Zuse Institute Berlin

Takustr. 7
14195 Berlin
Germany

MARTIN WEISER AND SUNAYANA GHOSH

Theoretically optimal inexact SDC methods

ZIB Report 16-52 (October 2016)

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

Theoretically optimal inexact SDC methods

Martin Weiser and Sunayana Ghosh

November 1, 2016

Abstract

In several inital value problems with particularly expensive right hand side computation, there is a trade-off between accuracy and computational effort in evaluating the right hand sides. We consider inexact spectral deferred correction (SDC) methods for solving such non-stiff initial value problems. SDC methods are interpreted as fixed point iterations and, due to their corrective iterative nature, allow to exploit the accuracy-work-tradeoff for a reduction of the total computational effort. On one hand we derive an error model bounding the total error in terms of the right hand side evaluation errors. On the other hand, we define work models describing the computational effort in terms of the evaluation accuracy. Combining both, a theoretically optimal tolerance selection is worked out by minimizing the total work subject to achieving the requested tolerance.

1 Introduction

In several initial value problems of the form

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

the evaluation of the right hand side f involves a significant amount of computation, and approximate results can be obtained much faster than exact ones. Examples are reaction-diffusion equations, where implicit time stepping schemes rely on iterative solvers [22, 26], molecular and stellar dynamics, where the exact evaluation of long range interactions is $\mathcal{O}(N^2)$ but can be approximated by clustering or fast multipole methods in $\mathcal{O}(N \log N)$ or $\mathcal{O}(N)$ time [4, 6], or cycle jump techniques for highly oscillatory problems of wear or fatigue [10, 13].

While the possibilities to exploit the trade-off between accuracy and computational cost in right hand side evaluation for improved simulation performance is rather limited in usual time stepping schemes such as explicit Runge-Kutta, extrapolation, or multistep schemes, iterative methods for solving implicit Runge-Kutta equations [3, 12, 24] can in principle correct inexact right hand side evaluations in subsequent iterations. Spectral deferred correction methods (SDC) [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 [22], where a small fixed number of V-cycles has been found to be sufficient. In this paper we will analyze the right hand side error propagation through the SDC iteration and, following the approach of Alfeld [1] for inexact fixed point iterations, derive an a priori selection of tolerances for the right hand side evaluation that leads to theoretically optimal efficiency of the overall integration scheme.

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 right hand side evaluations. The main Section 3 introduces an error model for quantifying the error propagation, work models for quantifying the computational cost, and the optimization of the accuracy per work to derive an optimal selection of tolerances. Effectivity 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

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

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, 23]. 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 just 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 can therefore restrict the presentation to a single time step $[0,T]$. Also without loss of generality, we assume (2.1) to be autonomous.

2.1 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, $i = 1,\dots,N$ within the interval $[0,T]$:

$$\begin{cases} y_c'(t_i) = f(y_c(t_i)), & i = 1,\dots,N \\ y_c(0) = y_0, \end{cases} \tag{2.2}$$

For simplicity of indexing, we define $t_0 = 0$. Popular choices for collocation points are equidistant nodes or Gauss-Legendre, Lobatto and Radau points. For a detailed discussion of collocation methods we refer to [9, 16].

The inital value problem (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 [17]:

$$\begin{cases} y_c(t_i) = y_c(t_{i-1}) + \sum_{k=1}^{N} S_{ik} f(y_c(t_k)), & i = 1,\dots,N \\ y_c(0) = y_0, \end{cases} \tag{2.3}$$

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

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

2.2 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 anyway present, an exact solution of (2.2) is not required. Thus, iterative methods form an interesting class of solvers, see, e.g., [7,8,18]. Here we consider Spectral Deferred Correction Methods (SDC). 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 [17, 25]. 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, see [25].

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

$$\begin{aligned} y_c(t_i) - y_i^{[j]} &= (y_c - y^{[j]})(t_{i-1}) + \sum_{k=1}^{N} S_{ik} \left(f(y_c(t_k)) - {y^{[j]}}'(t_k) \right) \\ &= (y_c - y^{[j]})(t_{i-1}) + \sum_{k=1}^{N} S_{ik} \left(f(y_c(t_k)) - f(y_k^{[j]}) \right) + \sum_{k=1}^{N} S_{ik} \left(f(y_k^{[j]}) - {y^{[j]}}'(t_k) \right) \end{aligned}$$

for $i = 1, \ldots, N$ with initial condition $(y_c - y^{[j]})(0) = 0$. As y_c is computationally unavailable, a correction $\delta^{[j]} \approx y_c - y^{[j]}$ can be defined and computed explicitly as

$$\delta_i^{[j]} = \delta_{i-1}^{[j]} + (t_i - t_{i-1}) \left(f(y_{i-1}^{[j]} + \delta_{i-1}^{[j]}) - f(y_{i-1}^{[j]}) \right) + \sum_{k=1}^{N} S_{ik} f(y_k^{[j]}) - (y_i^{[j]} - y_{i-1}^{[j]}), \quad i = 1,\ldots,N, \tag{2.4}$$

by replacing the spectral quadrature of the first sum by a simple left-looking rectangular rule corresponding to the explicit Euler time stepping scheme suitable for non-stiff problems. Of course, the initial value is $\delta_0^{[j]} = 0$. Here $\delta^{[j]}$ is the polynomial approximation of the exact error function $y_c - y^{[j]}$.

