Communications in Applied Mathematics and Computational Science

THEORETICALLY OPTIMAL INEXACT SPECTRAL DEFERRED CORRECTION METHODS

MARTIN WEISER AND SUNAYANA GHOSH

vol. 13 no. 1 2018

msp

COMM. APP. MATH. AND COMP. SCI.
Vol. 13, No. 1, 2018
dx.doi.org/10.2140/camcos.2018.13.53

THEORETICALLY OPTIMAL INEXACT SPECTRAL DEFERRED CORRECTION METHODS

MARTIN WEISER AND SUNAYANA GHOSH

In several initial value problems with particularly expensive right-hand side evaluation or implicit step computation, there is a tradeoff between accuracy and computational effort. We consider inexact spectral deferred correction (SDC) methods for solving such initial value problems. SDC methods are interpreted as fixed-point iterations and, due to their corrective iterative nature, allow one to exploit the accuracy-work tradeoff for a reduction of the total computational effort. First we derive error models bounding the total error in terms of the evaluation errors. Then we define work models describing the computational effort in terms of the evaluation accuracy. Combining both, a theoretically optimal local tolerance selection is worked out by minimizing the total work subject to achieving the requested tolerance. The properties of optimal local tolerances and the predicted efficiency gain compared to simpler heuristics, and reasonable practical performance, are illustrated with simple numerical examples.

1. Introduction

The numerical solution of initial value problems of the form

$$y'(t) = f(y(t)), \quad y(0) = y_0,$$

can involve a significant amount of computation, where the most effort is usually spent either on evaluating complex right-hand sides in nonstiff problems or on solving large linear equation systems in stiff systems. Often, there is an accuracy-effort tradeoff, such that inexact results can be obtained much faster than exact results. Examples for the first type of problem are molecular and stellar dynamics, where the exact evaluation of long-range interactions is $\mathbb{O}(N^2)$ but can be approximated by clustering or fast multipole methods in $\mathbb{O}(N \log N)$ or $\mathbb{O}(N)$ time [4; 6], or cycle jump techniques for highly oscillatory problems of wear or fatigue [10; 13]. Typical examples of the second type of problem are reaction-diffusion equations, where implicit time-stepping schemes rely on iterative solvers [24; 28],

A preprint of this paper appeared as Zuse Institute Berlin report 16-52.
MSC2010: 65L05, 65L20, 65L70, 65M70.
Keywords: spectral deferred corrections, initial value problems, error propagation, adaptive control of tolerances, inexact, work models, accuracy models.

While the possibilities to exploit the tradeoff between accuracy and computational effort for improved simulation performance are rather limited in usual time-stepping schemes such as explicit or implicit Runge–Kutta, extrapolation, or multistep schemes, iterative methods for solving implicit Runge–Kutta equations [3; 12; 26] can in principle correct inexact evaluations of intermediate quantities in subsequent iterations. Spectral deferred correction (SDC) methods [11] as iterative solvers for collocation systems have a particularly simple structure and are therefore considered here. Inexact implicit SDC methods with errors due to truncation of multigrid iterations have been investigated numerically in [20; 24], where a small fixed number of V-cycles has been found to be sufficient for convergence. Mixed-precision arithmetic for SDC has been proposed in [15] and found to save some computational effort. In this paper, we will analyze the error propagation through the SDC iteration and, following the approach of Alfeld [1] for inexact fixed-point iterations, derive an a priori selection of local tolerances for right-hand side evaluation and substep computation that leads to theoretically optimal efficiency of the overall integration scheme. Usually, explicit Runge–Kutta methods are hard to beat in efficiency by more complex methods such as SDC, but the results derived here indicate that this might be possible if inexactness can be exploited.

The remainder of the paper is organized as follows. Section 2 states the precise problem setting before briefly recalling spectral deferred correction methods and discussing the impact of inexact evaluations. The main Section 3 introduces error models for quantifying the error propagation, work models for quantifying the computational cost, and the optimization of accuracy per work to derive an optimal selection of tolerances. Effectiveness and efficiency of the resulting methods are illustrated in Section 4 with some numerical examples.

2. Inexactness in spectral deferred correction methods

The autonomous initial value problem (IVP) to be solved is given by

$$\begin{cases} y'(t) = f(y(t)), \quad t \in [0, T], \\ y(0) = y_0, \end{cases} \tag{2-1}$$

where the right-hand side f is a mapping $f : Y \to Y$ on a Banach space Y, and $t \in [0, T]$ denotes the time variable. It is assumed that f is continuous and locally Lipschitz continuous. Under these assumptions, a unique solution $y(t)$ exists; see, e.g., [9; 25]. An approximate numerical solution can be determined with time-stepping schemes. We consider single-step methods, where the time interval $[0, T]$ is subdivided into individual steps and the connection between the subintervals consists of transferring the value of y at the end point of one subinterval as the initial value for the following subinterval. Without loss of generality, we therefore

restrict the presentation to a single time step $[0, T]$. Also, without loss of generality, we assume (2-1) to be autonomous.

2A. ***Collocation conditions.*** Given the IVP (2-1), a collocation method approximates the exact solution y over the interval $[0, T]$ by a polynomial y_c satisfying (2-1) at N discrete collocation points t_i, $i = 1, \ldots, N$, within the interval $[0, T]$:

$$\begin{cases} y_c'(t_i) = f(y_c(t_i)), & i = 1, \ldots, N, \\ y_c(0) = y_0. \end{cases} \tag{2-2}$$

For simplicity of indexing, we define $t_0 = 0$. Popular choices for collocation points are equidistant nodes or Gauss–Legendre, Lobatto, or Radau points. For a detailed discussion of collocation methods, we refer to [9; 17].

The IVP (2-1) can be written equivalently as the Picard integral equation

$$y(t) = y_0 + \int_0^t f(y(\tau))\, d\tau,$$

which leads to corresponding Picard collocation conditions, as described in [18]:

$$\begin{cases} y_c(t_i) = y_c(t_{i-1}) + \sum_{k=1}^{N} S_{ik} f(y_c(t_k)), & i = 1, \ldots, N, \\ y_c(0) = y_0, \end{cases} \tag{2-3}$$

where the entries of the spectral quadrature matrix $S \in \mathbb{R}^{N \times N}$ are defined in terms of the Lagrange polynomials $L_k \in \mathbb{P}_{N-1}[\mathbb{R}]$ satisfying $L_k(t_i) = \delta_{ik}$ for $i = 1, \ldots, N$ as

$$S_{ik} = \int_{\tau = t_{i-1}}^{t_i} L_k(\tau)\, d\tau, \quad i, k = 1, \ldots, N.$$

2B. ***Spectral deferred correction method.*** The direct solution of the collocation system (2-2) or (2-3) can be quite involved if N is larger than one or two. As the time discretization error of the collocation method is present anyway, an exact solution of (2-2) is not required. Thus, iterative methods form an interesting class of solvers; see, e.g., [7; 8; 19]. Here we consider spectral deferred correction (SDC) methods. They were introduced by Dutt, Greengard, and Rokhlin [11] for fixed iteration number as time-stepping schemes in their own right, and only later on have been interpreted as fixed-point iterations for collocation systems [18; 27]. In SDC, the Picard collocation conditions (2-3) are solved iteratively by a defect-correction procedure. Using the Picard formulation has the advantage of faster convergence for nonstiff problems [27].

Approximate solutions are polynomials $y^{[j]} \in \mathbb{P}_N[Y]$, identified with vectors in Y^{N+1} by interpolation of their values $y_i^{[j]} := y^{[j]}(t_i)$ at the $N + 1$ grid points t_i. Given an approximate solution $y^{[j]}$, the *error* $y_c - y^{[j]}$ satisfies the Picard collocation

conditions

$$y_c(t_i) - y_i^{[j]} = (y_c - y^{[j]})(t_{i-1}) + \sum_{k=1}^{N} S_{ik}(f(y_c(t_k)) - y^{[j]\prime}(t_k))$$
$$= (y_c - y^{[j]})(t_{i-1}) + \sum_{k=1}^{N} S_{ik}(f(y_c(t_k)) - f(y_k^{[j]}))$$
$$+ \sum_{k=1}^{N} S_{ik}(f(y_k^{[j]}) - y^{[j]\prime}(t_k)) \quad (2\text{-}4)$$

for $i = 1, \ldots, N$ with initial condition $(y_c - y^{[j]})(0) = 0$. Defining the correction $d^{[j]} = y_c - y^{[j]}$ yields

$$d^{[j]}(t_i) = d^{[j]}(t_{i-1}) + \sum_{k=1}^{N} S_{ik}(f(y^{[j]}(t_k) + d^{[j]}(t_k)) - f(y_k^{[j]}))$$
$$+ \sum_{k=1}^{N} S_{ik} f(y_k^{[j]}) - (y_i^{[j]} - y_{i-1}^{[j]}),$$

which is not easier to solve than the original collocation problem (2-2) above. Different simple approximations of the middle integration term involving $d^{[j]}$, however, at least provide corrections that can be applied repeatedly to form a convergent stationary iteration.

Explicit SDC. Approximating the spectral integration term by the left-looking rectangular rule corresponding to the explicit Euler time-stepping scheme yields the explicit SDC correction

$$\delta_i^{[j]} = \delta_{i-1}^{[j]} + (t_i - t_{i-1})(f(y_{i-1}^{[j]} + \delta_{i-1}^{[j]}) - f(y_{i-1}^{[j]}))$$
$$+ \sum_{k=1}^{N} S_{ik} f(y_k^{[j]}) - (y_i^{[j]} - y_{i-1}^{[j]}), \quad i = 1, \ldots, N, \quad (2\text{-}5)$$

suitable for nonstiff problems. The initial value is $\delta_0^{[j]} = 0$. Now, the interpolant $\delta^{[j]}$ is a polynomial approximation of the exact error function $d^{[j]}$. An improved approximation $y^{[j+1]}$ is then obtained as $y^{[j+1]} = y^{[j]} + \delta^{[j]}$. Note that the value $f(y_{i-1}^{[j]} + \delta_{i-1}^{[j]})$ appears again as $f(y_{i-1}^{[j+1]})$ in the next iteration, such that for each iteration only N right-hand side evaluations are required.

The expensive part in the explicit SDC method is usually the evaluation of the right-hand sides $f(y_i^{[c]})$. As mentioned above, an exact evaluation of the right-hand side $f(y_i^{[j]})$ is not necessary, because SDC iteration errors are already present due to the replacement of the spectral quadrature term by the rectangular rule. If approximate values $f_i^{[j]} \approx f(y_i^{[j]})$ can be computed faster, we can exploit the allowed inaccuracy for a reduction of the total computation effort.

It is clear that the evaluation error $f_i^{[j]} - f(y_i^{[j]})$ must be controlled in an appropriate way such as not to destroy convergence of the fixed-point scheme. We assume that for evaluation of $f(y_i^{[j]})$ we can prescribe a local absolute tolerance $\epsilon_i^{[j]}$ such that the computed value $f_i^{[j]}$ satisfies $\|f_i^{[j]} - f(y_i^{[j]})\|_Y \le \epsilon_i^{[j]}$.

As a consequence, the explicit SDC correction $\hat{\delta}^{[j]}$ for inexact right-hand sides $f_i^{[j]}$ is obtained as

$$\hat{\delta}_i^{[j]} = \hat{\delta}_{i-1}^{[j]} + (t_i - t_{i-1})(f_{i-1}^{[j+1]} - f_{i-1}^{[j]}) + \sum_{k=1}^{N} S_{ik} f_k^{[j]} - (y_i^{[j]} - y_{i-1}^{[j]}), \qquad (2\text{-}6)$$

for $j = 0, \ldots, J-1$ and $i = 1, \ldots, N$ with $\hat{\delta}_0^{[j]} = 0$.

Implicit SDC. Assuming f is differentiable, linearizing f around $y_i^{[j]}$ and using the right-looking rectangular rule corresponds to the linearly implicit Euler scheme and leads to the implicit SDC correction

$$\delta_i^{[j]} = \delta_{i-1}^{[j]} + (t_i - t_{i-1}) f'(y_i^{[j]})\delta_i^{[j]} + \sum_{k=1}^{N} S_{ik} f(y_k^{[j]}) - (y_i^{[j]} - y_{i-1}^{[j]}), \quad i = 1, \ldots, N, \qquad (2\text{-}7)$$

suitable for stiff problems. As in the explicit case, N right-hand side evaluations are required, but additionally N evaluations of f' and N linear system solves with the matrices $I - (t_i - t_{i-1}) f'(y_i^{[j]})$.

Solving these systems, usually by an iterative solver, is often the expensive operation in the implicit SDC method. Early termination of the linear solver can reduce the computational effort significantly, but incurs a truncation error that must be controlled appropriately in terms of local tolerances $\epsilon_i^{[j]}$. Assuming the residuals $r_i^{[j]}$ are bounded by $\|r_i^{[j]}\|_Y \le \epsilon_i^j$, the implicit SDC correction $\hat{\delta}^{[j]}$ for inexact system solves is obtained as

$$(I - (t_i - t_{i-1}) f'(y_i^{[j]}))\hat{\delta}_i^{[j]} = \hat{\delta}_{i-1}^{[j]} + \sum_{k=1}^{N} S_{ik} f(y_k^{[j]}) - (y_i^{[j]} - y_{i-1}^{[j]}) + r_i^{[j]}, \quad i = 1, \ldots, N, \qquad (2\text{-}8)$$

for $j = 0, \ldots, J-1$ and $i = 1, \ldots, N$ with $\hat{\delta}_0^{[j]} = 0$.

