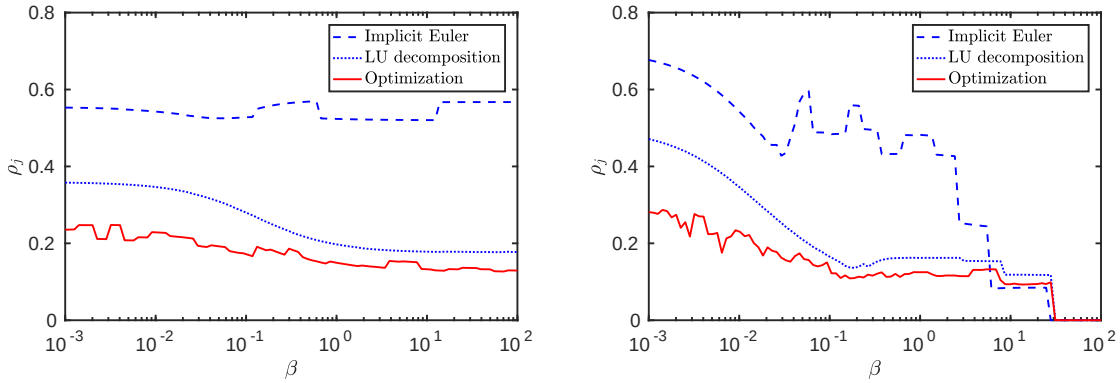

Fig. 5.1: Contraction factor ρ_j (5.4) of implicit Euler, LU decomposition and optimization based SDC methods for convection-diffusion problems (5.1) depending on $\beta = \alpha/(\tau w^2)$. Changing the problem by varying α where $\tau = 0.1$ (left), and varying τ where $\alpha = 0.001$ (right). Forcing a constant $\mathrm{Pe}_h < 1$ in the part of the finer Shishkin discretization. The optimization problem is given by (4.21).

spatial discretization has been done, we obtain the IVP

$$\begin{aligned} \frac{\partial u_h(t)}{\partial t} &= L_h u_h(t) + r_h \quad \text{in } (0,T), \\ u_h(0) &= u_{h,0}, \end{aligned} \tag{5.2}$$

where $r_h = [0,1,\dots,1,0]^T \in \mathbb{R}^N$ and $u_{h,0} = [0,\dots,0]^T \in \mathbb{R}^N$. The vector $u_h(t) \in \mathbb{R}^N$ approximates the solution $u(x_h,t)$ at the grid points $x_h = [x_0,...,x_{N-1}]^T$. The matrix L_h and the vector r_h are modified for the homogeneous Dirichlet boundary conditions. To apply the SDC framework as derived in the last chapters, the large time interval $[0,T]$ is subdivided and the IVP (5.2) is solved on the resulting subintervals $[0,\tau],[\tau,2\tau],\dots,[T-\tau,T]$. The SDC matrix formulation (2.12) with the constant Jacobian $J_f = L_h$ and exploiting the linearity of the right hand side of (5.2) lead to the SDC fixed point iteration

$$\begin{aligned} u_h^{[j+1]} &= G(Z)\, u_h^{[j]} + g, \\ G(Z) &= \left[I_N \otimes I_n - \left(I_N \otimes \hat{D} - Z \left(I_N \otimes \hat{S} \right) \right)^{-1} \left(I_N \otimes D - Z \left(I_N \otimes S^r \right) \right) \right], \\ g &= \left(I_N \otimes \hat{D} - Z \left(I_N \otimes \hat{S} \right) \right)^{-1} \left(\frac{\tau}{\tau_1} u_{h,0} \otimes e_1 + \tau \left(I_N \otimes S^r \right) \left(r_h \otimes \mathbf{1}_n \right) \right) \end{aligned} \tag{5.3}$$

on the first subinterval $[0,\tau]$, where $Z = \tau L_h \otimes I_n \in \mathbb{R}^{Nn \times Nn}$. The vector $u_h^{[j]} \in \mathbb{R}^{Nn}$, which is arranged as described in Section 2.2, approximates the solution $u(x,t)$ at a certain spatial grid $\{x_i\}_{i=0}^{N-1}$ and time grid $\{\tau_i\}_{i=0}^{n}$. The vector $\mathbf{1}_n = [1,\dots,1]^T \in \mathbb{R}^n$ and the identity matrix $I_n \in \mathbb{R}^n$ arise from the Radau-IIa discretization with s collocation points. In the following, the j-th approximation of this SDC fixed point iteration at the right interval end point $t_n = \tau$ is denoted by the vector $u_{h,\tau}^{[j]} \in \mathbb{R}^N$.
In the following, the Shishkin grids are constructed so that the matrices L_h have a dimension of 10001×10001 for all experiments. Due to the variable transition point

of these grids, a constant mesh Peclet number with $\mathrm{Pe}_h \approx 0.03$ is obtained in the finer discretized part, i.e., near the singularity. Based on this, we assume that the spatial discretizations for the convection-diffusion problems are highly accurate.
The first experiments concern the asymptotic behavior of the SDC methods where the SDC sweeps are performed with the abort condition $\|u_{h,\tau}^{[j+1]} - u_{h,\tau}^{[j]}\|_h \leq 10^{-10}$. The norm $\|\cdot\|_h$ is a discrete L^2-norm and takes the large differences in the Shishkin grid sizes into account, see Section 2.1 and, in particular, Definition (1.5) in [28]. This leads in the last performed SDC sweep to the contraction factor