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 computation of the correction $\delta^{[j]}$ realizes a fixed point iteration for the operator $F : Y^{N+1} \to Y^{N+1}$ with $F(y^{[j]}) = y^{[j+1]}$. For convergence analysis, we equip Y^{N+1} with a norm $\|y\| = \|[\|y_0\|_Y, \ldots, \|y_N\|_Y]\|_p$ in terms of the usual p-norm on $\mathbb{R}^{N+1}$ with p to be specified later. If F is Lipschitz continuous with constant $\rho < 1$, i.e.

$$\|F(x) - F(y)\| \le \rho \, \|x - y\|, \qquad \forall x, y \in Y^{N+1},$$

(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, and the time step size T. For sufficiently small time steps, convergence is guaranteed.

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 $\|y_c - y^{[J]}\| \leq \text{TOL}$. Given the contraction rate ρ, and assuming that $\|y_c - y^{[0]}\| > \text{TOL}$, the number of iterations is then bounded by

$$J \leq \left\lceil \frac{\log \frac{\text{TOL}}{\|y_c - y^{[0]}\|}}{\log \rho} \right\rceil .$$

The choice of the initial iterate $y^{[0]}$ can have not only 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 are used as proposed in [25]. Focusing on non-stiff problems, we simply choose $y_i^{[0]} \equiv y_0$ in this paper.

2.3 Perturbations of the right hand side

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 replacing the implicit spectral quadrature term by the explicit 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 $\|f_i^{[j]} - f(y_i^{[j]})\|_Y \leq \epsilon_i^{[j]}$.

Thus the SDC correction $\hat{\delta}^{[j]}$ for inexact right hand sides $f_i^{[j]}$ is obtained as

$$\hat{\delta}_i^{[j]} = \hat{\delta}_{i-1}^{[j]} + (t_i - t_{i-1}) \left(f_{i-1}^{[j+1]} - f_{i-1}^{[j]} \right) + \sum_{k=1}^{N} S_{ik} f_k^{[j]} - (y_i^{[j]} - y_{i-1}^{[j]}), \tag{2.5}$$

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

Given the requirement of computing a final iterate $y^{[J]}$ satisfying the requested accuracy $\|y_c - y^{[J]}\| \leq \text{TOL}$, 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 $\|y^{[J]} - y_c\| \leq \Phi(\epsilon, J)$:

$$\min_{J \in \mathbb{N}, \epsilon \in \mathcal{E} \subset \mathbb{R}^{N \times J+1}} W(\epsilon, J) \quad \text{subject to} \quad \Phi(\epsilon, J) \leq \text{TOL} \tag{3.1}$$

Here, ϵ denotes the $N \times J+1$ matrix of local tolerances $\epsilon_i^{[j]}$, restricted to an admissible set $\mathcal{E}$. We will consider different admissible sets in sections 3.3 to 3.5 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.

3.1 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, see [1,20]. Using the inexact right hand sides $f_i^{[j]}$ realizes a perturbed fixed point operator $\hat{F} : Y^{N+1} \times (\mathbb{R}_+^N \cup 0) \times (\mathbb{R}_+^N \cup 0) \to Y^{N+1}$ with the exact limit case $\hat{F}(y, 0, 0) = F(y)$. The $(j+1)$-th iterate is given by

$$y^{[j+1]} = \hat{F}(y^{[j]}, \epsilon^{[j]}, \epsilon^{[j+1]}), \qquad j = 0, \dots, J-1, \tag{3.2}$$

where $\epsilon^{[j]} \in \mathbb{R}_+^N$ denotes the right hand side evaluation tolerances $(\epsilon_1^{[j]}, \dots, \epsilon_N^{[j]})^T$. Below we consider the convergence of (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 an estimate how the right hand side errors bounded by $\epsilon_i^{[j]}$ are transported through the SDC sweep.

Definition 3.1. Let us assume there is a nonnegative function $L_f : \mathbb{R}_+ \to \mathbb{R}_+$ such that the right hand side f satisfies the Lipschitz condition

$$\|\delta + \tau(f(y+\delta) - f(y))\|_Y \le L_f(\tau)\|\delta\|_Y \quad \text{for all } \tau > 0 \text{ and } \delta, y \in Y. \tag{3.3}$$

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 $\kappa \in \mathbb{R}^{N \times N}$.

Theorem 3.2. *Assume that the ODE's right hand side satisfies condition (3.3). Then, for* $\epsilon^{[j]}, \epsilon^{[j+1]} \in \mathbb{R}_+^N$,

$$\|\hat{F}(y, \epsilon^{[j]}, \epsilon^{[j+1]}) - F(y)\| \le \|\kappa(\epsilon^{[j]} + \epsilon^{[j+1]}) + |S|\epsilon^{[j]}\|_L \tag{3.4}$$

holds 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-*δ. $|S| \in \mathbb{R}^{N \times N}$ *denotes the entry-wise absolute value of the integration matrix* S.