In both cases, the update $y^{[j]} \mapsto y^{[j+1]} := y^{[j]} + \hat{\delta}^{[j]}$ defines a parametrized fixed-point operator

$$\widehat{F} : Y^{N+1} \times \mathbb{R}_+^{N \times J+1} \times \mathbb{N} \to Y^{N+1}, \quad \widehat{F}(y^{[j]}; \epsilon, j) := y^{[j+1]},$$

with the exact limit case $F(y) := \widehat{F}(y; 0, 0)$.

For convergence analysis, we equip Y^{N+1} with a norm

$$\|y\| := \|[\|y_0\|_Y, \ldots, \|y_N\|_Y]\|_p \qquad (2\text{-}9)$$

in terms of the usual p-norm on $\mathbb{R}^{N+1}$ with $p \in [1, \infty]$ to be specified later. If F is Lipschitz continuous with constant $\rho < 1$, i.e.,

$$\|F(x) - F(y)\| \le \rho \|x - y\| \quad \text{for all } x, y \in Y^{N+1}$$

(which we will assume throughout the paper), Banach's fixed-point theorem yields q-linear convergence of the iteration to the unique collocation solution y_c independently of the initial iterate $y^{[0]}$. Note that the contraction property of F and hence the convergence of SDC depends on f, the collocation points t_i, the time step size T, and whether we use explicit or implicit SDC. For sufficiently small time steps, however, convergence is guaranteed if f is Lipschitz continuous.

Termination of the fixed-point iteration at iterate J can be based on either a fixed iteration count, resulting in a particular Runge–Kutta time-stepping scheme, or on an accuracy request of the form $\|y_c - y^{[J]}\| \le \mathrm{TOL}$. Given the contraction rate ρ, and assuming that $\|y_c - y^{[0]}\| > \mathrm{TOL}$, the number of exact iterations is then bounded by

$$J \le \left\lceil \frac{\log(\mathrm{TOL}/\|y_c - y^{[0]}\|)}{\log \rho} \right\rceil.$$

The choice of the initial iterate $y^{[0]}$ can not only have a significant impact on the number J of iterations needed to achieve the requested accuracy, but also on the properties of intermediate solutions. In particular for stiff problems, L-stability of intermediate solutions is obtained only if $y^{[0]}$ is computed by an L-stable basic scheme, e.g., implicit Euler, or special DIRK sweeps as proposed in [27]. For simplicity, however, we choose $y_i^{[0]} \equiv y_0$ in this paper.

Given the requirement of computing a final iterate $y^{[J]}$ satisfying the requested accuracy $\|y_c - y^{[J]}\| \le \mathrm{TOL}$, the immediate questions that arise are how to select the local tolerances $\epsilon_i^{[j]}$, and how many iterations to perform, in order to obtain the most efficient method. This question will be addressed in the following section.

3. A priori tolerance selection

Following the approach taken by Alfeld [1], an attractive choice of local tolerances $\epsilon_i^{[j]}$ and iteration count J is to minimize the overall computational effort $W(\epsilon, J)$ while bounding the final error $\|y^{[J]} - y_c\| \le \Phi(\epsilon, J)$:

$$\min_{J \in \mathbb{N},\, \epsilon \in \mathscr{E} \subset \mathbb{R}^{N \times J+1}} W(\epsilon, J) \quad \text{subject to } \Phi(\epsilon, J) \le \mathrm{TOL}. \tag{3-1}$$

Here, ϵ denotes the $N \times (J+1)$ matrix of local tolerances $\epsilon_i^{[j]}$, restricted to an admissible set $\mathscr{E}$. We will consider different admissible sets in Sections 3C–3E below.

For this abstract framework to be useful, a *work model* W and an *error model* Φ are needed. These two building blocks will be established in the following two subsections.

Remark 3.1. Both the error model and the work model will involve quantities that are not explicitly known in actual computation. For the error model, these parameters are in particular the SDC contraction factor ρ, Lipschitz constants, and the initial iteration error $\|y^{[0]} - y_c\|$. The work models derived below rely on typical or asymptotic computational effort, which may not very well describe the actual effort spent on a concrete problem. Therefore, the efficiency predicted by the solution of the optimization problem (3-1) may not be reached.

Moreover, even if the assumptions are satisfied and the parameters entering into the error model are known, the estimates are not sharp. The actual error will typically be smaller than its bound, which means that the local tolerances derived from the error bound will be smaller than necessary, and the computational effort in turn higher than need be. Therefore, solving (3-1) provides only theoretically optimal local tolerances.

Nevertheless, the approximation results developed in the subsequent sections provide not only theoretical insight, but can also guide algorithmic choices, if computable estimates for the required parameters are available. For example, the SDC contraction factor ρ can be assumed not to change quickly over the integration time, such that the convergence on the previous time step could provide sufficient information. Lipschitz constants can at least be bounded from below by inspecting the evaluated right-hand sides. Inserting such estimates into the optimization problem can yield reasonable heuristics for choosing local tolerances in actual computations. Of course, such heuristics will need to be complemented by a posteriori error estimators and heuristics for updating parameter values in case the estimated actual error is larger than predicted. This, however, is beyond the scope of the present work.

3A. ***Error model.*** The error model bounds the final iteration error by $\Phi(\epsilon, J)$ in terms of the local tolerances $\epsilon_i^{[j]}$ and the iteration count J. Focusing on SDC as a fixed-point iteration, we estimate Φ in terms of inexact fixed-point iterations [1; 22]. Below we consider the convergence of

$$y^{[j+1]} = \widehat{F}(y^{[j]}; \epsilon, j), \quad j = 0, \ldots, J-1, \tag{3-2}$$

to the fixed point y_c of F, and derive a bound on $\|y^{[J]} - y_c\|$ for given $y^{[0]}$, J, and ϵ.

First we establish estimates of how the errors bounded by $\epsilon_i^{[j]}$ are transported through the SDC sweep, and then address the complete iteration, both for explicit and implicit SDC schemes. For this we need some notation.

Definition 3.2. Let us assume there is a nonnegative function $L_f : \mathbb{R}_+ \to \mathbb{R}_+$ such that the right-hand side f satisfies the following Lipschitz-type conditions: for

explicit SDC

$$\|\delta + \tau(f(y+\delta) - f(y))\|_Y \le L_f(\tau)\|\delta\|_Y \quad \text{for all } \tau \in]0, T] \text{ and } \delta, y \in Y \tag{3-3}$$

and for implicit SDC

$$\|(I - \tau f'(y))^{-1}\|_Y \le L_f(\tau) \quad \text{for all } \tau \in]0, T] \text{ and } y \in Y. \tag{3-4}$$

Then we define the invertible lower-triangular matrix $L \in \mathbb{R}^{N\times N}$ as

$$L_{im} := \begin{cases} \prod_{l=m}^{i-1} L_f(t_{l+1} - t_l), & m \le i, \\ 0, & \text{otherwise}, \end{cases}$$

and introduce $\|e\|_L := \|Le\|_p$ for $e \in \mathbb{R}^N$ and $\|\kappa\|_L := \max_{\|e\|_L=1} \|\kappa e\|_L = \|L\kappa L^{-1}\|_p$ for $\kappa \in \mathbb{R}^{N\times N}$.

Note that the nonstandard Lipschitz condition (3-3) follows from the standard Lipschitz condition on f (because $\|f(y+\delta) - f(y)\|_Y \le L_*\|\delta\|_Y$ implies $L_f(\tau) \le 1 + \tau L_*$), but is weaker, in particular for slightly stiff systems. For example, for $f(y) = -y$ we obtain $L_f(\tau) = |1 - \tau| \ll 1 + \tau$ for $\tau \approx 1$. Nevertheless, the weaker condition (3-3) is sufficient for bounding the error transport through explicit SDC sweeps in the following theorem. Analogously, condition (3-4) describes the error transport through linearly implicit Euler sweeps in the implicit SDC method.

Explicit SDC. Now we derive error bounds, first for single SDC sweeps and then for the whole iteration.

Theorem 3.3. *Assume that the ODE's right-hand side satisfies the Lipschitz-like condition (3-3). Then, for* $\epsilon \in \mathbb{R}_+^{N\times J+1}$,

$$\|\widehat{F}(y; \epsilon, j) - F(y)\| \le \|\kappa(\epsilon^{[j]} + \epsilon^{[j+1]}) + |S|\epsilon^{[j]}\|_L \tag{3-5}$$

holds for the explicit SDC iteration with $\kappa \in \mathbb{R}^{N\times N}$, $\kappa_{mk} := \delta_{m-1,k}(t_m - t_{m-1})$, *where* $\delta_{m,k}$ *denotes the Kronecker delta.* $|S| \in \mathbb{R}^{N\times N}$ *denotes the entrywise absolute value of the integration matrix* S.

Proof. From (2-5) and (2-6) we obtain for the SDC corrections $\hat{\delta}_i$ the estimate

$$\begin{aligned} \|\widehat{F}(y; \epsilon, j)_i - F(y)_i\|_Y &= \|\hat{\delta}_i^{[j]} - \delta_i^{[j]}\|_Y \\ &\le \|\hat{\delta}_{i-1}^{[j]} - \delta_{i-1}^{[j]} + (t_i - t_{i-1})(f(y_{i-1} + \hat{\delta}_{i-1}^{[j]}) - f(y_{i-1} + \delta_{i-1}^{[j]}))\|_Y \\ &\quad + (t_i - t_{i-1})(\epsilon_{i-1}^{[j+1]} + \epsilon_{i-1}^{[j]}) + \sum_{k=1}^{N} |S_{ik}|\epsilon_k^{[j]} \\ &\le L_f(t_i - t_{i-1})\|\hat{\delta}_{i-1}^{[j]} - \delta_{i-1}^{[j]}\|_Y + (\kappa(\epsilon^{[j]} + \epsilon^{[j+1]}))_i + (|S|\epsilon^{[j]})_i \end{aligned}$$

with $\hat{\delta}_0^{[j]} - \delta_0^{[j]} = 0$. By induction we obtain the discrete Gronwall result

$$\begin{aligned}\|\hat{\delta}_i^{[j]} - \delta_i^{[j]}\|_Y &\le \sum_{m=1}^{i} \prod_{l=m}^{i-1} L_f(t_{l+1} - t_l)([\kappa(\epsilon^{[j]} + \epsilon^{[j+1]})]_m + (|S|\epsilon^{[j]})_m) \\ &= \sum_{m=1}^{i} L_{im}([\kappa(\epsilon^{[j]} + \epsilon^{[j+1]})]_m + (|S|\epsilon^{[j]})_m) \\ &= [L(\kappa(\epsilon^{[j]} + \epsilon^{[j+1]}) + |S|\epsilon^{[j]})]_i.\end{aligned}$$

Taking the norm over $i = 1, \dots, N$ yields the claim (3-5). □

With the error bound (3-5) for a single inexact SDC sweep at hand, we are in the position to bound the final time error.

Theorem 3.4. *Let $y^{[0]} \in Y^N$ be given, and let $y^{[j+1]}$ be defined by*

$$y^{[j+1]} = \widehat{F}(y^{[j]}, \epsilon, j), \quad j = 0, \dots, J-1,$$

for some $J \in \mathbb{N}$ and some local tolerance matrix $\epsilon \in \mathbb{R}^{N \times J+1}$. Then

$$\|y^{[J]} - y_c\| \le \alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \|\epsilon^{[j]}\|_L + \|\kappa\epsilon^{[J]}\|_L + \rho^J \|y^{[0]} - y_c\| =: \Phi(\epsilon, J) \quad (3\text{-}6)$$

holds with $\alpha = \|\kappa + |S|\|_L + \rho\|\kappa\|_L$ and κ and $|S|$ as defined in Theorem 3.3.