$$\rho_j = \frac{\left\| u_{h,\tau}^{[j+1]} - u_{h,\tau}^{[j]} \right\|_h}{\left\| u_{h,\tau}^{[j]} - u_{h,\tau}^{[j-1]} \right\|_h}, \tag{5.4}$$

which is an approximation of the asymptotic contraction factor Φ_∞ of Definition 2.15. For the left plot of Figure 5.1, the diffusion coefficient α is varied, which leads to different β-parabola-regions, and $\tau = 0.1$ is fixed. The results of the experiments match with the study of Section 4.4. Due to the small grid sizes, the resulting matrices L_h have eigenvalues with a large negative real part. For implicit Euler based SDC methods, these stiff components of L_h combined with the moderate time interval length τ lead to a worse contraction factor ρ_j compared to LU decomposition based SDC methods. The direct optimization (4.21) is the best approach. The resulting SDC methods consist of matrices which take the β-parabola-region of the convection-diffusion problems into account, see Figure 4.8. It can be further discovered that the convergence of the LU decomposition based SDC methods become slower if β decreases. We observed this before in bottom plots of Figure 4.7 as the regions of maximum $\rho(G(z))$.
Similar results can be observed in the right plot of Figure 5.1, where we vary the time interval length τ and set $\alpha = 0.001$. The results for $\tau \to 0$ can be explained as follows: The β-parabola-regions become narrower by decreasing τ. Furthermore, by considering the fixed point iteration (5.3), there holds $Z(\tau \to 0) \to 0$ and thus $G(Z) \to 0$ as $\tau \to 0$. By decreasing τ, the stiff components of L_h have less influence on the convergence behavior of the fixed point iterations and this finally leads to $\rho_j \to 0$ as $\tau \to 0$.
A greater practical relevance than the asymptotic contraction factor may be that of the relative global error of SDC methods in the first few SDC sweeps. In the following, we consider two different settings for this. In the first experiments, SDC-7 methods are constructed, i.e., seven SDC sweeps are performed. The implicit Euler, LU decomposition and optimization approach is applied for different β by varying α and setting $\tau = 0.1$. This leads to different convection-diffusion problems. For all experiments, the reference solution is the collocation solution. The optimization approach is based on (4.23) with SDC sweep blocks of the size $m = 7$, see Subsection 3.2.4. For each β, the maximal global pre-asymptotic contraction factor is minimized after seven successively performed SDC sweeps. Thus, the minimized variable is $\left\| e_4^T G(z)^7 \right\|_2^{1/7}$ for $z \in \Pi_z(\beta)$. The results of the SDC-7 methods are presented in the left plot of Figure 5.2. The smallest relative global errors can be observed in the case of the optimization approach. Furthermore, the implicit Euler approach leads to significantly worse results compared to the LU decomposition approach.
The second experiments consider one certain convection-diffusion problem with a fixed $\beta = 0.01$ where $\alpha = 0.001$ and $\tau = 0.1$. For this problem, we construct SDC-j methods

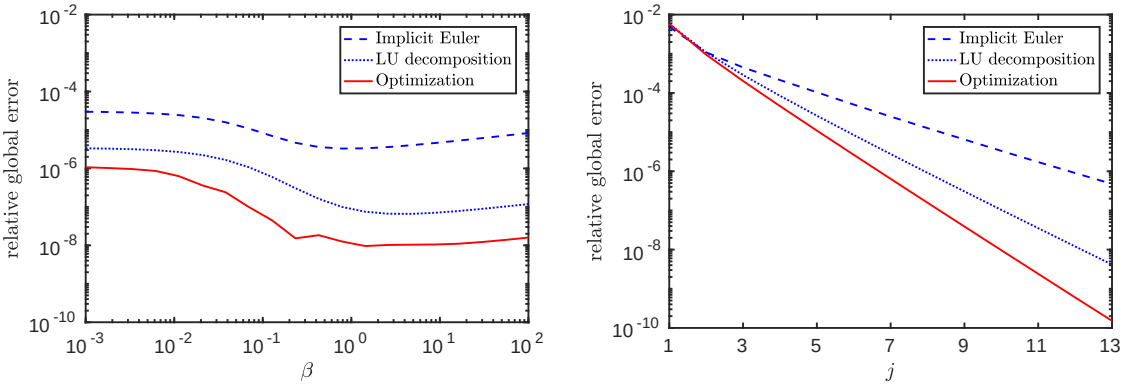

Fig. 5.2: Relative global error of SDC-7 methods for convection-diffusion problems (5.1) depending on $\beta = \alpha/(\tau w^2)$ where α is variable and $\tau = 0.1$ (left). Relative global error of SDC-j methods where $\beta = 0.01$ with $\alpha = 0.001$ and $\tau = 0.1$ (right). Forcing a constant $\mathrm{Pe}_h < 1$ in the part of the finer Shishkin discretization and the reference solution is the collocation solution. The optimization problem is given by (4.23) with SDC sweep blocks of size $m = 7$ (both plots), see Subsection 3.2.4.

and study the behavior of the relative global error. The results of these SDC-j methods are presented in the right plot of Figure 5.2. For $j \geq 2$ iterations, the LU decomposition and optimization based SDC methods have a lower relative global error than the implicit Euler based SDC methods. The direct optimization is the most promising approach for all SDC-j methods.