Proof. From (2.4) and (2.5) we obtain for the SDC corrections $\hat{\delta}_i$ the estimate

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

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

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

Taking the norm over $i = 1, \ldots, N$ yields the claim (3.4). □

With (3.4) at hand, we are in the position to bound the final time error.

Theorem 3.3. *Let $y^{[0]} \in Y^N$ be given and let $y^{[j+1]}$ be defined by*

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

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

$$\|y^{[J]} - y_c\| \leq \alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \|\epsilon^{[j]}\|_L + \|\kappa\epsilon^{[J]}\|_L + \rho^J \|y^{[0]} - y_c\| =: \Phi(\epsilon, J) \tag{3.5}$$

holds with $\alpha = \|\kappa + |S|\|_L + \rho\|\kappa\|_L$.

Proof. First we show the (slightly stronger) result

$$\|y^{[J]} - y_c\| \leq \sum_{j=1}^{J} \rho^{J-j} \|\kappa(\epsilon^{[j-1]} + \epsilon^{[j]}) + |S|\epsilon^{[j-1]}\|_L + \rho^J \|y^{[0]} - y_c\| \tag{3.6}$$

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

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

which is just (3.6). Applying the triangle inequality and rearranging terms in the sum yields

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

and thus the claim (3.5). □

The error model Φ as defined in (3.5) 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 $\Phi(\epsilon, J) \leq \mathrm{TOL}$ implies $\rho^J\|y^{[0]} - y_c\| \leq \mathrm{TOL}$ and

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

3.2 Work models

Let us assume that the computational effort of evaluating $f_i^{[j]}$ 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,

$$W_{\mathrm{total}}(\epsilon) = \sum_{j=0}^{J}\sum_{i=1}^{N} W_i^{[j]}(\epsilon_i^{[j]}), \tag{3.7}$$

is just the sum of all right hand side evaluation efforts. Hence, common positive factors can be neglected, as they will not affect the minimizer of (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 right hand side evaluation 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

$$W_i^{[j]}(\epsilon_i^{[j]}) := \frac{1}{d}(\epsilon_i^{[j]})^{-d}. \tag{3.8}$$

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

Of course, the asymptotic behavior $W_i^{[j]} \to 0$ for $\epsilon_i^{[j]} \to 0$ 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 $\epsilon_i^{[j]} \leq \epsilon_{\max} = (dW_{\min})^{-1/d}$. We will address this in Section 3.6.

Truncation errors. First we assume the evaluation of the right hand side $f(y_i^{[j]})$ involves the solution of a large sparse linear equation system that is solved approximately by a linearly convergent iterative solver with contraction rate $\rho_{\mathrm{it}} < 1$. In general, a good initial value can be derived from the previous SDC iteration, such that we assume the error bound $\epsilon_i^{[j]}$ is given as

$$\epsilon_i^{[j]} \approx \|f_i^{[j]} - f(y_i^{[j]})\|_Y \leq \rho_{\mathrm{it}}^m\|f_i^{[j-1]} - f(y_i^{[j]})\|_Y,$$

where $m \geq 0$ denotes the iteration count. Assuming $\|f_i^{[j-1]} - f(y_i^{[j]})\|_Y \approx \|f(y_i^{[j-1]}) - f(y_i^{[j]})\|_Y$ and linear convergence $\|y_i^{[j-1]} - y_i^{[j]}\| \leq (1+\rho)\rho^j\|y^{[0]} - y_c\|$ (which will be justified in (3.20)), this can be approximated as

$$\epsilon_i^{[j]} \approx \rho_{\mathrm{it}}^m L_*(1+\rho)\rho^j\|y^{[0]} - y_c\|$$

where $L_* = \lim_{\tau\to\infty} L_f(\tau)/\tau$ is the usual Lipschitz constant of f. As the work is proportional to the number m of solver iterations, we can define the work model

$$W_i^{[j]}(\epsilon_i^{[j]}) := -\log \epsilon_i^{[j]} + \log(L_*(1+\rho)\rho^j \|y^{[0]} - y_c\|), \tag{3.9}$$

where the common positive factor $-\log \rho_{\text{it}}$ has been dropped.

Again, the validity of this work model is limited to local tolerances $\epsilon_i^{[j]} \leq \epsilon_{\max}^{[j]} = L_*(1+\rho)\rho^{j+1}\|y^{[0]} - y_c\|$, as at least one iteration has to be performed in each solver call, corresponding to $W_{\min} = 1$.

Remark 3.4. We have formulated this work model for explicit SDC methods, but it applies to implicit schemes as well. In diffusive processes such as reaction-diffusion equations spatially discretized by the method of lines, implicit time stepping schemes are usually necessary to ensure stability with reasonably large step sizes. However, a highly accurate solution of the arising systems is often not necessary to guarantee stability, e.g., in cardiac simulations [26]. In that case, iterative solvers can be terminated early, and the truncation error corresponds directly to some iteration residual.