Proof. First we show the (slightly stronger) result

$$\|y^{[J]} - y_c\| \le \sum_{j=1}^{J} \rho^{J-j} \|\kappa(\epsilon^{[j-1]} + \epsilon^{[j]}) + |S|\epsilon^{[j-1]}\|_L + \rho^J \|y^{[0]} - y_c\| \quad (3\text{-}7)$$

by induction over J. The claim holds trivially for $J = 0$. Otherwise, we obtain

$$\begin{aligned}\|y^{[J]} - y_c\| &= \|\widehat{F}(y^{[J-1]}; \epsilon, j) - F(y_c)\| \\ &\le \|\widehat{F}(y^{[J-1]}; \epsilon, J-1) - F(y^{[J-1]})\| + \|F(y^{[J-1]}) - F(y_c)\| \\ &\le \|\kappa(\epsilon^{[J-1]} + \epsilon^{[J]}) + |S|\epsilon^{[J-1]}\|_L + \rho\|y^{[J-1]} - y_c\| \\ &\le \|\kappa(\epsilon^{[J-1]} + \epsilon^{[J]}) + |S|\epsilon^{[J-1]}\|_L \\ &\quad + \rho\left(\sum_{j=1}^{J-1} \rho^{J-1-j} \|\kappa(\epsilon^{[j-1]} + \epsilon^{[j]}) + |S|\epsilon^{[j-1]}\|_L + \rho^{J-1}\|y^{[0]} - y_c\|\right),\end{aligned}$$

which is just (3-7). Applying the triangle inequality and rearranging terms in the sum yields

$$\begin{aligned}\|y^{[J]} - y_c\| &\le \rho^{J-1}\|(\kappa + |S|)\epsilon^{[0]}\|_L + \sum_{j=1}^{J-1} \rho^{J-1-j}(\|(\kappa + |S|)\epsilon^{[j]}\|_L + \rho\|\kappa\epsilon^{[j]}\|_L) \\ &\quad + \|\kappa\epsilon^{[J]}\|_L + \rho^J\|y^{[0]} - y_c\| \\ &\le \sum_{j=0}^{J-1} \rho^{J-1-j} \underbrace{(\|\kappa + |S|\|_L + \rho\|\kappa\|_L)}_{=\alpha}\|\epsilon^j\|_L + \|\kappa\epsilon^{[J]}\|_L + \rho^J\|y^{[0]} - y_c\|\end{aligned}$$

and thus the claim (3-6). □

Note that $\epsilon^{[J]}$ enters the error bound $\Phi(\epsilon, J)$ given in (3-6) in a different way than $\epsilon_i^{[j]}$ for $j < J$. This is due to the fact that all right-hand sides evaluated enter twice into the computation (see (2-6)) except for the very last sweep evaluations, which enter only once. This turns out to be quantitatively important in Section 4.

Implicit SDC. Error bounds for inexact implicit SDC follow the same line of argument as for the explicit method, but are slightly simpler.

Theorem 3.5. *Assume that the ODE's right-hand side satisfies the Lipschitz-like condition (3-4). Then, for* $\epsilon \in \mathbb{R}_+^{N\times J+1}$,

$$\|\widehat{F}(y; \epsilon, j) - F(y)\| \le \|\sigma\epsilon^{[j]}\|_L \tag{3-8}$$

holds for the implicit SDC iteration with $\sigma \in \mathbb{R}^{N\times N}$, $\sigma_{kk} = L_f(t_k - t_{k-1})$.

Proof. From (2-7) and (2-8) we obtain for the SDC corrections $\hat{\delta}_i$ the estimate

$$\begin{aligned}\|\widehat{F}(y; \epsilon, j)_i - F(y)_i\|_Y &\le \|(I + (t_i - t_{i-1}))^{-1}(\hat{\delta}_{i-1}^{[j]} - \delta_{i-1}^{[j]} + r_i^{[j]})\|_Y \\ &\le L_f(t_i - t_{i-1})(\|\hat{\delta}_{i-1}^{[j]} - \delta_{i-1}^{[j]}\|_Y + \epsilon_i^{[j]}) \\ &= L_f(t_i - t_{i-1})\|\hat{\delta}_{i-1}^{[j]} - \delta_{i-1}^{[j]}\|_Y + (\sigma\epsilon^{[j]})_i\end{aligned}$$

with $\hat{\delta}_0^{[j]} - \delta_0^{[j]} = 0$. As before, induction provides the discrete Gronwall result

$$\|\widehat{F}(y; \epsilon, j)_i - F(y)_i\|_Y \le [L\sigma\epsilon^{[j]}]_i$$

and hence the claim (3-8). □

With the error bound (3-8) for a single implicit inexact SDC sweep at hand, we are in the position to bound the final time error.

Theorem 3.6. *Let* $y^{[0]} \in Y^N$ *be given, and let* $y^{[j+1]}$ *be defined by*

$$y^{[j+1]} = \widehat{F}(y^{[j]}; \epsilon, j), \quad j = 0, \dots, J-1,$$

for some $J \in \mathbb{N}$ and some local tolerance matrix $\epsilon \in \mathbb{R}^{N\times J+1}$. Then

$$\|y^{[J]} - y_c\| \le \alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \|\epsilon^{[j]}\|_L + \|\kappa\epsilon^{[J]}\|_L + \rho^J \|y^{[0]} - y_c\| =: \Phi(\epsilon, J) \quad (3\text{-}9)$$

holds with $\alpha = \|\sigma\|_L$ and $\kappa = 0$, where σ is defined in Theorem 3.5.

Proof. First we show the (slightly stronger) result

$$\|y^{[J]} - y_c\| \le \sum_{j=1}^{J} \rho^{J-j} \|\sigma\epsilon^{[j-1]}\|_L + \rho^J \|y^{[0]} - y_c\| \quad (3\text{-}10)$$

by induction over J. The claim holds trivially for $J = 0$. Otherwise, we obtain as in the proof of Theorem 3.4, now using (3-8),

$$\begin{aligned}\|y^{[J]} - y_c\| &\le \|\widehat{F}(y^{[J-1]}; \epsilon, J-1) - F(y^{[J-1]})\| + \|F(y^{[J-1]}) - F(y_c)\| \\ &\le \|\sigma\epsilon^{[J-1]}\|_L + \rho\|y^{[J-1]} - y_c\|,\end{aligned}$$

which implies (3-10). An index shift in j is all it takes to obtain the claim (3-9). □

Note that the error bounds $\Phi(\epsilon, J)$ as given in (3-6) and (3-9) for explicit and implicit SDC, respectively, have identical structure, and differ only in the values of the parameters α and κ. This allows a uniform analytical treatment of both explicit and implicit schemes in the following sections.

Remark 3.7. The choice of collocation nodes t_i affects the error bound (3-6) in three ways. First, the substep sizes $t_{i+1} - t_i$ enter into L_{ki} and hence into $\|\cdot\|_L$. Second, the integration matrix S enters into the factor α, and third, the contraction factor ρ depends on the collocation nodes in a nontrivial and up to now not well understood way.

The error model Φ as defined in (3-6) is an upper bound of the inexact SDC iteration for arbitrary errors bounded by the local tolerances $\epsilon_i^{[j]}$, and hence also an upper bound for the error $\rho^J\|y^{[0]} - y_c\|$ of the exact SDC iteration. Consequently, meeting the accuracy requirement $\Phi(\epsilon, J) \le \mathrm{TOL}$ implies $\rho^J\|y^{[0]} - y_c\| \le \mathrm{TOL}$ and

$$J \ge J_{\min} := \frac{\log \mathrm{TOL} - \log\|y^{[0]} - y_c\|}{\log \rho}.$$

3B. ***Work models.*** Let us assume that the computational effort of evaluating $f_i^{[j]}$ (in explicit SDC) or of solving for $\delta_i^{[j]}$ (in implicit SDC) is given in terms of the work $W_i^{[j]} : \mathbb{R}_+ \to \mathbb{R}_+$ as $W_i^{[j]}(\epsilon_i^{[j]})$. The total work to spend for J SDC iterations,

$$W(\epsilon) = \sum_{j=0}^{J} \sum_{i=1}^{N} W_i^{[j]}(\epsilon_i^{[j]}), \quad (3\text{-}11)$$

is just the sum of all efforts. Hence, common positive factors can be neglected, as they will not affect the minimizer of the optimization problem (3-1) at all. Note that, for optimizing just ϵ with fixed J, additive terms in the work model can also be neglected.

First we will discuss a few prototypical work models that cover common sources of controlled inaccuracy.

Finite element discretization. If the computation of basic steps within the SDC sweeps involves a PDE solution realized by adaptive finite elements, the discretization error can be expected to be proportional to $n^{-1/d}$, where n is the number of grid points and d is the spatial dimension. Assuming the work to be proportional to the number of grid points, we obtain

$$W_i^{[j]}(\epsilon_i^{[j]}) := \frac{1}{d}(\epsilon_i^{[j]})^{-d}. \tag{3-12}$$

The arbitrary factor d^{-1} has been introduced for notational convenience only.

Of course, the asymptotic behavior $W_i^{[j]} \to 0$ for $\epsilon_i^{[j]} \to \infty$ is not realistic, as there is a fixed amount $W_{\min}$ of work necessary on the coarse grid. Thus, the work model is valid only for $\epsilon_i^{[j]} \le \epsilon_{\max} = (d\,W_{\min})^{-1/d}$. We will address this in Section 3F.

Truncation errors. Let us assume the basic step computation involves the solution of a linear equation system by a linearly converging iterative solver. This is usually the case in implicit SDC methods applied to PDE problems. Starting the iterative solver at zero, the residual after m iterations may be assumed to be bounded by $\|r_i^{[j]}\|_Y \le \rho_{\mathrm{it}}^m R^{[j]}$, where $\rho_{\mathrm{it}} < 1$ is the contraction rate of the linear solver and $R^{[j]} \sim \|\delta_i^{[j]}\|_Y$ the size of the initial residual. The number of iterations necessary to reach the local tolerance $\|r_i^{[j]}\|_Y \le \epsilon_i^{[j]}$ is expected to be

$$m \ge \frac{\log \epsilon_i^{[j]} - \log R^{[j]}}{\log \rho_{\mathrm{it}}}.$$

If the outer SDC iteration converges linearly with unperturbed contraction factor ρ, an assumption that will be justified in (3-27), the initial residual is roughly $R^{[j]} \approx \rho^j \|y^{[0]} - y_c\|$, which leads to

$$W_i^{[j]}(\epsilon_i^{[j]}) := -\log \epsilon_i^{[j]} + \log\|y^{[0]} - y_c\| + j \log \rho. \tag{3-13}$$

Of course, a negative number of iterations cannot be realized, and therefore, this work model is limited to $\log \epsilon_i^{[j]} < \epsilon_{\max}^{[j]} = \rho^j \|y^{[0]} - y_c\|$.

Remark 3.8. Simplifying the work model (3-13) by ignoring the additive contribution $\log(c\mathrm{TOL}) + (j - J)\log\rho$ is sufficient for optimizing the local tolerances ϵ, but affects optimizing the number J of iterations and renders work ratios $W(\epsilon)/W(\hat{\epsilon})$ for comparing different local tolerance choices ϵ and $\hat{\epsilon}$ meaningless.

Stochastic sampling. In case the right-hand side contains a high-dimensional integral to be evaluated by Monte Carlo sampling, the accuracy can be expected to be proportional to the inverse square root of the number of samples. The work proportional to the number of samples is then

$$W_i^{[j]}(\epsilon_i^{[j]}) := \tfrac{1}{2}(\epsilon_i^{[j]})^{-2},$$

just a special case of (3-12). Of course, as the error bound of Monte Carlo sampling is not strict, the error model from the previous section gives no guarantee in this case.

The prototypical work models presented here share a common structure. Minimization of the total work is based on derivatives of the work with respect to the local tolerances. Here we see that all three models satisfy

$$(W_i^{[j]})'(\epsilon_i^{[j]}) = (\epsilon_i^{[j]})^{-(d+1)},$$

with $d = 0$ for truncation of iterative solvers, $d = 2$ for Monte Carlo sampling, and d giving the spatial dimension in adaptive linear finite element computations. This will allow us to treat all work models uniformly in the work optimization.

Moreover, the work models exhibit some qualitative properties, which we conjecture to be general properties of plausible work models.

Definition 3.9. A *work model* is a family of strictly convex, positive, and monotonically decreasing functions $W_i^{[j]} :]0, (\epsilon_{\max})_i^{[j]}] \to \mathbb{R}_+$ mapping requested tolerances to the associated computational effort. The functions $W_i^{[j]}$ exhibit the barrier property $W_i^{[j]}(\epsilon) \to \infty$ for $\epsilon \to 0$.

The properties of $W_i^{[j]}$ are inherited by the total work W of (3-11), which is strictly convex and monotone.

3C. ***Fixed local tolerance.*** To begin with, we consider heuristic choices of the admissible set $\mathscr{E}$ of local tolerances. The simplest possibility is to take the same value $\epsilon_i^{[j]} \equiv \epsilon_{\text{fix}}$ for all right-hand side evaluations. This corresponds to a fixed *absolute* solver tolerance in inexact implicit SDC.

In this case, the error bound (3-6) reduces to

$$\|y^{[J]} - y_c\| \le \epsilon_{\text{fix}}\left(\alpha\|\mathbf{1}\|_L \frac{1-\rho^J}{1-\rho} + \|\kappa\mathbf{1}\|_L\right) + \rho^J\|y^{[0]} - y_c\|,$$

where $\mathbf{1} \in \mathbb{R}^N$ with $\mathbf{1}_i = 1$. Consequently,

$$\epsilon_{\text{fix}} = \min\left(\epsilon_{\max}, \frac{\text{TOL} - \rho^J\|y^{[0]} - y_c\|}{\alpha\|\mathbf{1}\|_L(1-\rho^J)/(1-\rho) + \|\kappa\mathbf{1}\|_L}\right) \tag{3-14}$$

provides the largest admissible choice, and hence the one that incurs the least computational effort, for given J. With $\epsilon_{\text{fix}}(J)$ fixed, what remains is to choose the

number J of SDC sweeps such that the overall work is minimized. To this extent, we consider the slightly more restrictive but easier to analyze variant

$$\epsilon_{\text{fix}} = \min\left(\epsilon_{\max}, \frac{\text{TOL} - \rho^J \|y^{[0]} - y_c\|}{\alpha\|\mathbf{1}\|_L/(1-\rho) + \|\kappa\mathbf{1}\|_L}\right).$$

For the general work model (3-12), the total work is just $W = N(J+1)\epsilon_{\text{fix}}(J)^{-d}/d$. Assuming $\epsilon_{\text{fix}} < \epsilon_{\max}$ and eliminating constant factors, we need to minimize $W(J) \sim (J+1)/(\text{TOL} - \rho^J\|y^{[0]} - y_c\|)^d$. A simple analysis reveals that $W(J)$ is quasiconvex, such that there is exactly one minimizer in $]J_{\min}, \infty[$; see the Appendix. Unfortunately, no closed expression seems to exist, but a numerical computation is straightforward. Due to the quasiconvexity, the optimal $J \in \mathbb{N}$ is one of the neighboring integer values.