Remark 5.1. Considering inhomogeneous Dirichlet boundary conditions, only the vector f_h will change and the iteration matrix $G(Z)$ remains the same. Thus, the studies on $G(Z)$ are independent of the values of the Dirichlet boundary conditions and also of possible sources or sinks quantified by r_h.

Remark 5.2. Once there is a sufficient approximate solution on the first subinterval $[0, \tau]$, we can easily go to the next subintervals and start the SDC fixed point iteration again. For this, the last result of the SDC method at the end point of the current subinterval can be used as the initial solution for the new subinterval. A common choice for the initial guess $u_h^{[0]}$ is a constant initial solution for all grid points.

5.2 Two-dimensional finite element discretization

The following experiments cover a problem from Chapter 10 in [19]. This two-dimensional convection-diffusion problem is given by

$$\frac{\partial u}{\partial t} = \alpha \Delta u - w \cdot \nabla u \quad \text{in } \Omega \times (0, T) \tag{5.5}$$

where $\Omega = (-1, 1) \times (-1, 1)$ is a square domain. The wind $w(x, y)$ is recirculating on this domain and is of the form $w(x, y) = \hat{w}[2y(1 - x^2), -2x(1 - y^2)]^T$, where we added a scalar parameter $\hat{w} \in \mathbb{R}^+$. There are no sources or sinks present and Dirichlet boundary conditions are imposed everywhere. From the limit $t \to \infty$ follows the steady-state case

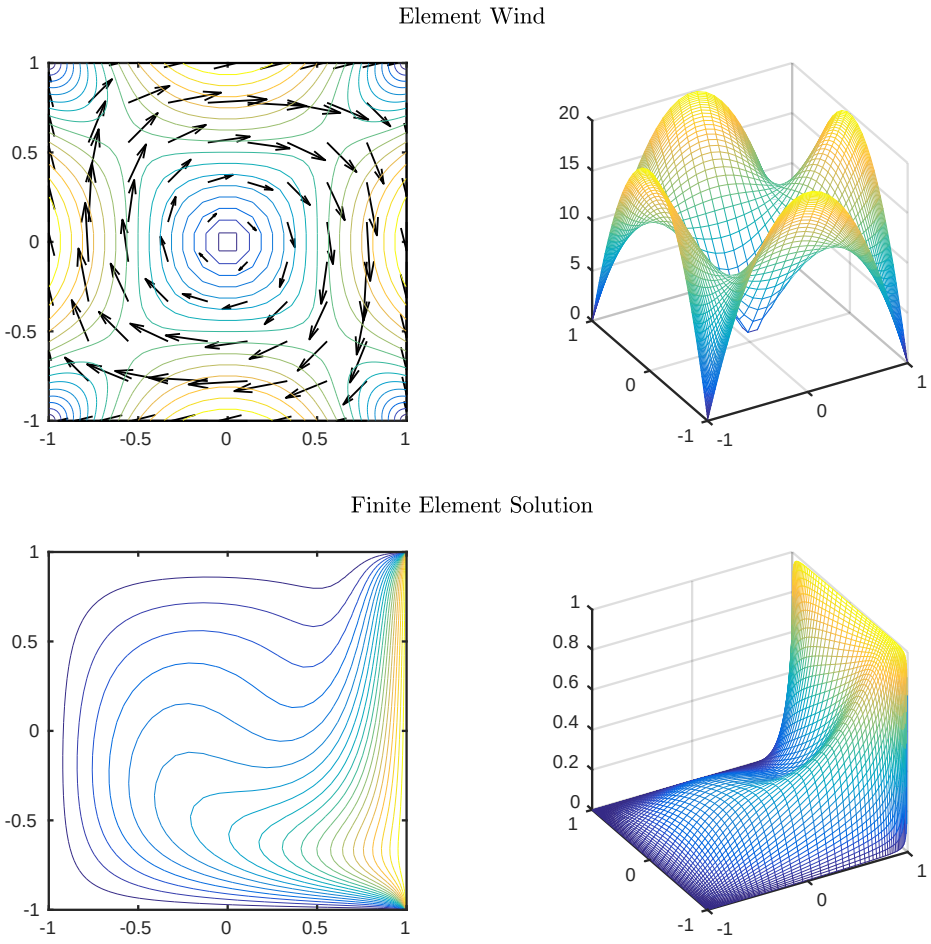

Fig. 5.3: The 2-norm of the wind at each element (top) and the steady-state finite element solution of the convection-diffusion problem (5.5) (bottom) for $\alpha = 1$ and $\hat{w} = 10$. Isolines are shown in the left plots and a 64×64 stretched mesh is used.