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

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

just a special case of (3.8). 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 worked out above exhibit some qualitative properties, which we conjecture to be general properties of plausible work models.

Definition 3.5. A *work model* is a family of strictly convex, positive, and monotonically decreasing function $W_i^{[j]} :]0, (\epsilon_{\max})_i^{[j]}] \to \mathbb{R}_+$ 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 W_{total}, which is strictly convex and monotone.

3.3 Fixed local tolerance

To begin with, we consider heuristic choices of the admissible set $\mathcal{E}$ of local tolerances. The simplest possibility is to take the same value $\epsilon_i^{[j]} \equiv \epsilon_0$ for all right hand side evaluations. In this case, (3.5) reduces to

$$\|y^{[J]} - y_c\| \leq \epsilon_0 \left(\alpha \|\mathbf{1}\|_L \frac{1-\rho^J}{1-\rho} + \|\kappa \mathbf{1}\|_L\right) + \rho^J \|y^{[0]} - y_c\|,$$

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

$$\epsilon_0 = \min\left(\epsilon_{\max}, \frac{\text{TOL} - \rho^J \|y^{[0]} - y_c\|}{\alpha \|\mathbf{1}\|_L \frac{1-\rho^J}{1-\rho} + \|\kappa \mathbf{1}\|_L}\right) \tag{3.10}$$

provides the largest admissible choice, and hence the one that incurs the least computational effort, for given J. With $\epsilon_0(J)$ 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

$$\epsilon_0 = \min\left(\epsilon_{\max}, \frac{\mathrm{TOL} - \rho^J \|y^{[0]} - y_c\|}{\alpha\|\mathbf{1}\|_L/(1-\rho) + \|\kappa\mathbf{1}\|_L}\right).$$

For the finite element work model (3.8), the total work is just $W = N(J+1)\epsilon_0(J)^{-d}/d$. Assuming $\epsilon_0 < \epsilon_{\max}$ and eliminating constant factors, we need to minimize $W(J) \sim (J+1)/(\mathrm{TOL} - \rho^J\|y^{[0]} - y_c\|)^d$. A simple analysis reveals that $W(J)$ is quasi-convex, such that there is exactly one minimizer in $]J_{\min}, \infty[$, see Appendix A. Unfortunately, no closed expression seems to exist, but a numerical computation is straightforward. Due to the quasi-convexity, the optimal $J \in \mathbb{N}$ is one of the neighboring integer values.

The local tolerance is bounded by $\epsilon_0 \leq c\mathrm{TOL}$ for some generic constant c independent of J and TOL. Consequently, the total work is at least

$$W \geq c(J_{\min}+1)\mathrm{TOL}^{-d} = c\left(\frac{\log(\mathrm{TOL}/\|y^{[0]} - y_c\|)}{\log\rho} + 1\right)\mathrm{TOL}^{-d}. \tag{3.11}$$

Apparently, a complexity of $\mathcal{O}(\mathrm{TOL}^{-d})$ is unavoidable, as this is already required for a single right hand side evaluation to the requested accuracy. The logarithmic factor in (3.11), 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 3.4 and 3.5 and investigate it numerically in Section 4.

For completeness we note that in the less interesting case $\epsilon_0 = \epsilon_{\max}$, J is determined by minimizing $W = N(J+1)\epsilon_{\max}^{-d}/d$ subject to

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

i.e.

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

3.4 Geometrically decreasing local tolerances

The next step is to exploit 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 $\epsilon_i^{[j]} = \min(\epsilon_{\max}, \beta\rho^{\gamma j})$ for some $\beta, \gamma > 0$. This has been considered in [5] for $\gamma = 1$ as "adaptive strategy" and is closely related to evaluating implicit Euler steps up to a fixed relative precision in implicit SDC methods, as suggested in [15] or realized in [22] 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_0. Ignoring the impact of $\epsilon_{\max}$, (3.5) results in the slightly stronger accuracy requirement

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

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

$$\beta \leq \frac{\mathrm{TOL} - \rho^J \|y^{[0]} - y_c\|}{\rho^{\gamma(J-1)}(\alpha \|\mathbf{1}\|_L \frac{1-\rho^{(1-\gamma)J}}{1-\rho^{1-\gamma}} + \rho^\gamma \|\kappa \mathbf{1}\|_L)}. \tag{3.12}$$

The total work $W(\epsilon)$ is monotonely decreasing in β, such that (3.1) is solved by equality in (3.12). Of course, $\epsilon_0 = \lim_{\gamma \to 0} \beta$ is recovered in the limit.

Optimizing the iteration count J for the finite element work model, 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 neglect constant factors independent of J and derive the upper bound

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

decreasing monotonically with J towards $\lim_{J\to\infty} W \lesssim \mathrm{TOL}^{-d}$. Compared to (3.11), the complexity to reach the requested tolerance is improved from $\mathcal{O}(\mathrm{TOL}^{-d}|\log \mathrm{TOL}|)$ to $\mathcal{O}(\mathrm{TOL}^{-d})$ independently of γ. 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 $\mathcal{E}$.