The local tolerance is bounded by $\epsilon_{\text{fix}} \le c\text{TOL}$ for some generic constant c independent of J and TOL. Consequently, the total work is at least

$$W \ge c(J_{\min}+1)\text{TOL}^{-d} = c\left(\frac{\log(\text{TOL}/\|y^{[0]} - y_c\|)}{\log\rho} + 1\right)\text{TOL}^{-d}. \tag{3-15}$$

Apparently, a complexity of $\mathcal{O}(\text{TOL}^{-d})$ is unavoidable, as this is already required for a single right-hand side evaluation to the requested accuracy. The logarithmic factor in (3-15), however, appears to be suboptimal. As this corresponds to the number J of SDC sweeps, which, depending on the concrete problem, can easily exceed a factor of ten, the suboptimality may induce a significant inefficiency in actual computation. We will address this shortcoming in the following Sections 3D and 3E and investigate it numerically in Section 4.

For completeness we note that in the less interesting case $\epsilon_{\text{fix}} = \epsilon_{\max}$, J is determined by minimizing $W = N(J+1)\epsilon_{\max}^{-d}/d$ subject to

$$\text{TOL} \ge \epsilon_{\max}(\alpha\|\mathbf{1}\|_L/(1-\rho) + \|\kappa\mathbf{1}\|_L) + \rho^J\|y^{[0]} - y_c\| \ge \Phi(\epsilon_{\max}, J) \ge \|y^{[J]} - y_c\|,$$

i.e.,

$$J \ge (\log\rho)^{-1}\log\frac{\text{TOL} - \epsilon_{\max}(\alpha\|\mathbf{1}\|_L/(1-\rho) + \|\kappa\mathbf{1}\|_L)}{\|y^{[J]} - y_c\|}.$$

3D. ***Geometrically decreasing local tolerances.*** The next step is to exploit the fact that, due to the linear convergence of the SDC iteration, larger evaluation errors are acceptable in the early iterations, and to make the heuristic choice

$$(\epsilon_{\text{geo}})_i^{[j]} = \min(\epsilon_{\max}, \beta\rho^{\gamma j}) \quad \text{for some } \beta, \gamma > 0. \tag{3-16}$$

This has been considered in [5] for $\gamma = 1$ as an "adaptive strategy" and is closely related to evaluating implicit Euler steps up to a fixed *relative* precision in implicit SDC methods, as suggested in [16] or realized in [24] by a fixed number of multigrid V-cycles.

We will assume that γ is given and optimize β as we have done before with ϵ_{fix}. Ignoring the impact of $\epsilon_{\max}$, (3-6) results in the slightly stronger accuracy requirement

$$\|y^{[J]} - y_c\| \leq \beta\left(\alpha\|\mathbf{1}\|_L \sum_{j=0}^{J-1} \rho^{J-1-j+\gamma j} + \rho^{\gamma J}\|\kappa\mathbf{1}\|_L\right) + \rho^J\|y^{[0]} - y_c\| \overset{!}{\leq} \text{TOL}.$$

Note that this implies a convergence rate of $\|y^{[J]} - y_c\| = \mathbb{O}(\rho^{\min(1,\gamma)J})$. For $\gamma \neq 1$ (there is a continuous extension to $\gamma = 1$, though) we obtain

$$\beta \leq \frac{\text{TOL} - \rho^J\|y^{[0]} - y_c\|}{\rho^{\gamma(J-1)}(\alpha\|\mathbf{1}\|_L(1-\rho^{(1-\gamma)J})/(1-\rho^{1-\gamma}) + \rho^\gamma\|\kappa\mathbf{1}\|_L)}. \tag{3-17}$$

The total work $W(\epsilon)$ is monotonically decreasing in β due to (3-16) and Definition 3.9, such that the work-minimization problem (3-1) is solved by equality in (3-17), and we obtain

$$(\epsilon_{\text{geo}})_i^{[j]} = \min\left(\epsilon_{\max}, \frac{\text{TOL} - \rho^J\|y^{[0]} - y_c\|}{\rho^{\gamma(J-1)}(\alpha\|\mathbf{1}\|_L(1-\rho^{(1-\gamma)J})/(1-\rho^{1-\gamma}) + \rho^\gamma\|\kappa\mathbf{1}\|_L)}\rho^{\gamma j}\right). \tag{3-18}$$

Of course, $\epsilon_{\text{fix}} = \lim_{\gamma\to 0}\epsilon_{\text{geo}}$ is recovered in the limit.

Optimizing the iteration count J for the generic work model (3-12) with $d > 0$, we minimize

$$W = N\beta^{-d}\sum_{j=0}^{J}\rho^{-d\gamma j}/d \sim \beta^{-d}\frac{1-\rho^{-d\gamma(J+1)}}{1-\rho^{-d\gamma}}.$$

We distinguish between $\gamma < 1$ and $\gamma > 1$. In the first case, we obtain

$$\frac{1-\rho^{(1-\gamma)J}}{1-\rho^{1-\gamma}} \leq (1-\rho^{1-\gamma})^{-1} \quad \text{and thus} \quad \beta \geq \frac{\text{TOL} - \rho^J\|y^{[0]} - y_c\|}{\rho^{\gamma(J-1)}c}$$

for some $c > 0$ independent of J. Neglecting constant factors independent of J yields the upper bound

$$W \lesssim \left(\frac{\text{TOL} - \rho^J\|y^{[0]} - y_c\|}{\rho^{\gamma(J-1)}}\right)^{-d}\rho^{-d\gamma(J+1)}$$

decreasing monotonically with J towards $\lim_{J\to\infty} W \lesssim \text{TOL}^{-d}$. Compared to (3-15), the complexity to reach the requested tolerance is improved, independently of γ, from $\mathbb{O}(\text{TOL}^{-d}|\log\text{TOL}|)$ to $\mathbb{O}(\text{TOL}^{-d})$. In the next section we will see that this complexity is indeed optimal, but the constants can be improved further by considering a larger admissible set $\mathfrak{E}$.

In the second case $\gamma > 1$, we obtain the upper bound

$$W \lesssim \left(\frac{\mathrm{TOL} - \rho^J \|y^{[0]} - y_c\|}{c\rho^J + \rho^{\gamma(J-1)}} \right)^{-d} \rho^{-d\gamma(J+1)} \sim \left(\frac{c\rho^{(1-\gamma)J} + b}{\mathrm{TOL} - \rho^J \|y^{[0]} - y_c\|} \right)^{d}$$

for some generic constants b and c independent of J and TOL. Inserting $J \geq \log(\mathrm{TOL}/\|y^{[0]} - y_c\|)/\log\rho$ reveals a complexity of $\mathcal{O}(\mathrm{TOL}^{-\gamma d})$, indeed worse than the fixed choice $\epsilon_i^{[j]} \equiv \epsilon_{\mathrm{fix}}$ before. As a certain number of SDC iterations have to be performed with sufficient accuracy, increasing the accuracy too quickly is a waste of resources. Fortunately, a fixed relative accuracy will always lead to $\gamma \leq 1$.

3E. ***Variable local tolerances.*** Finally, let us consider the most general admissible set $\mathcal{E} = \{\epsilon \in \mathbb{R}_+^{N\times J+1} \mid \epsilon_i^{[j]} \leq \epsilon_{\max}\}$ in greater detail than we have treated the heuristic choices. Again, we will proceed in two steps, first assuming J to be given, optimizing only the local tolerances ϵ, and considering the integer variable J of the mixed integer program later on.

We obtain the nonlinear program

$$\min_{\epsilon\in\mathbb{R}_+^{N\times J+1}} W(\epsilon) \quad \text{subject to } \Phi(\epsilon, J) \leq \mathrm{TOL},\ \epsilon \leq \epsilon_{\max}. \tag{3-19}$$

From the properties of W and Φ, we immediately obtain the following result.

Theorem 3.10. *If $\rho^J \|y^{[0]} - y_c\| < \mathrm{TOL}$, i.e., the exact SDC iteration converges to the given tolerance, the optimization problem (3-19) has a unique solution $\epsilon(y^{[0]}, J)$. In the generic case $\epsilon_i^{[j]} < (\epsilon_{\max})_i^{[j]}$ for some i and j, i.e., if not all of the local tolerance constraints are active, the accuracy constraint is active, i.e., $\Phi(\epsilon(y^{[0]}, J), J) = \mathrm{TOL}$.*

Proof. From (3-6) it is apparent that sufficiently small values $\epsilon_i^{[j]} > 0$ lead to

$$\alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \|\epsilon^{[j]}\|_L + \|\kappa\epsilon^{[J]}\|_L \leq \mathrm{TOL} - \rho^J \|y^{[0]} - y_c\|,$$

such that the admissible set is nonempty. Strict convexity of W and convexity of Φ imply uniqueness of a solution. Strict convexity and monotonicity of W imply its strict monotonicity, and hence, the constraint must be active unless all local tolerances are actively bound by $\epsilon \leq \epsilon_{\max}$. □

The activity of the accuracy constraint in the generic case means that, as expected, no effort is wasted on reducing the error below the requested tolerance.

We may reasonably expect the local tolerances to decrease monotonically. This is indeed true in general, as the following result shows.

Theorem 3.11. *Assume that $\rho \in (0,1)$, $J \in \mathbb{N}$, and $\mathrm{TOL} \in \mathbb{R}_+$ are given constants. Let the local tolerance matrix ϵ be the minimizer of (3-19). Then $\|\epsilon^{[j]}\|_L \le \|\epsilon^{[j-1]}\|_L$ holds for all $j = 1, \dots, J-1$.*

For the norm exponent $p = 1$ in (2-9), componentwise monotonicity holds as well, i.e., $\epsilon_i^{[j]} \le \epsilon_i^{[j-1]}$ holds for all i and $j < J$.

Proof. Let $\tilde{\epsilon}$ be an admissible point for (3-19) with $\|\tilde{\epsilon}^{[k_1]}\|_L < \|\tilde{\epsilon}^{[k_2]}\|_L$ for some $1 \le k_1 < k_2 < J$. Then we consider ϵ with $\epsilon^{[j]} = \tilde{\epsilon}^{[j]}$ except for $\epsilon^{[k_2]} = \tilde{\epsilon}^{[k_1]}$ and $\epsilon^{[k_1]} = \tilde{\epsilon}^{[k_2]}$. Obviously, $W(\epsilon) = W(\tilde{\epsilon})$.

The error bound (3-6), however, is reduced:

$$\begin{aligned}
&\Phi(\tilde{\epsilon}, J) - \Phi(\epsilon, J) \\
&\qquad = \alpha(\rho^{J-1-k_1}(\|\tilde{\epsilon}^{[k_1]}\|_L - \|\epsilon^{[k_1]}\|_L) + \rho^{J-1-k_2}(\|\tilde{\epsilon}^{[k_2]}\|_L - \|\epsilon^{[k_2]}\|_L)) \\
&\qquad = \alpha(\rho^{J-1-k_1}(\|\tilde{\epsilon}^{[k_1]}\|_L - \|\tilde{\epsilon}^{[k_2]}\|_L) + \rho^{J-1-k_2}(\|\tilde{\epsilon}^{[k_2]}\|_L - \|\tilde{\epsilon}^{[k_1]}\|_L)) \\
&\qquad = \alpha(\rho^{J-1-k_1} - \rho^{J-1-k_2})(\|\tilde{\epsilon}^{[k_1]}\|_L - \|\tilde{\epsilon}^{[k_2]}\|_L) > 0,
\end{aligned}$$

as $\alpha > 0$ and the other two factors on the last line are negative. Since $\Phi(\epsilon, J) < \Phi(\tilde{\epsilon}, J) \le \mathrm{TOL}$, ϵ is feasible. The constraint, however, is inactive, such that ϵ cannot be the minimizer $\epsilon(y^{[0]}, J)$. We conclude that

$$W(\epsilon(y^{[0]}, J)) < W(\epsilon) = W(\tilde{\epsilon}),$$

such that $\tilde{\epsilon} \ne \epsilon(y^{[0]}, J)$. The same line of argument holds for $p = 1$ and componentwise monotonicity, where however ϵ is constructed such that only $\epsilon_i^{[k_1]}$ and $\epsilon_i^{[k_2]}$ are swapped. □

Below the necessary and, due to convexity, also sufficient conditions for the solution of the constrained optimization problem are derived for $p < \infty$.