with $u(x,-1,t) = u(x,1,t) = u(-1,y,t) = 0$ and $u(1,y,t) = 1$, see Chapter 6 in [19] with the example 6.1.4. The initial condition is given by $u(x,y,0) = 0$ and the Dirichlet boundary conditions change in time by the factor $1 - e^{-10t}$ so that the values of the boundary conditions smoothly go to the values of the steady-state case. This problem, the so-called *double-glazing problem*, can be interpreted as follows: The solution could be a temperature field in a cavity with a hot wall at $x = 1$. The wind distributes the heat in the cavity in a recirculating way. There are two discontinuities at the hot wall in $x = 1$ and $y = \pm 1$ and they lead to boundary layers. For further information on this problem and in particular on the mathematical meaning of its boundary layers we refer to Chapter 6 in [19].
In this section, the convection-diffusion operator is semi-discretized in space by finite elements. For this, the open-source software package IFISS [18],[51] is applied. This MATLAB® implementation is developed by the authors of [19]. The meshes consist of rectangular elements, see Subsection 1.3.2 in [19], and they are stretched so that there

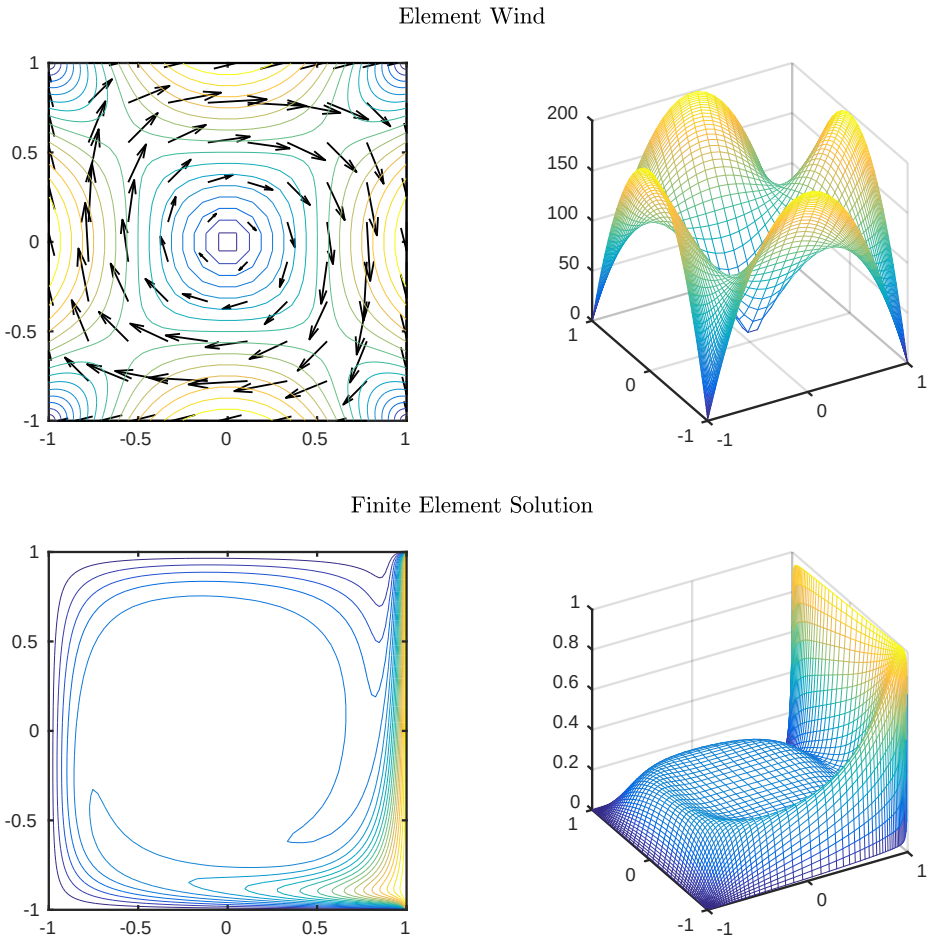

Fig. 5.4: The 2-norm of the wind at each element (top) and the steady-state finite element solution of the convection-diffusion problem (5.5) (bottom) for $\alpha = 1$ and $\hat{w} = 100$. Isolines are shown in the left plots and a 64×64 stretched mesh is used.

are more elements near the boundaries. If the mesh Peclet number Pe_h becomes too large for some elements, the streamline diffusion method is applied, see Subsection 4.2.2. We refer to the Sections 6.2 and 6.3 of [19] for further information on the functionality of the IFISS code for convection-diffusion problems. Two specific examples of finite element solutions, which are computed by the IFISS software, are plotted in the Figures 5.3 and 5.4. The 2-norm of the wind $w(x, y) = \hat{w}[2y(1 - x^2), -2x(1 - y^2)]^T$ at each element and furthermore, the steady-state finite element solution for a fixed $\alpha = 1$ and different $\hat{w}$ are presented. In Figure 5.3, where $\alpha = 1$ and $\hat{w} = 10$, the diffusion has a large enough influence to prevent visible recirculating flows in the bottom plots. In Figure 5.4, the convection effect becomes stronger by setting $\hat{w} = 100$. This leads to plots which are more typical for a double-glazing problem. The implementation of the SDC methods is done a way analogous to the previous section, i.e., the discrete operator L_h obtained by the IFISS software is applied for fixed point iterations, which are similiar to (5.3). We set $\alpha = 1$ for all following experiments.