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, 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 $\epsilon_i^{[j]} \equiv \epsilon_0$ 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 $\gamma \leq 1$.

3.5 Variable local tolerances

Finally, let us consider the most general admissible set $\mathcal{E} = \{\epsilon \in \mathbb{R}_+^{N\times J} \mid \epsilon_i^{[j]} \leq \epsilon_{\max}\}$ 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 consider the integer variable J of the mixed integer program later on.

We obtain the nonlinear program

$$\min_{\epsilon \in \mathbb{R}_+^{N\times J+1}} W_{\mathrm{total}}(\epsilon) \quad \text{subject to} \quad \Phi(\epsilon, J) \leq \mathrm{TOL}, \quad \epsilon \leq \epsilon_{\max}. \tag{3.13}$$

From the properties of W_{total} and Φ, we immediately obtain the following result.

Theorem 3.6. *If $\rho^J \|y^{[0]} - y_c\| < \mathrm{TOL}$, i.e. the exact SDC iteration converges to the given tolerance, the optimization problem (3.13) 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 (3.5) 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_{total} and convexity of Φ imply uniqueness of a solution. Strict convexity and monotonicity of W_{total} 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.

Next we prove that the optimal sequence of local tolerances is monotonically decreasing.

Theorem 3.7. *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 (3.13). Then $\|\epsilon^{[j]}\|_L \leq \|\epsilon^{[j-1]}\|_L$ holds for all $j = 1, \ldots, J-1$.*

For $p = 1$, componentwise monotonicity holds as well, i.e. $\epsilon_i^{[j]} \leq \epsilon_i^{[j-1]}$ holds for all i and j.

Proof. Let $\tilde{\epsilon}$ be an admissible point for (3.13) with $\|\tilde{\epsilon}^{[k_1]}\|_L < \|\tilde{\epsilon}^{[k_2]}\|_L$ for some $1 \leq k_1 < k_2 < J$. 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, $W_{\mathrm{total}}(\epsilon) = W_{\mathrm{total}}(\tilde{\epsilon})$.

The error bound (3.5), however, is reduced,

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

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

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

such that $\tilde{\epsilon} \neq \epsilon(y^{[0]}, J)$. 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 convexitx, also sufficient conditions for the solution of the constrained optimization problem are derived for $p < \infty$.

Theorem 3.8. *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 (3.13), if and only if there exist multipliers $\mu \in \mathbb{R}$ and $\eta \in \mathbb{R}^{N \times J+1}$ such that*

$$\begin{aligned}
&(W_i^{[j]})'(\epsilon_i^{[j]}) + \mu \, \alpha \rho^{J-1-j} \|\epsilon^{[j]}\|_L^{1-p} \sum_{k=1}^{N} (L\epsilon^{[j]})_k^{p-1} L_{ki} + \eta_i^{[j]} = 0, \qquad j = 0, 1, \ldots, J-1, \\
&(W_i^{[J]})'(\epsilon_i^{[J]}) + \mu \|\kappa\epsilon^{[J]}\|_L^{1-p} \sum_{k=1}^{N} (L\kappa\epsilon^{[J]})^{p-1} (L\kappa)_{ki} + \eta_i^{[J]} = 0, \\
&(\mathrm{TOL} - \Phi(\epsilon, J))\mu = 0, \quad \mu \geq 0, \\
&(\epsilon_{\max} - \epsilon) : \eta = 0, \quad \eta \geq 0.
\end{aligned} \tag{3.14}$$

Here, $\epsilon : \eta$ denotes contraction or Frobenius product.

Proof. Necessary and also sufficient conditions for optimality of ϵ is the stationarity of the Lagrangian

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

for some multiplier $\mu \in \mathbb{R}$ and $\eta \in \mathbb{R}^{N \times J+1}$, see, e.g., [19]. According to (3.5) and (3.7), its partial derivatives are just the expressions in (3.14). □

At this point, the unique minimizer $\epsilon(y^{[0]}, J)$ of the convex program (3.13) can in principle be computed numerically. For $p = 1$, however, explicit analytical expressions can be derived easily due to (3.5) reducing to

$$\begin{aligned}\Phi(\epsilon, J) &= \alpha \sum_{j=0}^{J-1} \rho^{J-1-j} \sum_{k=1}^{N} \sum_{i=1}^{N} L_{ki} \epsilon_i^{[j]} + \sum_{k=1}^{N} \sum_{i=1}^{N} (L\kappa)_{ki} \epsilon_i^{[J]} + \rho^J \|y^{[0]} - y_c\| \\ &= q : \epsilon + \rho^J \|y^{[0]} - y_c\| \end{aligned} \tag{3.15}$$

with

$$q_i^{[j]} = \begin{cases} \alpha \rho^{J-1-j} \sum_{k=1}^{N} L_{ki}, & j < J \\ \sum_{k=1}^{N} (L\kappa)_{ki}, & j = J. \end{cases} \tag{3.16}$$

Then, (3.14) assumes the particularly simple form

$$(W_i^{[j]})'(\epsilon_i^{[j]}) + \mu q_i^{[j]} + \eta_i^{[j]} = 0. \tag{3.17}$$

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 3.2, with $d = 0$ for iterative solvers and $d = 2$ for stochastic sampling.