Theorem 3.12. *Let the norm exponent p be finite. Assume $\rho^J \|y^{[0]} - y_c\| < \mathrm{TOL}$ and $W_i^{[j]} \in C^1(0, \infty)$. Then $\epsilon \in \mathbb{R}_+^{N \times J+1}$ solves (3-19), if and only if there exist multipliers $\mu \in \mathbb{R}$ and $\eta \in \mathbb{R}^{N \times J+1}$ such that*

$$\begin{aligned}
(W_i^{[j]})'(\epsilon_i^{[j]}) + \mu\alpha\rho^{J-1-j}\|\epsilon^{[j]}\|_L^{1-p} \sum_{k=1}^{N} (L\epsilon^{[j]})_k^{p-1} L_{ki} + \eta_i^{[j]} = 0, \\
j = 0, 1, \dots, J-1, \\
(W_i^{[J]})'(\epsilon_i^{[J]}) + \mu\|\kappa\epsilon^{[J]}\|_L^{1-p} \sum_{k=1}^{N} (L\kappa\epsilon^{[J]})^{p-1} (L\kappa)_{ki} + \eta_i^{[J]} = 0, \\
(\mathrm{TOL} - \Phi(\epsilon, J))\mu = 0, \quad \mu \ge 0, \\
(\epsilon_{\max} - \epsilon) : \eta = 0, \quad \eta \ge 0.
\end{aligned} \tag{3-20}$$

Here, $\epsilon : \eta$ denotes contraction or Frobenius product, and we use the convention $0^0 := 0$ (for $\kappa = 0$ and $p = 1$ this expression can formally arise).

Proof. The necessary and also sufficient condition for optimality of ϵ is the stationarity of the Lagrangian

$$L(\epsilon, \mu, \eta) = W(\epsilon, J) + \mu(\Phi(\epsilon, J) - \mathrm{TOL}) + \eta : (\epsilon_{\max} - \epsilon)$$

for some multiplier $\mu \in \mathbb{R}$ and $\eta \in \mathbb{R}^{N\times J+1}$; see, e.g., [21]. According to the (structurally identical) error bounds (3-6) and (3-9), and the total work (3-11), its partial derivatives are just the expressions in (3-20). □

At this point, the unique minimizer $\epsilon(y^{[0]}, J)$ of the convex program (3-19) can in principle be computed numerically. For an exponent $p = 1$ in the norm definition (2-9), however, explicit analytical expressions can easily be derived due to (3-6) reducing to

$$\begin{aligned}\Phi(\epsilon, J) &= \alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \sum_{k=1}^{N}\sum_{i=1}^{N} L_{ki}\epsilon_i^{[j]} + \sum_{k=1}^{N}\sum_{i=1}^{N}(L\kappa)_{ki}\epsilon_i^{[J]} + \rho^J \|y^{[0]} - y_c\| \\ &= q : \epsilon + \rho^J \|y^{[0]} - y_c\| \end{aligned} \tag{3-21}$$

with

$$q_i^{[j]} = \begin{cases} \alpha\rho^{J-1-j} \sum_{k=1}^{N} L_{ki}, & j < J, \\ \sum_{k=1}^{N}(L\kappa)_{ki}, & j = J. \end{cases} \tag{3-22}$$

Then, the first-order necessary conditions (3-20) assume the particularly simple form

$$(W_i^{[j]})'(\epsilon_i^{[j]}) + \mu q_i^{[j]} + \eta_i^{[j]} = 0. \tag{3-23}$$

Below we will derive the analytical structure of solutions for $p = 1$ and different work models, which also sheds some more light on the structure of the solution as well as on the achieved efficiency. The following theorem applies to all work models from Section 3B, with $d = 0$ for iterative solvers and $d = 2$ for stochastic sampling.

Theorem 3.13. *Let $p = 1$ and $(W_i^{[j]})'(\epsilon_i^{[j]}) = -(\epsilon_i^{[j]})^{-(d+1)}$. Then there is $\mu > 0$ such that the solution $\epsilon = \epsilon(y^{[0]}, J)$ of (3-19) is given by*

$$\epsilon_i^{[j]} = \begin{cases} (\epsilon_{\max})_i^{[j]}, & q_i^{[j]} = 0, \\ \min((\epsilon_{\max})_i^{[j]}, (\mu q_i^{[j]})^{-1/(d+1)}), & \textit{otherwise}, \end{cases} \tag{3-24}$$

with $q_i^{[j]}$ given in (3-22). Locally unconstrained tolerances $\epsilon_i^{[j]} < (\epsilon_{\max})_i^{[j]}$ decrease linearly up to $j = J - 1$:

$$\epsilon_i^{[j]} \sim \rho^{j/(d+1)}. \tag{3-25}$$

Proof. From the necessary condition (3-23) we obtain a multiplier $\hat{\mu} \geq 0$. If $\hat{\mu} = 0$, then $\eta_i^{[j]} > 0$ for all i and j due to $(W_i^{[j]})' < 0$, which implies $\epsilon = \epsilon_{\max}$ via complementarity in (3-20). Choosing $\mu > 0$ sufficiently small verifies the claim (3-24).

Otherwise we choose $\mu = \hat{\mu} > 0$ and obtain

$$\epsilon_i^{[j]} = (\mu q_i^{[j]} + \eta_i^{[j]})^{-1/(d+1)} \tag{3-26}$$

from (3-23). In case $\epsilon_i^{[j]} < (\epsilon_{\max})_i^{[j]}$, $\eta_i^{[j]} = 0$ holds due to complementarity in (3-20), such that the claim (3-24) is satisfied. For $j < J$, the definition (3-22) of $q_i^{[j]}$ implies

$$\epsilon_i^{[j]} = (\mu\alpha\rho^{J-1-j} \sum_{k=1}^{N} L_{ki})^{-1/(d+1)} \sim \rho^{j/(d+1)}$$

and hence the geometric decrease (3-25).

In case $\epsilon_i^{[j]} = (\epsilon_{\max})_i^{[j]}$ and $q_i^{[j]} \neq 0$, (3-26) implies

$$(\mu q_i^{[j]})^{-1} = (((\epsilon_{\max})_i^{[j]})^{-(d+1)} - \eta)^{-1} \geq ((\epsilon_{\max})_i^{[j]})^{d+1}$$

and hence the claim (3-24). □

The result (3-25) reveals that the heuristic of geometrically decreasing local tolerances is indeed of optimal complexity, at least for $\gamma < 1$, and now theoretically justified. Beyond that, an optimal value of $\gamma = (d+1)^{-1}$ and different accuracies for the collocation points are provided. We will see in Section 4 that the last issue can have a nonnegligible impact on the computational effort. Moreover, the result (3-25) shows that the contraction rate of optimal inexact SDC iterations depends on the work model: ρ for the truncation of linearly convergent iterations and $\rho^{1/(d+1)}$ for linear finite element solutions. The latter convergence is actually slower than the exact SDC iteration. This is a consequence of the different work required to reduce the error: while a reduction of the SDC iteration error is relatively cheap, reducing finite element discretization errors is rather expensive. An optimal tolerance selection therefore assigns a larger portion of the total error to the discretization and has to ensure that the SDC iteration error is by a certain factor smaller than the discretization error.

As expected, the geometric decrease (3-25) translates directly into linear convergence of the inexact SDC iteration:

Corollary 3.14. *If $\epsilon < \epsilon_{\max}$ holds, there is some c independent of j (though it depends on J) such that*

$$\|y^{[j]} - y_c\| \leq c\rho^{j/(d+1)}. \tag{3-27}$$

Proof. The result (3-25) yields

$$\begin{aligned}\|y^{[j]} - y_c\| &\le \alpha \sum_{k=0}^{j-1} \rho^{j-1-k} \|\epsilon^{[k]}\|_L + \|\kappa\epsilon^{[j]}\|_L + \rho^j \|y^{[0]} - y_c\| \\ &\le c\left(\sum_{k=0}^{j-1} \rho^{j-1-k}\rho^{k/(d+1)} + \rho^{j/(d+1)}\right) + \rho^j \|y^{[0]} - y_c\| \\ &\le c\rho^{j/(d+1)} \end{aligned} \tag{3-28}$$

and hence the claim. □

For $d = 0$, this linear convergence justifies the contraction rate assumed in defining the work model (3-13) for iterative solvers.

Let us state two more observations. First, it pays off to treat the final local tolerances $\epsilon_i^{[J]}$ separately in Theorem 3.4: now $\epsilon_i^{[J]} > \epsilon_i^{[J-1]}$ holds instead of $\epsilon_i^{[J]} = \rho\epsilon_i^{[J-1]}$. Thus, the effort for the otherwise greatest expense, due to having the most accurate right-hand side evaluations, is reduced, as illustrated in Figure 1. Similarly, for implicit SDC schemes with $\kappa = 0$ defined in Theorem 3.6, $q_i^{[J]} = 0$ holds, which implies $\epsilon_i^{[J]} = (\epsilon_{\max})_i^{[J]}$.

Second, (3-24) is monotone in μ, such that the actual value of μ and in turn ϵ is easily computed numerically by solving $\Phi(\epsilon, J) = \mathrm{TOL}$. In case $\epsilon_{\max}$ is sufficiently large such that $\epsilon < \epsilon_{\max}$ holds, combining (3-24), (3-22), and (3-21) yields an explicit expression

$$\epsilon_i^{[j]} = \frac{\mathrm{TOL} - \rho^J \|y^{[0]} - y_c\|}{\sum_{j=0}^{J}\sum_{i=0}^{N}(q_i^{[j]})^{d/(d+1)}} (q_i^{[j]})^{-1/(d+1)}. \tag{3-29}$$

3F. ***Iteration count optimization.*** As in the case of uniform local tolerances, the number J of inexact SDC iterations has to be selected in order to minimize the total work. For the generic work model (3-12), and stripping it of common factors and additive terms, we obtain with Theorem 3.13

$$W(J) = \sum_{j=0}^{J}\sum_{i=1}^{N}(\epsilon_i^{[j]})^{-d} = \sum_{j=0}^{J}\sum_{i=1}^{N}(\mu q_i^{[j]})^{d/(d+1)}$$

as long as $\epsilon_i^{[j]} < (\epsilon_{\max})_i^{[j]}$ for all i and j. Inserting the definition (3-22) of $q_i^{[j]}$ and neglecting constant factors independent of J and N yields

$$\begin{aligned} W \le \sum_{j=0}^{J}\sum_{i=1}^{N}\left(\mu\alpha\rho^{J-1-j}\sum_{k=1}^{N} L_{ki}\right)^{d/(d+1)} &\sim N\mu^{d/(d+1)} \sum_{j=0}^{J} \rho^{(d/(d+1))(J-1-j)} \\ &\sim N\mu^{d/(d+1)} \frac{1-\rho^{d(J+1)/(d+1)}}{1-\rho^{d/(d+1)}} \end{aligned} \tag{3-30}$$

as long as $\max_i(t_i - t_{i-1}) \le c/N$ for some constant c.

The multiplier μ is obtained from $\Phi(\epsilon, J) = \mathrm{TOL}$ with $\epsilon_i^{[j]} = (\mu q_i^{[j]})^{-1/(d+1)}$. We obtain

$$\begin{aligned}
\mathrm{TOL} &= \alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \|\epsilon^{[j]}\|_L + \|\kappa \epsilon^{[J]}\|_L + \rho^J \|y^0 - y_c\| \\
&= \mu^{-1/(d+1)} \left(\alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \|(q^{[j]})^{-1/(d+1)}\|_L + \|\kappa (q^{[J]})^{-1/(d+1)}\|_L \right) + \rho^J \|y^0 - y_c\| \\
&= \mu^{-1/(d+1)} \left(a \sum_{j=0}^{J-1} \rho^{(d/(d+1))(J-1-j)} + b \right) + \rho^J \|y^0 - y_c\| \\
&= \mu^{-1/(d+1)} \left(a \frac{1 - \rho^{dJ/(d+1)}}{1 - \rho^{d/(d+1)}} + b \right) + \rho^J \|y^0 - y_c\|
\end{aligned}$$

with constants $a = \alpha \|(\alpha \sum_{k=1}^N L_{ki})^{-1/(d+1)}\|_L$ and $b = \|\kappa (q^{[J]})^{-1/(d+1)}\|_L$ independent of J. Consequently,

$$\mu^{d/(d+1)} = \left(\frac{a(1 - \rho^{dJ/(d+1)})/(1 - \rho^{d/(d+1)}) + b}{\mathrm{TOL} - \rho^J \|y^0 - y_c\|} \right)^d$$

holds. Entering this into the work bound (3-30) yields

$$W \lesssim N \left(\frac{a(1 - \rho^{dJ/(d+1)})/(1 - \rho^{d/(d+1)}) + b}{\mathrm{TOL} - \rho^J \|y^0 - y_c\|} \right)^d \frac{1 - \rho^{d(J+1)/(d+1)}}{1 - \rho^{d/(d+1)}}.$$

Replacing $1 - \rho^{dJ/(d+1)}$ by 1 and neglecting constant factors independent of J and TOL provides the upper bound

$$W \lesssim N(\mathrm{TOL} - \rho^J \|y^0 - y_c\|)^{-d}. \tag{3-31}$$

The upper bound (3-31) is monotonically decreasing and suggests choosing J as large as possible. In the limit $J \to \infty$, the total work is bounded by

$$W \lesssim N \mathrm{TOL}^{-d}. \tag{3-32}$$

Compared to the work bound (3-15) for uniform local tolerances, the logarithmic factor $\log \mathrm{TOL}$ is missing, which yields the optimal complexity of evaluating N steps of the basic Euler scheme up to the requested tolerance.