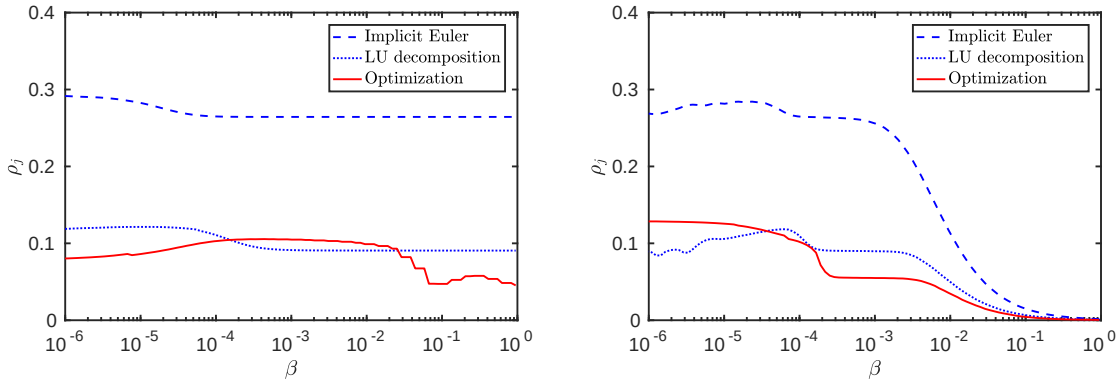

Fig. 5.5: Contraction factor ρ_j (5.4) of implicit Euler, LU decomposition and optimization based SDC methods for convection-diffusion problems (5.5) depending on β. Changing the problem by varying $\hat{w}$ where $\tau = 1$ and $\alpha = 1$ (left). Varying τ where $\hat{w} = 100$ and $\alpha = 1$ (right). The optimization problem is given by (4.21).

The first experiments address the asymptotic behavior of SDC methods. For this, the contraction factor ρ_j (5.4) is computed where the SDC fixed point iterations have the abort condition $||u_{h,\tau}^{[j+1]} - u_{h,\tau}^{[j]}||_h \leq 10^{-12}$ in the last SDC sweep. For the computation of $\beta = \alpha/(\tau W^2)$, see Definition 4.7, we set W to the mean of $\|w(x,y)\|_2$ for $(x,y) \in \Omega$. The results of varying $\hat{w}$ are presented in the left plot of Figure 5.5. As for the experiments regarding the finite difference discretization, a significant benefit by the LU decomposition and optimization approach is obtained compared to implicit Euler SDC methods. The optimization approach leads to the best results for $\beta \to 0$ and $\beta \to 1$, but for intermediate values, there is a slower convergence speed than for LU decomposition based SDC methods. A possible reason for this is the choice of the scalar value W for the β-parabola-region. The problem (5.5) has a variable wind $w(x,y)$ and thus the computation of β with the mean of $\|w(x,y)\|_2$ is not the optimal choice for all possible β. A further observation is that in both cases of the LU decomposition and implicit Euler approach decreasing β leads to worse contraction factors ρ_j. We assume again that the reason for this are the regions of maximum $\rho(G(z))$, which are observed before in Figure 4.7. For the right plot of Figure 5.5, we vary τ and this results in $\rho_j \to 0$ as $\beta \to 1$, which was already discovered in the last section.

The next experiments cover the relative global error of different SDC methods for the double-glazing problem (5.5). Compared to the implicit Euler based SDC methods, the LU decomposition and optimization based SDC methods have a smaller relative global error after seven performed SDC sweeps, see the left plot of Figure 5.6. The direct optimization approach is the most promising approach and this holds for all β. As in the previous section, the maximal global pre-asymptotic contraction factor is minimized for SDC sweep blocks of size $m = 7$. In the right plot of Figure 5.6, the relative global error is presented over the number j of SDC sweeps. In this experiment, the β-parabola-region is fixed with $\beta \approx 0.0001$ where $\alpha = 1$, $\hat{w} = 100$ and $\tau = 1$. We optimize again for a fixed SDC sweep block size $m = 7$. The LU decomposition and optimization approach lead to smaller relative global errors than the implicit Euler approach, for all SDC-j methods.

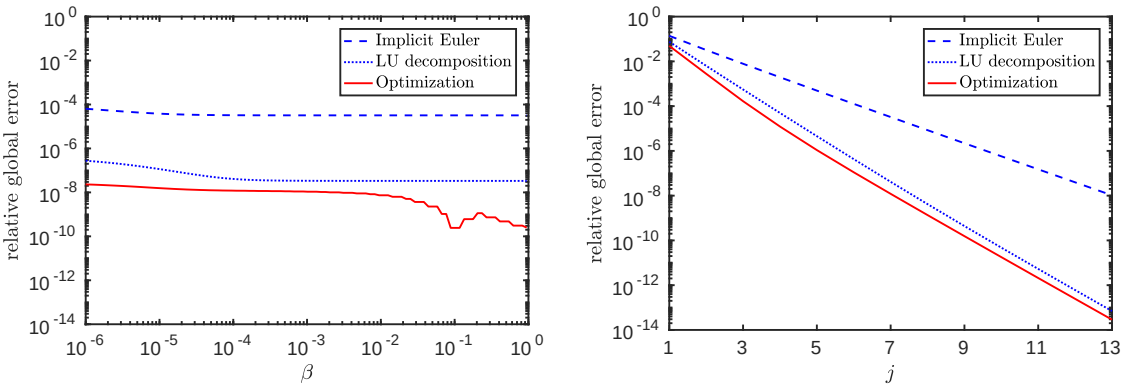

Fig. 5.6: Relative global error of SDC-7 methods for convection-diffusion problems (5.5) depending on β where $\hat{w}$ is variable, $\tau = 1$ and $\alpha = 1$ (left). Relative global error of SDC-j methods where $\beta \approx 0.0001$ with $\alpha = 1$, $\tau = 1$ and $\hat{w} = 100$ (right). The reference solution is the collocation solution. The optimization problem is given by (4.23) with SDC sweep blocks of size $m = 7$ (both plots), see Subsection 3.2.4.