Theorem 3.9. *Let $p = 1$ and $(W_i^{[j]})'(\epsilon_i^{[j]}) = -(\epsilon_i^{[j]})^{-(d+1)}$. Then the solution $\epsilon = \epsilon(y^{[0]}, J)$ of (3.13) is given by*

$$\epsilon_i^{[j]} = \min((\epsilon_{\max})_i^{[j]}, (\mu q_i^{[j]})^{-1/(d+1)}). \tag{3.18}$$

Locally unconstrained tolerances $\epsilon_i^{[j]} < (\epsilon_{\max})_i^{[j]}$ decrease linearly up to $j = J - 1$:

$$\epsilon_i^{[j]} \sim \rho^{j/(d+1)} \tag{3.19}$$

Proof. From (3.17) we obtain $\epsilon_i^{[j]} = \left(\mu q_i^{[j]} + \eta_i^{[j]}\right)^{-1/(d+1)}$. For $\epsilon_i^{[j]} < (\epsilon_{\max})_i^{[j]}$, $\eta_i^{[j]} = 0$ holds due to complementarity in (3.14), such that (3.18) is satisfied. For $j < J$, (3.16) 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 (3.19).

The other case, $\epsilon_i^{[j]} = (\epsilon_{\max})_i^{[j]}$, implies $(\mu q_i^{[j]})^{-1} = (((\epsilon_{\max})_i^{[j]})^{-(d+1)} - \eta)^{-1} \geq ((\epsilon_{\max})_i^{[j]})^{d+1}$ and hence (3.18). □

The result (3.19) reveals that the heuristic of geometically deceasing local tolerances is indeed of optimal complexity, at least for $\gamma < 1$, and now theoretically justified. Beyond that, an optimal value of γ and different accuracies for the collocation points are provided. We will see in Section 4 that the last issue can have a non-negligible impact on the computational effort.

Let us state two observations. First, it pays off to treat the final local tolerances $\epsilon_i^{[J]}$ separately in Theorem 3.3: 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 most expensive since most accurate right hand side evaluations is reduced, as illustrated in Fig. 1. Second, (3.18) is monotone in μ, such that the actual value of μ is easily computed numerically by solving $\Phi(\epsilon, J) = \mathrm{TOL}$.

In case $\epsilon < \epsilon_{\max}$, the result (3.19) yields

$$\begin{aligned}\|y^{[j]} - y_c\| &\leq \alpha \sum_{k=0}^{j-1} \rho^{j-1-k} \|\epsilon^{[k]}\|_L + \|\kappa\epsilon^{[j]}\|_L + \rho^j \|y^{[0]} - y_c\| \\ &\leq c \left(\sum_{k=0}^{j-1} \rho^{j-1-k} \rho^{k/(d+1)} + \rho^{j/(d+1)} \right) + \rho^j \|y^{[0]} - y_c\| \\ &\leq c\rho^{j/(d+1)} \end{aligned} \tag{3.20}$$

with some generic constant c independent of j (though it depends on J). As expected, the geometric decrease (3.19) translates directly into linear convergence of the inexact SDC iteration. For $d = 0$, this justifies the contraction rate assumed in defining the work model (3.9). Note that Theorem 3.9 does not depend on that assumption.

Moreover, the results (3.19) and (3.20) show 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 the FE discretization error is rather expensive. An optimal tolerance selection therefore assigns a larger portion of the total error to the FE discretization and has to ensure that the SDC iteration error is by a certain factor smaller than the discretization error.

3.6 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 finite element work model, we obtain

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

as long as $\epsilon_i^{[j]} \leq (\epsilon_{\max})_i^{[j]}$ for all i, j. Inserting (3.16) and neglecting constant factors independent of J yields

$$W \leq \sum_{j=0}^{J} \sum_{i=1}^{N} \left(\mu\alpha\rho^{J-1-j} \sum_{k=1}^{N} L_{ki} \right)^{\frac{d}{d+1}} \sim \mu^{\frac{d}{d+1}} \sum_{j=0}^{J} \rho^{\frac{d}{d+1}(J-1-j)} \sim \mu^{\frac{d}{d+1}} \frac{1 - \rho^{\frac{d(J+1)}{d+1}}}{1 - \rho^{\frac{d}{d+1}}}. \tag{3.21}$$

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

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

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

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

holds. Entering this into (3.21) yields

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

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

$$W \lesssim (\mathrm{TOL} - \rho^J \|y^0 - y_c\|)^{-d}. \tag{3.22}$$

The upper bound (3.22) is monotonically decreasing and suggests choosing J as large as possible. In the limit $J \to \infty$, the total work is bounded by

$$W \lesssim \mathrm{TOL}^{-d}.$$

Compared to the result (3.11) for uniform local tolerances, the logarithmic factor $\log \mathrm{TOL}$ is missing, which yields the optimal complexity of evaluating a single right hand side evaluation up to the requested tolerance.