Remark 3.15. The result (3-32) suggests that inexact explicit SDC methods might be able to reach or even surpass the efficiency of standard explicit Runge–Kutta methods.

However, the practical bound $\epsilon \le \epsilon_{\max}$ induces a lower bound $W_i^j \ge W_{\min}$ on the work per iteration, and hence, the total work $W(J)$ grows linearly with J

for $J \to \infty$. This contradicts the asymptotic work bound (3-32), which means that the assumption $\epsilon < \epsilon_{\max}$ used to derive (3-32) can only hold up to some finite iteration count. As closed expressions for a global minimizer of $W(J)$ when taking the local tolerance constraint $\epsilon \leq \epsilon_{\max}$ into account are hard to get, a heuristic selection of J appears to be most promising in practice. The convexity of (3-31) and linear growth of W for large J suggest that we may select J as

$$J = \min\{j \in \mathbb{N} \mid W(j+1) > W(j)\}.$$

4. Numerical examples

Here we will illustrate and compare the effectiveness of the inexact SDC strategies worked out above. First, the properties of the strategies will be explored using a simple academic test problem. Then, inexactness due to iterative solvers and Monte Carlo sampling are considered with the heat equation and a molecular dynamics example, respectively.

4A. *An illustrative example.*

Problem setup. As a particularly simple example that allows a detailed investigation of the theoretical predictions, we consider the harmonic oscillator

$$\dot{u} = v,$$
$$\dot{v} = -u,$$

with initial values $u_0 = 0$ and $v_0 = 1$, on the time interval $[0, \pi]$ subdivided into n equidistant time steps. The Lipschitz constant of the right-hand side is $L_* = 1$, and we estimate $L_f(\tau) = 1 + \tau$ using the triangle inequality. We use N Gauss–Legendre collocation points in each of the n time steps. The collocation error e_c at final time π can easily be obtained by comparing the result with the exact solution $u(t) = \sin(t)$, $v(t) = \cos(t)$. The contraction rate ρ of the exact SDC iteration is estimated numerically, and is virtually independent of the actual time t.

Exact right-hand side evaluation is of course straightforward, such that artificial inexactness and associated computational work are quite arbitrary. Here we use normally distributed random additive errors and the generic work model (3-12) with parameter d unless otherwise stated.

Aiming at a final time error comparable to the collocation error, we choose a tolerance $\mathrm{TOL} = e_c/\sqrt{n}$ for each time step, based on the assumption that the random errors of each time step simply add up, and yield a standard deviation of the final result of $\sqrt{n}\mathrm{TOL}$. With this setting, the quantities entering into the computation of the local tolerances ϵ are the same for all time steps. Unless otherwise stated $N = 3$ is used throughout, such that the collocation scheme is of order 6.

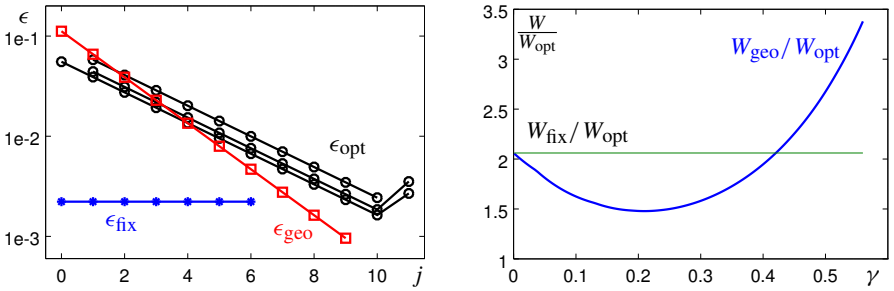

Figure 1. Left: exemplary local tolerances versus iteration number j for the different admissible design sets: fixed (ϵ_{fix}, stars), geometrically decreasing (ϵ_{geo}, squares, $\gamma = 0.5$), and variable (ϵ_{opt}, circles). Here $n = 2$ steps have been used to define the problem data $\rho = 0.35$, TOL $= 0.05$. Right: relative work $W_{\text{geo}}/W_{\text{opt}}$ for geometrically decreasing local tolerances versus the exponent γ. For larger γ, the work grows exponentially. The horizontal line denotes the relative work $W_{\text{fix}}/W_{\text{opt}}$ for fixed local tolerances.

Theoretical predictions. Let us first investigate the structure of local tolerances and the predicted efficiency gain in different situations.

To begin, we fix the iteration count J and time step T and compare local tolerances ϵ_{fix}, ϵ_{geo}, and ϵ_{opt} as given by the three considered strategies in (3-14), (3-18), and (3-24), respectively. Exemplary values for TOL $= 0.05$, $J = 11$, $\epsilon_{\max} = \infty$, $n = 2$, and estimated $\|y^{[0]} - y_c\| \approx 2.4$ are shown in Figure 1, left, versus the iteration number j. For the geometrically decreasing local tolerances, an exponent $\gamma = \frac{1}{2}$ has been chosen arbitrarily, but less than one due to the worse computational complexity for $\gamma > 1$; see Section 3D. For optimal variable tolerances, ϵ_{opt} has been obtained via (3-29). Clearly visible is the slow geometric decrease of the optimal variable local tolerances $\epsilon_{\text{opt}}^{[j]}$ with an order $\rho^{j/3}$, even slower than the explicitly chosen geometrical decrease $\rho^{\gamma j}$ with $\gamma = \frac{1}{2}$. The relative predicted work is $W_{\text{fix}}/W_{\text{opt}} = 2.06$ and $W_{\text{geo}}/W_{\text{opt}} = 2.67$. Somewhat surprisingly, exploiting the linear convergence of the SDC iteration does not necessarily pay off compared to a fixed accuracy, depending on the chosen parameter γ. The variable local tolerances approach achieves its low work by (i) choosing the appropriate decrease rate $\gamma = 1/(d+1)$, (ii) allowing for larger errors in later collocation points with less global impact, and (iii) imposing less restrictive requirements on the final sweep according to the definition (3-22) of $q_i^{[j]}$. The latter two aspects make up a reduction of work by a factor of 1.67 compared to the geometrically decreasing local tolerances with $\gamma = 1/(d+1)$. The relative work for different values of γ is shown in Figure 1, right, where the predicted total work induced by geometrically decreasing tolerances is plotted over the exponent γ. The optimum with a relative work of 1.48 is attained around $\gamma = 0.21$, even less than $1/(d+1)$. This can be attributed to avoiding high costs in the very last sweep, where high accuracy is actually not necessary, while ensuring sufficient accuracy in the next to last sweep.

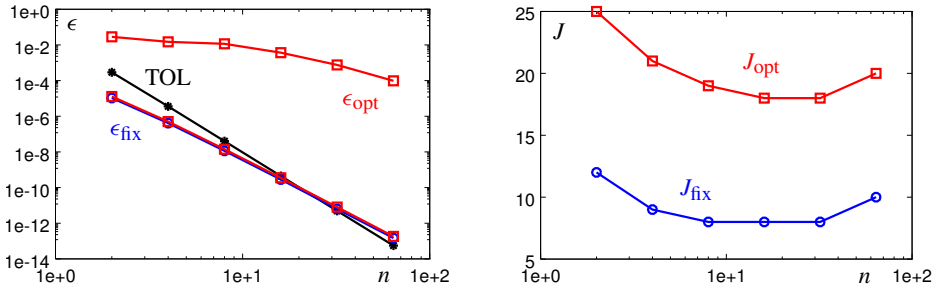

Figure 2. Left: local tolerances ϵ for the inexact SDC iterations versus number n of time steps. For optimal variable local tolerances (ϵ_{opt}, squares), the range between minimal and maximal local tolerance is shown. The requested tolerance TOL is shown with stars, the fixed local tolerance ϵ_{fix} with circles. Right: optimal number J of inexact SDC iterations versus number n of time steps for optimal variable (J_{opt}, squares) and fixed (J_{fix}, circles) local tolerances.

Next we look into the dependence of local tolerances and optimal iteration counts on the time step size T. Let us consider tolerances $\text{TOL} = e_c/\sqrt{n}$ depending on the time step size π/n. As shown in Figure 2, left, they decrease as $n^{-2N-1/2}$ according to the sixth-order collocation error and the error accumulation of order $\frac{1}{2}$. As expected, the fixed local tolerance ϵ_{fix} and the minimal variable local tolerance $\min_{i,j} \epsilon_i^{[j]}$ stay very close to each other and also close to TOL, but decrease roughly one order slower. This is due to $\alpha, \kappa = \mathbb{O}(t_N) = \mathbb{O}(n^{-1})$, and leads to the surprising fact that for small time steps the allowed evaluation error can be larger than the requested tolerance. Obviously, the heuristic choice $\epsilon_{\text{fix}} = c\text{TOL}$ for some fixed $c < 1$ is suboptimal for small time steps.

As intended, the maximal local tolerance, encountered in the very first inexact SDC sweep, is much larger than the minimal one, which is the basis for the envisioned performance gain. It also decreases much slower than the step tolerance TOL due to the fact that $\rho \to 0$ for $t_N \to 0$.

The optimal number of sweeps shown in Figure 2, right, is rather different for fixed and variable local tolerances, with a factor of two in between. This is due to the intended slower contraction rate in (3-24) compared to (3-14). As each sweep increases the order of the SDC integrator by one, and the tolerance TOL is of order $n^{-6.5}$, we expect at least seven sweeps to be necessary. This is nicely reflected by the fixed local tolerance scheme resorting to an optimal value of eight sweeps over a range of step sizes. For larger step sizes, the growth in the contraction rate ρ destroys this asymptotic property.

Finally, we take a look at the predicted efficiency gain over the simple fixed local tolerance strategy in dependence on time step size and overall tolerance. The total work per step induced by the choices of local tolerances is shown in Figure 3. The ratio of more than 10^{15} of computational effort between $n = 2$ and $n = 64$

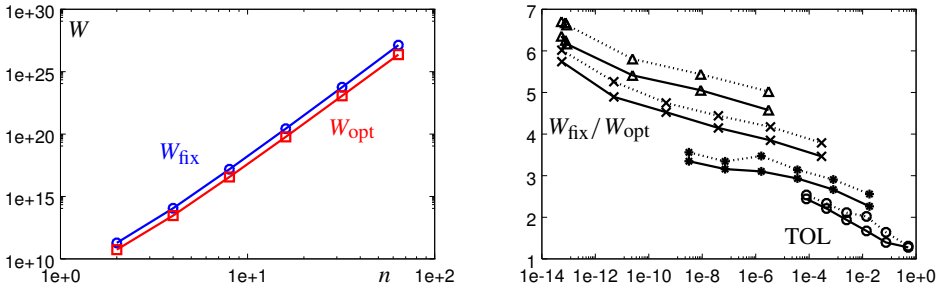

Figure 3. Left: total work per time step for fixed (W_{fix}, circles) and variable (W_{opt}, squares) local tolerances versus the number of time steps. Right: ratio of total work of fixed and variable local tolerances versus the requested tolerance TOL, for work model parameter $d = 2$ (solid lines) and $d = 3$ (dotted lines), number of collocation points $N \in \{1, 2, 3, 4\}$ (circles, stars, crosses, triangles), and different number n of time steps.

is due to the high accuracy of the Gauss collocation and the slow convergence of linear finite elements assumed with $d = 2$. According to (3-12), the work is of order $\mathbb{O}(\epsilon^{-d}) = \mathbb{O}(n^{d(2N-1/2)})$, which amounts here to a growth of n^{11}. Obviously, the high accuracies reached in the model problem are unrealistic in practical finite element computation. The ratio between the work for fixed and local tolerances shown in detail in Figure 3, right, adheres to the theoretical order $-\log \text{TOL}$, with minor differences due to different collocation order N. A small but consistent impact of spatial dimension d can be observed, with slightly larger efficiency gain for higher dimension.

Numerical computations. Up to here, the results were just predictions, theoretical values obtained from the work and error models derived in Section 3. Of particular interest is whether these model predictions coincide with actual computation.

In Figure 4, contraction rate and final time error of inexact SDC computations are shown. Inexact evaluation of the right-hand side is imitated by adding a random perturbation of size $\epsilon_i^{[j]}$ and uniformly distributed direction. On the left, estimated contraction rates are shown, obtained by regression over the complete SDC iteration. As expected, the exact SDC contraction factor ρ decreases roughly linearly with the time step size. The fixed local tolerance iteration converges with a very similar rate, since the rather small allowed errors can only affect the last sweeps. The optimal rate for variable local tolerances is larger: from (3-24) we expect a rate of $\rho^{1/(d+1)}$, which is indeed achieved. The slightly faster convergence can be attributed to the errors in actual computation not realizing the theoretical worst case.