Remark 5.3. To get different convection-diffusion problems for the experiments concerning the finite element discretization, we vary the wind w by choosing $\hat{w}$. Forcing a stronger convection by decreasing α leads to the result that for all β, a similar convergence behavior is obtained. It makes no difference whether using implicit Euler, LU decomposition or optimization based SDC methods. The reason for this observation is that the matrices L_h have no stiff components for very small α and reasonable fine discretizations. This leads to the interesting result that all the constructed SDC methods for problems with $\|w\|_2 = \mathcal{O}(1)$ seem to work very well for a large range of α. For the two-dimensional case in Section 5.2, the use of very small grid sizes leads to very large linear systems. It can be difficult to solve these systems.

6 Conclusion

The approaches in [55] with application to the problem class of reaction-diffusion equations work for convection-diffusion problems as well. Several experiments demonstrate that it is possible to accelerate the convergence speed and to improve the error reduction in the first few SDC sweeps. In addition to this experimental study, a contribution of this thesis is the derivation of a framework for constructing suitable approximate differentiation and integration matrices of SDC methods for convection-diffusion problems. This framework uses the β-parabola-region of Definition 4.7 and is the basis for the direct optimizations (4.21), (4.22) and (4.23). Although the β-parabola-region is based on a simple one-dimensional convection-diffusion operator with constant coefficients, the numerical solution of more realistic convection-diffusion problems in Section 5.2 leads to promising results. The direct optimization with respect to SDC sweep blocks is the best approach compared to linearly implicit Euler and LU decomposition based SDC methods. The direct optimization approach applies a local optimization with initial matrices from the LU decomposition approach.
On the other hand, the LU decomposition approach and its implementation are very simple. The experiments show significant benefits compared to implicit Euler based SDC methods, most of all for problems with very stiff components in the eigenvalues of the partial differential operator. In the limit case of stiff problems, the asymptotic contraction factor of the LU decomposition based SDC methods is equal to zero, see Lemma 4.9. Furthermore, we prove in Theorem 3.6 that the LU decomposition of the spectral integration matrix is unique and thus pivoting never has to be considered. This guarantees that the SDC methods run forward in time.
If a convection-diffusion operator $\mathcal{L} = \alpha\Delta + w \cdot \nabla$ has non-constant coefficients, there is more than one possibility to compute the β-parabola-region for the considered problem. To achieve a satisfying framework for this, other experiments and considerations are necessary. Further improvements are conceivable by applying the direct optimization approach with such a resulting β-parabola-region. Additionally, the analysis of the arising approximate differentiation and integration matrices is a reasonable next step. This offers a deeper insight into the SDC sweep structure, which is a worthwhile goal in its own right, and it is also the basis for an efficient implementation of SDC methods.

References

[1] Kendall Atkinson and Weimin Han. *Theoretical numerical analysis. A functional analysis framework.* 3rd ed. Berlin: Springer-Verlag, 2009, pp. xvi + 625.

[2] W. Auzinger et al. "Modified defect correction algorithms for ODEs. II: Stiff initial value problems." In: *Numer. Algorithms* 40.3 (2005), pp. 285–303.

[3] G.K. Batchelor. *An introduction to fluid dynamics.* 2nd pbk-ed. Cambridge: Cambridge University Press, 1999, pp. xviii + 615.

[4] Matthias Bolten, Dieter Moser, and Robert Speck. "A multigrid perspective on the parallel full approximation scheme in space and time." In: *Numer. Linear Algebra Appl.* (Submitted on 11 Mar 2016), pp. 1–26.

[5] Matthias Bolten, Dieter Moser, and Robert Speck. "Asymptotic convergence of the parallel full approximation scheme in space and time for linear problems." In: *Numer. Linear Algebra Appl.* (Submitted on 21 Mar 2017), pp. 1–24.

[6] Sunyoung Bu, Jingfang Huang, and Michael L. Minion. "Semi-implicit Krylov deferred correction methods for differential algebraic equations." In: *Math. Comput.* 81.280 (2012), pp. 2127–2157.

[7] J. C. Butcher. *Numerical methods for ordinary differential equations.* 2nd revised ed. Hoboken, NJ: John Wiley & Sons, 2008, pp. xix + 463.

[8] Andrew J. Christlieb, Colin B. MacDonald, and Benjamin W. Ong. "Parallel high-order integrators." In: *SIAM J. Sci. Comput.* 32.2 (2010), pp. 818–835.

[9] E.B. Davies. "Semi-classical analysis and pseudo-spectra." In: *J. Differ. Equations* 216.1 (2005), pp. 153–187.