Choosing J very large, however, violates the above assumption of $\epsilon < \epsilon_{\max}$, as $\epsilon^{[j]} \to \infty$ for $J \to \infty$. Due to $\epsilon < \epsilon_{\max}$ and hence $W_i^j \geq W_{\min}$, the total work $W(J)$ grows linearly with J. As closed expressions for a global minimizer of $W(J)$ when taking the local tolerance constraint into account are hard to get, a heuristic selection of J appears to be most promising in practice. The convexity of (3.22) and linear growth of W for large J suggest that the first local minimizer encountered when evaluating $W(J)$ starting at $J_{\min}$ should be a good candidate for the global minimizer.

4 Numerical examples

Here we will illustrate and compare the effectivity of the inexact SDC strategies worked out above. First, the properties of the strategies will be explored using a simple academic test problem. Second, an example from molecular dynamics is considered.

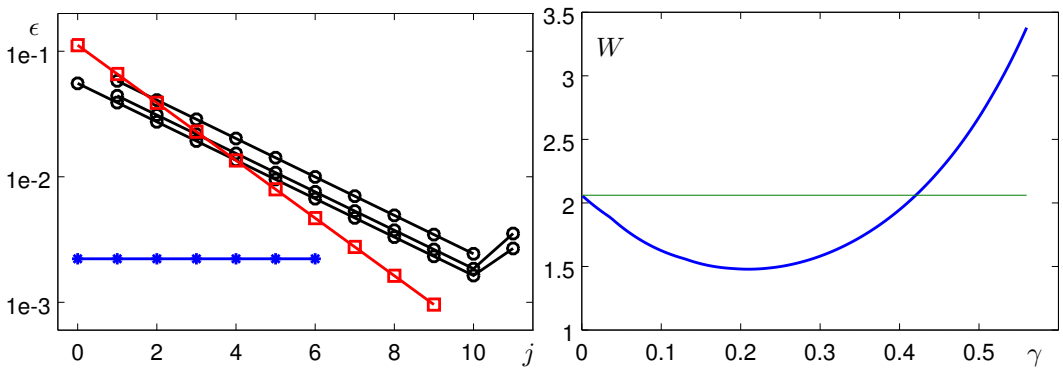

Figure 1: *Left:* Exemplary local tolerances versus iteration number j for the different admissible design sets: fixed (stars), geometrically decreasing (squares, $\gamma = 0.5$), and variable (circles). $n = 2$ steps have been used to define the problem data $\rho = 0.35$, TOL $= 0.05$. *Right:* Relative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

4.1 An illustrative example

As a particularly simple example we consider the harmonic oscillator

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

with initial value $u_0 = 0$, $v_0 = 1$, 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.

Aiming at a final time error comparable to the collocation error, we choose a tolerance $\text{TOL} = e_c/\sqrt{n}$ 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 $\sqrt{n}\text{TOL}$. We use the finite element work model (3.8) with spatial dimension d. With this setting, the quantities entering into the computation of the local tolerances ϵ are the same for all time steps. Unless otherwise stated, $d = 2$ and $N = 3$ are used throughout, such that the collocation scheme is of order 6.

Let us first turn to the local tolerances ϵ prescribed due to (3.10), (3.12), and (3.18), respectively. Exemplary values are shown in Fig. 1, left, versus the iteration number j. Clearly visible is the slow geometric decrease of the variable local tolerances $\epsilon^{[j]}$ with an order $\rho^{j/3}$, slower than the explicitly chosen geometrical decrease $\rho^{\gamma j}$ with $\gamma = 1/2$. The normalized predicted work is 2.06, 2.67, and 1, respectively. 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)

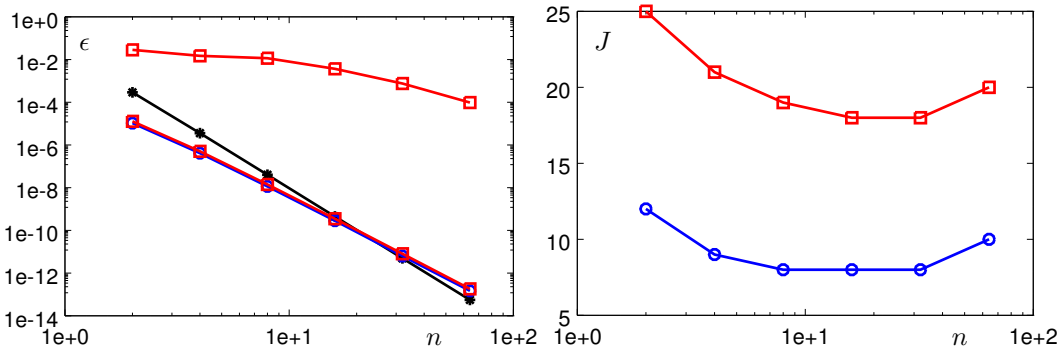

Figure 2: *Left:* Local tolerances ϵ for the inexact SDC iterations versus number n of time steps. For variable local tolerances (squares), the range between minimal and maximal local tolerance is shown. The requested tolerance TOL is shown with stars, the fixed local tolerance with circles. *Right:*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

by imposing less restrictive requirements on the final sweep. 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 Fig. 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.