In Figure 4, right, the final time deviation of the inexact SDC iterations from the limit point, the collocation solution, is shown, relative to the error of the collocation solution itself. The sample mean of twenty realizations is plotted together with the standard deviation, since, in contrast to all other figures, the actual errors depend

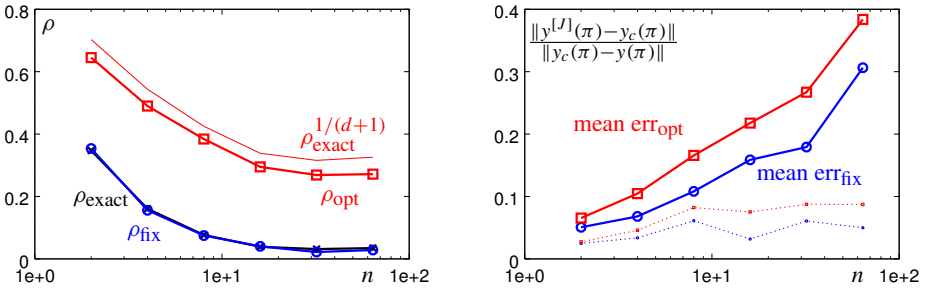

Figure 4. Left: observed contraction factors ρ for exact SDC (ρ_{exact}, crosses), fixed local tolerances (ρ_{fix}, circles), and optimal variable local tolerances (ρ_{opt}, squares) versus number n of time steps. The theoretical contraction rate of $\rho^{1/(d+1)}$ for variable local tolerances is plotted for reference. Right: final time difference between inexact SDC methods and collocation solution, relative to the collocation error. Solid lines are sample means, and dotted lines show the standard deviation.

significantly on the random inexactness of the realizations. We observe that the error model used in defining local tolerances works reasonably well, with comparable final time errors for fixed and optimal variable local tolerances. Again, numerical computations are more accurate than predicted by the worst case estimates. The slow but steady increase with the number n of time steps suggests that the normally distributed local errors do not simply add up, as has been assumed when choosing the tolerance $\text{TOL} \sim n^{-1/2}$.

4B. ***Iterative solver example: heat equation.*** Diffusion processes like heat conduction are usually solved by implicit time-stepping schemes, where solving the arising sparse large-scale linear systems may dominate the computational effort. Here we consider as a prototypical example the linear heat equation

$$\begin{aligned} \dot{y} &= \Delta y, \quad && \text{in } \Omega, \\ y &= 0, && \text{on } \partial\Omega, \\ y &= y_0, && \text{for } t = 0, \end{aligned}$$

on the domain $\Omega =]0, 2\pi[$ and the initial value $y_0 = \chi_{]0,\pi]}$. For spatial discretization, we employ second-order equidistant finite differences on $n = 128$ intervals. We consider a single SDC time step of length $T = 1$ using $N = 4$ Radau-IIa collocation nodes and implicit Euler as the basic method. The exact SDC contraction factor can thus be assumed to be $\rho \approx 0.62$ [27].

The arising linear systems (2-7) assume the form $(I - (t_i - t_{i-1})A)\delta_i^{[j]} = R_i^{[j]}$ with stiffness matrix A. Even though these tridiagonal systems can be solved efficiently with a direct solver, we use iterative solvers in order to evaluate the impact of truncation on inexact SDC performance. As extreme cases we consider simple Jacobi iterations with asymptotic contraction rate $\rho_{\text{Jac}} \approx 1 - 50/n^2$ and multigrid

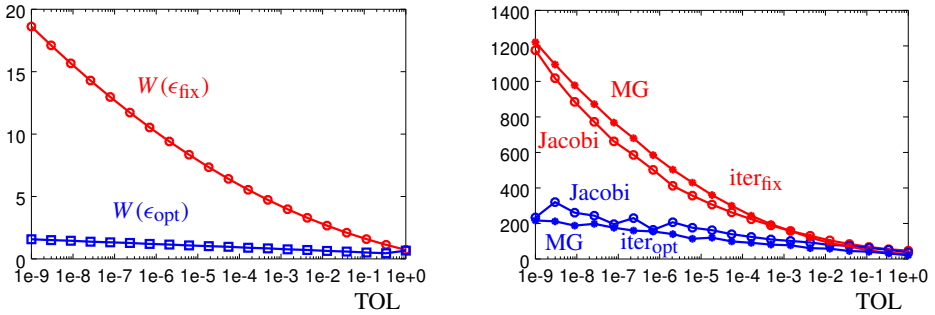

Figure 5. Computational effort of computing an inexact SDC step for the heat equation versus the desired accuracy TOL $\in [10^{-9}, 10^{-1}]\|y^{[0]} - y_c\|$. Left: predicted computational effort $W(\epsilon)$ in arbitrary work units for fixed absolute tolerance and optimized local tolerances. Right: total number of linear solver iterations for fixed and optimized local tolerances, for both Jacobi and multigrid linear solvers. The numbers of Jacobi iterations have been scaled down by a common factor such that they have the same mean as the multigrid iteration numbers, in order to allow a comparison of relative effort between fixed and optimized tolerance choices.

V-cycles with two damped Jacobi presmoothing steps resulting in a contraction rate $\rho_{\text{MG}} \approx 0.25$. The truncation work model (3-13) and consequently also the optimal local tolerances are, however, independent of the iterative solver's contraction rate.

Let us focus on the computational effort incurred by the different local tolerance choices, both predicted and realized. The predicted work $W(\epsilon)$ for fixed and optimized local tolerances is plotted versus the requested SDC iteration accuracy TOL in Figure 5, left, and shows a significant expected benefit of local tolerance optimization. The choice of geometrically decreasing local tolerances $\epsilon_i^{[j]} = \beta\rho^{\gamma j}$ as considered in Section 3D with optimal value $\gamma = 1$ leads to results almost indistinguishable from the optimized tolerances, and is therefore not considered separately.

The actually required work, in terms of number of linear solver iterations, is shown in Figure 5, right, for both Jacobi and multigrid solvers. Note that the iteration numbers of the Jacobi solver have been scaled down by a common factor, such that the relative work between fixed and optimized local tolerances can be observed. While the actual work reduction is less than the predicted one, a factor of five rather than ten for high accuracy, the qualitative behavior is captured very well by the theoretical work model. Moreover, despite the huge difference in convergence speed between Jacobi and multigrid solvers, the relative effort between fixed and optimized local tolerances, and between different required SDC tolerances, is essentially unaffected by the choice of solver, which agrees rather well with the truncation work model derivation in Section 3B.

As before, the accuracy actually achieved is better than the requested tolerance. The ratio $\|y^{[J]} - y_c\|/\text{TOL}$ of actual error and tolerance is (almost) always less

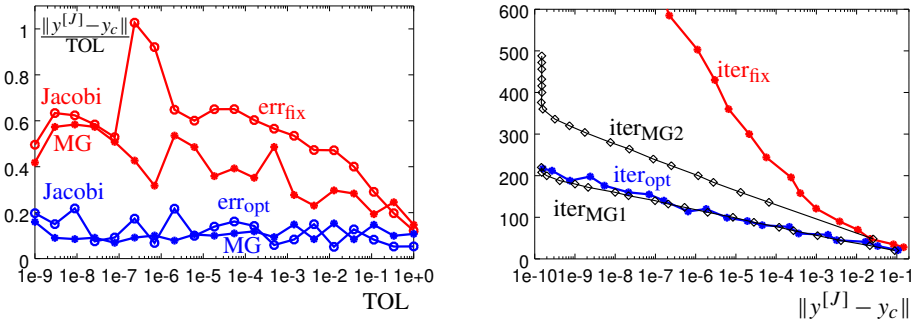

Figure 6. Left: relative deviation of achieved accuracy $\|y^{[J]} - y_c\|$ from the requested tolerance TOL for different choices of local tolerances and linear solvers. Right: total number of linear solver iterations versus the achieved SDC accuracy $\|y^{[J]} - y_c\|$ for different strategies of solving linear systems. Shown are fixed and optimized local tolerances with multigrid solver, as well as simple heuristics of performing exactly one or two multigrid V-cycles.

than one, as predicted by the error bound (3-9). The inefficiency incurred by the error bound not being sharp is, however, moderate, since the achieved accuracy is less than TOL by a factor between two and ten. It depends on the choice of local tolerances, but not much on the linear solver; see Figure 6, left. Consequently, the computational effort required to achieve a certain accuracy, shown in Figure 6, right, resembles very much the work versus requested tolerance shown in Figure 5.

Using a fixed number of linear solver iterations is a simple heuristic for inexact implicit SDC methods [20; 24]. With this choice, linearly convergent solvers can lead to a convergent scheme with expected contraction factor $\max(\rho, \rho_{\text{it}})$, and resembles the geometrically decreasing local tolerances with $\gamma \leq 1$. The efficiency on the heat equation example is comparable to optimized local tolerances for just one V-cycle, and slightly worse for two V-cycles. The best number of iterations will, of course, depend on the problem. A reasonable value can be assumed to be $\lceil \log \rho / \log \rho_{\text{it}} \rceil$.

4C. ***Monte Carlo example: smoothed molecular dynamics.*** Classical molecular dynamics [2] is generally described by Newtonian mechanics of the positions $x \in \mathbb{R}^{nd}$ of n atoms in $\mathbb{R}^d$ with mass M influenced by a potential V:

$$M\ddot{x} = -\nabla V(x). \tag{4-1}$$

One interesting quantity is the time it takes to exit a given potential well or to move between two wells. The computation of these times is expensive as the transitions are rare events, and long trajectories need to be computed before such an event is observed. Statistic reweighting techniques [23] allow one to compute the exit times of interest from exit times induced by a modified potential $\bar{V}$ with shorter

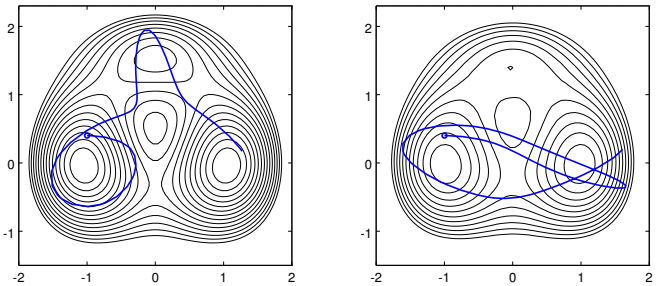

Figure 7. Potential and considered trajectory. Left: original potential V from (4-2). Right: the smoothed potential $\bar{V}$ for $\lambda = 0.316$. The equipotential lines are at the same levels in both pictures.

exit times. One of the modifications in use is potential smoothing by diffusion, i.e., $\bar{V} := V(\lambda)$ with $\partial V/\partial\lambda = \Delta V$. As the number n of involved atoms is usually large, computing $\bar{V}$ by finite element or finite difference methods is out of the question. Instead, pointwise evaluation by convolution with the Green's function is performed [14] using importance sampling:

$$\begin{aligned}\nabla\bar{V}(x) &= (\lambda\sqrt{2\pi})^{-nd}\int_{\mathbb{R}^{nd}} \nabla V(x+s)\exp(-s^2/(2\lambda^2))\,ds\\ &= (\lambda\sqrt{2\pi})^{-nd}\int_{\mathbb{R}^{nd}} (\nabla V(x+s) - Hs)\exp(-s^2/(2\lambda^2))\,ds\\ &\approx \frac{1}{m}\sum_{i=1}^{m}(\nabla V(\xi_i) - H(\xi_i - x)) =: \nabla\widehat{V}_m(x),\end{aligned}$$

where the random variable ξ is normally distributed with mean x and covariance λI, and $H \in \mathbb{R}^{nd}$ is arbitrary. The expected error is proportional to $m^{-1/2}$ and can be estimated in terms of the sample covariance

$$\sigma_m^2 = \frac{1}{m-1}\sum_{i=1}^{m} s_i s_i^T, \quad s_i = \nabla V(\xi_i) - H(\xi_i - x) - \nabla\widehat{V}(x),$$

as

$$E[\|\nabla\bar{V}(x) - \nabla\widehat{V}_m(x)\|] \approx \frac{\|\sigma_m\|}{\sqrt{m}}.$$

Obviously, s_i and consequently σ_m are particularly small if H is the Hessian of V.

When evaluating $\widehat{V}$ with a requested local tolerance ϵ, the number of sampling points is doubled until $\|\sigma_m\| \le \sqrt{m}\epsilon$. This defines a realization of $\widehat{V}_\epsilon(x)$. Note that this does not give an actual error bound, such that the error analysis and tolerance selection from Section 3 only hold in a probabilistic sense.

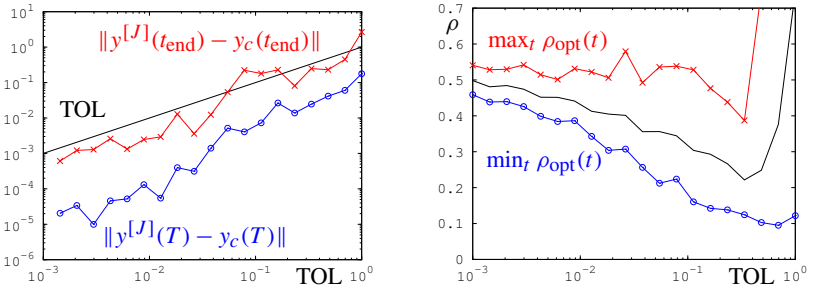

Figure 8. Numerical result averaged over fifteen realizations. Left: estimated error after the first time step of length T (circles) and at final time t_{end} (crosses) versus the requested tolerance TOL. Right: maximal, average, and minimal observed contraction factors ρ_{opt} of the inexact SDC method in all time steps versus the requested step tolerance.