[10] E.B. Davies. *Spectral theory and differential operators.* Cambridge: Cambridge University Press, 1995, pp. ix + 182.

[11] Peter Deuflhard and Folkmar Bornemann. *Scientific computing with ordinary differential equations.* Transl. from the German by Werner C. Rheinboldt. New York: NY: Springer, 2002, pp. xix + 485.

[12] Peter Deuflhard and Andreas Hohmann. *Numerical analysis in modern scientific computing. An introduction.* 2nd rev. ed. New York: NY: Springer, 2003, pp. xviii + 337.

[13] Peter Deuflhard and Martin Weiser. *Adaptive numerical solution of PDEs.* Berlin: de Gruyter, 2012, pp. xi + 421.

[14] Nelson Dunford and Jacob T. Schwartz. *Linear operators. Part I: General theory. With the assistance of William G. Bade and Robert G. Bartle.* Repr. of the orig., publ. 1959 by John Wiley & Sons Ltd., Paperback ed. New York etc.: John Wiley & Sons Ltd., 1988, pp. xiv + 858.

[15] Nelson Dunford and Jacob T. Schwartz. *Linear operators. Part II: Spectral theory, self adjoint operators in Hilbert space.* Repr. of the orig., publ. 1963 by John Wiley & Sons Ltd., Paperback ed. New York etc.: John Wiley & Sons Ltd./Interscience Publishers, Inc., 1988, pp. ix + 1064 +7.

[16] Nelson Dunford and Jacob T. Schwartz. *Linear operators. Part III: Spectral operators. With the assistance of William G. Bade and Robert G. Bartle.* Repr. of the orig., publ. 1971 by John Wiley & Sons Ltd., Paperback ed. New York etc.: John Wiley & Sons Ltd./Interscience Publishers, Inc., 1988, pp. xix + 667.

[17] Alok Dutt, Leslie Greengard, and Vladimir Rokhlin. "Spectral deferred correction methods for ordinary differential equations." In: *BIT* 40.2 (2000), pp. 241–266.

[18] Howard C. Elman, Alison Ramage, and David J. Silvester. "IFISS: A computational laboratory for investigating incompressible flow problems." In: *SIAM Rev.* 56.2 (2014), pp. 261–273.

[19] Howard C. Elman, David J. Silvester, and Andrew J. Wathen. *Finite elements and fast iterative solvers. With applications in incompressible fluid dynamics.* 2nd ed. Oxford: Oxford University Press, 2014, pp. xiv + 479.

[20] Matthew Emmett and Michael L. Minion. "Toward an efficient parallel in time method for partial differential equations." In: *Commun. Appl. Math. Comput. Sci.* 7.1 (2012), pp. 105–132.

[21] Lawrence C. Evans. *Partial differential equations.* 2nd ed. Providence, RI: American Mathematical Society (AMS), 2010, pp. xxi + 749.

[22] Lisa Fischer, Sebastian Götschel, and Martin Weiser. "Lossy data compression reduces communication time in hybrid time-parallel integrators". In: *Tech. Rep. ZIB-Report* 17-25 (2017).

[23] Pascal Jean Frey and Paul-Louis George. *Mesh generation. Application to finite elements.* 2nd ed. London: ISTE; Hoboken, NJ: John Wiley & Sons, 2008, pp. 848.

[24] Seymour Goldberg. *Unbounded linear operators. Theory and applications.* Reprint of the 1966 edition. New York, NY: Dover Publications, Inc., 1985, pp. viii + 199.

[25] Gene H. Golub and Charles F. Van Loan. *Matrix computations.* 4th ed. Baltimore, MD: The Johns Hopkins University Press, 2013, pp. xxi + 756.

[26] David Gottlieb and Steven A. Orszag. *Numerical analysis of spectral methods: Theory and applications.* CBMS-NSF Regional Conference Series in Applied Mathematics. 26. Philadelphia. Philadelphia, PA: SIAM Society for Industrial and Applied Mathematics, 1977, pp. v+170.

[27] L. Greengard. "Spectral integration and two-point boundary value problems." In: *SIAM J. Numer. Anal.* 28.4 (1991), pp. 1071–1080.

[28] Christian Grossmann, Hans-Görg Roos, and Martin Stynes. *Numerical treatment of partial differential equations.* Berlin: Springer-Verlag, 2007, pp. xii + 591.

[29] Thomas Hagstrom and Ruhai Zhou. "On the spectral deferred correction of splitting methods for initial value problems." In: *Commun. Appl. Math. Comput. Sci.* 1 (2006), pp. 169–205.

[30] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration. Structure-preserving algorithms for ordinary differential equations.* Berlin: Springer-Verlag, 2002, pp. xiii + 515.

[31] Ernst Hairer, Syvert P. Nørsett, and Gerhard Wanner. *Solving ordinary differential equations. I: Nonstiff problems.* 2nd rev. ed. Berlin: Springer-Verlag, 1993, pp. xv + 528.

[32] Ernst Hairer and Gerhard Wanner. *Solving ordinary differential equations. II: Stiff and differential-algebraic problems.* Reprint of the 1996 2nd rev. ed. Berlin: Springer-Verlag, 2010, pp. xvi + 614.

[33] Jingfang Huang, Jun Jia, and Michael Minion. "Accelerating the convergence of spectral deferred correction methods." In: *J. Comput. Phys.* 214.2 (2006), pp. 633–656.

[34] Jingfang Huang, Jun Jia, and Michael Minion. "Arbitrary order Krylov deferred correction methods for differential algebraic equations." In: *J. Comput. Phys.* 221.2 (2007), pp. 739–760.

[35] Willem Hundsdorfer and Jan Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations.* Berlin: Springer-Verlag, 2003, pp. x + 471.

[36] Jun Jia and Jingfang Huang. "Krylov deferred correction accelerated method of lines transpose for parabolic problems." In: *J. Comput. Phys.* 227.3 (2008), pp. 1739–1753.

[37] Tosio Kato. *Perturbation theory for linear operators.* 2nd corr. print. of the 2nd ed. Grundlehren der Mathematischen Wissenschaften, 132. Berlin: Springer-Verlag., 1984, pp. XXI + 619.

[38] Peter Knabner and Lutz Angermann. *Numerical methods for elliptic and parabolic partial differential equations. Translation from the German.* New York: NY: Springer, 2003, pp. xv + 424.

[39] Natalia Kopteva and Eugene O'Riordan. "Shishkin meshes in the numerical solution of singularly perturbed differential equations." In: *Int. J. Numer. Anal. Model.* 7.3 (2010), pp. 393–415.

[40] Anita T. Layton and Michael L. Minion. "Implications of the choice of quadrature nodes for Picard integral deferred corrections methods for ordinary differential equations." In: *BIT* 45.2 (2005), pp. 341–373.

[41] Jörg Liesen and Volker Mehrmann. *Linear algebra.* Cham: Springer, 2015, pp. xi + 324.

[42] Jörg Liesen and Zdeněk Strakoš. *Krylov subspace methods. Principles and analysis.* Oxford: Oxford University Press, 2013, pp. xv + 391.

[43] Michael L. Minion. "A hybrid parareal spectral deferred corrections method." In: *Commun. Appl. Math. Comput. Sci.* 5.2 (2010), pp. 265–301.

[44] K.W. Morton. *Numerical solution of convection-diffusion problems.* London: Chapman & Hall, 1996, pp. xii + 372.

[45] Satish C. Reddy and Lloyd N. Trefethen. "Pseudospectra of the convection-diffusion operator." In: *SIAM J. Appl. Math.* 54.6 (1994), pp. 1634–1649.

[46] Satish C. Reddy, Lloyd N. Trefethen, and D. Pathria. "Pseudospectra of the convection-diffusion operator (extended version)." In: *Tech. Rep. CTC93TR126, Cornell Theory Center* (1993), pp. 1–32.

[47] Lothar Reichel and Lloyd N. Trefethen. "Eigenvalues and pseudo-eigenvalues of Toeplitz matrices." In: *Linear Algebra Appl.* 162-164 (1992), pp. 153–185.

[48] Daniel Ruprecht and Robert Speck. "Spectral deferred corrections with fast-wave slow-wave splitting." In: *SIAM J. Sci. Comput.* 38.4 (2016), a2535–a2557.

[49] Yousef Saad. *Iterative methods for sparse linear systems.* 2nd ed. Philadelphia, PA: SIAM Society for Industrial and Applied Mathematics, 2003, pp. xviii + 528.

[50] Heinz Schade and Ewald Kunz. *Strömungslehre. Bearbeitet von Frank Kameier und Christian Oliver Paschereit.* German. 3rd ed. Berlin: de Gruyter, 2007, pp. xv + 558.

[51] David Silvester, Howard Elman, and Alison Ramage. *Incompressible Flow and Iterative Solver Software (IFISS) version 3.5.* http://www.manchester.ac.uk/ifiss/. 2016.

[52] Lloyd N. Trefethen. "Pseudospectra of linear operators." In: *SIAM Rev.* 39.3 (1997), pp. 383–406.

[53] Lloyd N. Trefethen and David III Bau. *Numerical linear algebra.* Philadelphia, PA: SIAM Society for Industrial and Applied Mathematics, 1997, pp. xii + 361.

[54] Lloyd N. Trefethen and Mark Embree. *Spectra and pseudospectra. The behavior of nonnormal matrices and operators.* Princeton, NJ: Princeton University Press, 2005, pp. xviii + 606.

[55] Martin Weiser. "Faster SDC convergence on non-equidistant grids by DIRK sweeps." In: *BIT* 55.4 (2015), pp. 1219–1241.

[56] Martin Weiser. *Inside finite elements.* Berlin: de Gruyter, 2016, pp. xi + 146.

[57] Martin Weiser and Sunayana Ghosh. "Theoretically optimal inexact SDC methods." In: *Tech. Rep. ZIB-Report* 16-52 (2016, submitted to *Commun. Appl. Math. Comput. Sci.*).

[58] Martin Weiser and Simone Scacchi. "Spectral Deferred Correction methods for adaptive electro-mechanical coupling in cardiac simulation." In: *Tech. Rep. ZIB-Report* 14-22 (2014).

[59] O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu. *The finite element method: its basis and fundamentals.* 7th ed. Amsterdam: Elsevier/Butterworth Heinemann, 2013, pp. xxxviii + 714.