As a simple test problem of this type we consider $n = 1$ and $d = 2$ with $M = I$,

$$V(x) = 3\exp(-\|x - e_2\|^2) - 3\exp(-\|x - 5e_2\|^2) - 5\exp(-\|x - e_1\|^2) \\ - 5\exp(-\|x + e_1\|^2) + (x_1^4 + (x_2 - 1/3)^4)/5, \quad \text{where } (e_i)_j = \delta_{ij}, \quad (4\text{-}2)$$

initial value $x(0) = [-1, 0.4]^T$, $\dot{x}(0) = [2.1, 0]^T$ in the vicinity of one of the three local energy minimizers, final time $t_{\text{end}} = 6$, and variance $\lambda = 0.316$. Despite its simplicity, the potential (4-2) as shown in Figure 7 is an interesting test case, as the direct path between the two deep wells crosses a higher potential barrier than the indirect path via the third, shallow well.

Figure 7 shows the original potential V as defined in (4-2) and the considered trajectory on the left, and the smoothed potential $\bar{V}$ for $\lambda = 0.1$ on the right. The shallow well on the top has almost vanished, and the potential barrier between the two dominant wells is much lower. Consequently, the trajectory crosses the barrier easily now and alternates between the two wells.

The ODE (4-1) is transformed into a first-order system to fit into the setting (2-1). For the tests, $N = 4$ collocation points have been used and $n = 15$ equidistant time steps. The numerically observed exact SDC contraction factor varies roughly in a range $[0.15, 0.24]$. For simplicity, a fixed value of 0.2 has been used for computing local tolerances. For the Lipschitz condition (3-3), we notice that f' has values with purely imaginary spectrum, and estimate $L_f(\tau) = \max_{y \in B} \|I + \tau f'(y)\|$ numerically by evaluating $f'(y_0)$ in each step using Monte Carlo integration of V''. In each time step, the initial iteration error $\|y^{[0]} - y_c\|$ is estimated by substituting a single explicit Euler step for y_c, which here yields a reasonable estimation error of usually less than 50% with a minor impact on local tolerances.

The results shown in Figure 8 indicate that the inexact SDC method works essentially as expected, even though the obtained errors $\|y(T) - y_c(T)\|$ are smaller than the target value TOL by one to two orders of magnitude. This is probably

due to the error propagation result (3-5) reflecting the worst case rather than the average case. Replacing the generously used triangle inequality by sharper bounds, however, would require not only prescribing the magnitude of the evaluation error but also restricting its direction. If possible and practicable at all, this would require the error analysis to be very much specific for particular problems or right-hand side evaluation schemes.

The interpretation that the observed, better than desired accuracies are due to average versus worst case is supported by the observed inexact SDC contraction rates shown in Figure 8, right. With an exact SDC contraction rate $\rho \approx 0.2$, the targeted inexact contraction rate is $\rho^{1/(d+1)} \approx 0.58$, very close to the worst cases observed in actual computation. There is, however, a significant gap between the best and the worst encountered contraction rates, suggesting that the worst-case behavior is captured well by the theoretical derivations.

Conclusion

The theoretically optimal choice of iteration counts and local tolerances when evaluating basic steps in SDC methods as derived here allows significant savings in computational effort compared to naive strategies. Effort reduction factors between two and six have been observed in examples. Thus, exploiting the inexactness that is possible in SDC methods appears to be attractive for expensive simulations.

The local tolerances are defined in terms of problem-dependent quantities, in particular Lipschitz constants L_f, initial iteration error $\|y^{[0]} - y_c\|$, and contraction factor ρ of exact SDC iterations, which are usually not directly available a priori. For a practical implementation of the optimal choice, adaptive methods based on cheap a posteriori estimates of these quantities are needed. We have considered a particular weak model of error type: independent errors for each evaluation, which are likely to line up to the worst case. Correspondingly, worst-case error bounds have been derived and optimized. In concrete computational problems, often more of the error structure is known, and slightly different approaches would be more appropriate. In sampling problems such as the smoothed molecular dynamics example, the random errors tend to cancel out to some extent. Looking at the average behavior instead of the worst case allows us to use larger local tolerances. On the other hand, the errors are highly correlated in several finite element computations. Consequently, the error differences are small, which leads to different error propagation through the SDC iteration. Extending the approach to these settings is the subject of further research.

Acknowledgements

Partial funding by BMBF grant SOAK is gratefully acknowledged. The authors would like to thank Marcus Weber for providing the molecular dynamics example.

Appendix: Uniqueness of work minimizer

Here we prove that for fixed local tolerance ϵ_{fix}, the continuous relaxation of the work model with respect to the iteration count J is quasiconvex and thus has a unique minimizer.

Theorem. *Let*

$$W(J) = \frac{J+1}{(\text{TOL} - \rho^J \delta)^d}$$

with $\delta > \text{TOL} > 0$, $d > 0$, *and* $0 < \rho < 1$. *Then* W *has exactly one local minimizer on* $]J_{\min}, \infty[$, *where* $J_{\min} = \log(\text{TOL}/\delta)/\log\rho$.

Proof. The derivative of W is

$$W'(J) = \frac{(\text{TOL} - \rho^J\delta)^d - (J+1)d(\text{TOL} - \rho^J\delta)^{d-1}(-\delta)\rho^J \log\rho}{(\text{TOL} - \rho^J\delta)^{2d}}.$$

We are just interested in the zeros and the sign of the derivative, and multiply with $\delta^{-1}(\text{TOL} - \rho^J\delta)^{d+1} > 0$ for simplification, which gives $\operatorname{sgn} W'(J) = \operatorname{sgn} q(J)$ with

$$q(J) := \frac{\text{TOL}}{\delta} - \rho^J + (J+1)d\rho^J \log\rho.$$

We obtain $q(J_{\min}) = (J_{\min}+1)d\rho^{J_{\min}}\log\rho < 0$ and $q(J) \to \text{TOL}/\delta > 0$ for $J \to \infty$. Since q is continuous, it has an odd number of zeros in $]J_{\min}, \infty[$.

Next we consider

$$\begin{aligned} q'(J) &= \rho^J \log\rho((J+1)d\log\rho - 1) + \rho^J d\log\rho \\ &= \rho^J \log\rho((J+1)d\log\rho + d - 1). \end{aligned}$$

Any zeros of q' have to satisfy $(J+1)d\log\rho + d - 1 = 0$, such that there is at most one zero of q' and correspondingly at most one extremum of q. If q had more than one zero, i.e., at least three zeros, it would have at least two extrema, which is not the case. Thus, q has exactly one zero and consequently W exactly one extremum. The sign of W' changes from negative to positive there, such that W has exactly one local minimizer. □

References

[1] P. Alfeld, *Fixed point iteration with inexact function values*, Math. Comp. **38** (1982), no. 157, 87–98. MR Zbl

[2] M. P. Allen and D. J. Tildesley, *Computer simulation of liquids*, Oxford University, 1987. Zbl

[3] P. Amodio and L. Brugnano, *A note on the efficient implementation of implicit methods for ODEs*, J. Comput. Appl. Math. **87** (1997), no. 1, 1–9. MR Zbl

[4] J. Barnes and P. Hut, *A hierarchical $O(N \log N)$ force-calculation algorithm*, Nature **324** (1986), 446–449.

[5] P. Birken, *Termination criteria for inexact fixed-point schemes*, Numer. Linear Algebra Appl. **22** (2015), no. 4, 702–716. MR Zbl

[6] J. Carrier, L. Greengard, and V. Rokhlin, *A fast adaptive multipole algorithm for particle simulations*, SIAM J. Sci. Statist. Comput. **9** (1988), no. 4, 669–686. MR Zbl

[7] G. J. Cooper and J. C. Butcher, *An iteration scheme for implicit Runge–Kutta methods*, IMA J. Numer. Anal. **3** (1983), no. 2, 127–140. MR Zbl

[8] G. J. Cooper and R. Vignesvaran, *A scheme for the implementation of implicit Runge–Kutta methods*, Computing **45** (1990), no. 4, 321–332. MR Zbl

[9] P. Deuflhard and F. Bornemann, *Scientific computing with ordinary differential equations*, Texts Appl. Math., no. 42, Springer, 2002. MR Zbl

[10] F. P. E. Dunne and D. R. Hayhurst, *Efficient cycle jumping techniques for the modelling of materials and structures under cyclic mechanical and thermal loading*, Eur. J. Mech. A Solid. **13** (1994), no. 5, 639–660. Zbl

[11] A. Dutt, L. Greengard, and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT **40** (2000), no. 2, 241–266. MR Zbl

[12] B. V. Faleichik, *Analytic iterative processes and numerical algorithms for stiff problems*, Comput. Methods Appl. Math. **8** (2008), no. 2, 116–129. MR Zbl

[13] K. Frischmuth and D. Langemann, *Numerical calculation of wear in mechanical systems*, Math. Comput. Simulation **81** (2011), no. 12, 2688–2701. MR Zbl

[14] E. Gallicchio, S. A. Egorov, and B. J. Berne, *On the application of numerical analytic continuation methods to the study of quantum mechanical vibrational relaxation processes*, J. Chem. Phys. **109** (1998), no. 18, 7745–7755.

[15] R. W. S. Grout, *Mixed-precision spectral deferred correction*, preprint CP-2C00-64959, National Renewable Energy Laboratory, 2015.

[16] S. Güttel and J. W. Pearson, *A rational deferred correction approach to PDE-constrained optimization*, preprint, University of Kent, 2016.

[17] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations, I: Nonstiff problems*, 2nd ed., Springer Series Comput. Math., no. 8, Springer, 1993. MR Zbl

[18] J. Huang, J. Jia, and M. Minion, *Accelerating the convergence of spectral deferred correction methods*, J. Comput. Phys. **214** (2006), no. 2, 633–656. MR Zbl

[19] L. O. Jay and T. Braconnier, *A parallelizable preconditioner for the iterative solution of implicit Runge–Kutta-type methods*, J. Comput. Appl. Math. **111** (1999), no. 1–2, 63–76. MR Zbl

[20] M. L. Minion, R. Speck, M. Bolten, M. Emmett, and D. Ruprecht, *Interweaving PFASST and parallel multigrid*, SIAM J. Sci. Comput. **37** (2015), no. 5, S244–S263. MR Zbl

[21] J. Nocedal and S. J. Wright, *Numerical optimization*, Springer, 1999. MR Zbl

[22] J. M. Ortega and W. C. Rheinboldt, *Iterative solution of nonlinear equations in several variables*, Academic, 1970. MR Zbl

[23] C. Schütte, A. Nielsen, and M. Weber, *Markov state models and molecular alchemy*, Mol. Phys. **113** (2015), no. 1, 69–78.

[24] R. Speck, D. Ruprecht, M. Minion, M. Emmett, and R. Krause, *Inexact spectral deferred corrections*, Domain decomposition methods in science and engineering XXII (T. Dickopf, M. J. Gander, L. Halpern, R. Krause, and L. F. Pavarino, eds.), Lect. Notes Comput. Sci. Eng., no. 104, Springer, 2016, pp. 389–396. MR Zbl

[25] W. Tutschke, *Solution of initial value problems in classes of generalized analytic functions*, Springer, 1989. MR Zbl

[26] P. J. van der Houwen and J. J. B. de Swart, *Triangularly implicit iteration methods for ODE-IVP solvers*, SIAM J. Sci. Comput. **18** (1997), no. 1, 41–55. MR Zbl

[27] M. Weiser, *Faster SDC convergence on non-equidistant grids by DIRK sweeps*, BIT **55** (2015), no. 4, 1219–1241. MR Zbl

[28] M. Wilhelms, G. Seemann, M. Weiser, and O. Dössel, *Benchmarking solvers of the monodomain equation in cardiac electrophysiological modeling*, Biomed. Eng. **55** (2010), 99–102.

Received February 14, 2017. Revised October 30, 2017.

MARTIN WEISER: weiser@zib.de
Zuse Institute Berlin, Berlin, Germany

SUNAYANA GHOSH: sunayanag@gmail.com
Zuse Institute Berlin, Berlin, Germany

Communications in Applied Mathematics and Computational Science

msp.org/camcos

See inside back cover or msp.org/camcos for submission instructions.

The subscription price for 2018 is US $100/year for the electronic version, and $150/year (+$15, if shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscriber address should be sent to MSP.

Communications in Applied Mathematics and Computational Science (ISSN 2157-5452 electronic, 1559-3940 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840, is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

CAMCoS peer review and production are managed by EditFLOW® from MSP.

PUBLISHED BY

mathematical sciences publishers

nonprofit scientific publishing

http://msp.org/

Communications in Applied Mathematics and Computational Science

vol. 13 no. 1 2018