Now let us consider varying step tolerances TOL. As shown in Fig. 2, left, they decrease as $n^{-2N-1/2}$ according to the sixth order collocation error and the error accumulation of order $1/2$. As expected, the fixed local tolerance ϵ_0 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 $\alpha, \kappa = \mathcal{O}(t_N) = \mathcal{O}(n^{-1})$, 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 $\epsilon_0 = c\,\mathrm{TOL}$ 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 Fig. 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 (3.18) compared to (3.10). As each sweep increases the order of the SDC integrator by one, and the step tolerance is of order 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.

The total work per step induced by the choices of local tolerances is shown in Fig. 3. The ratio of more than 10^{15} of computational effort between $n = 2$ and $n = 64$ is due to the high accuracy of the Gauss collocation and the slow convergence of linear finite elements assumed

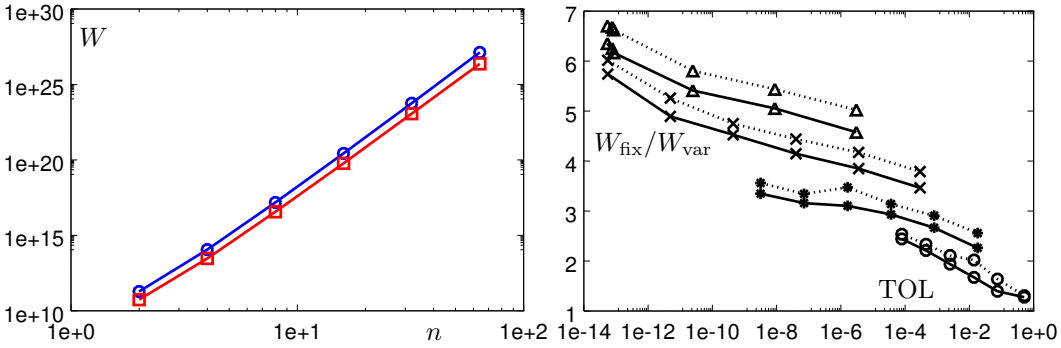

Figure 3: *Left:* Total work per time step for fixed (circles) and variable (squares) local tolerances versus the number of time steps. *Right:* Ratio of total work of fixed and variable local tolerances versus the requestd step tolerance TOL, for spatial dimensions $d = 2$ (solid lines) and $d = 3$ (dotted lines), number of collocation points $N \in \{1, 2, 3, 4\}$ (circles, stars, crosses, triangles), and different number n of time steps.

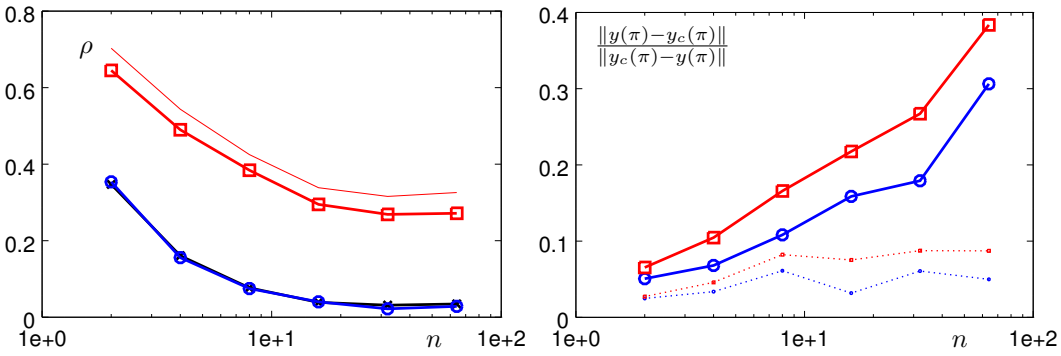

Figure 4: *Left:* Observed contraction factors ρ for exact SDC (crosses), fixed local tolerances (circles), and variable local tolerances (squares) versus number n of time steps. The theoretical contraction rate of $\rho^{1/(d+1)}$ for variable local tolerances is plotted for reference. *Right:* Final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

in the work model. According to (3.8), the work is of order $\mathcal{O}(\epsilon^{-d}) = \mathcal{O}(n^{d(2N-1/2)})$, 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 Fig. 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.

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 Fig. 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 (3.18) 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 Fig. 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 20 realizations is plotted together with the standard deviation, since, in contrast to all other figures, the actual errors depend significantly on the random inexactness realizations. We observe that the error model used in defining local tolerances works reasonably well, with comparable results for fixed and 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 TOL $\sim n^{-1/2}$.

4.2 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 [21] allow to compute the exit times of interest from exit times induced by a modified potential $\bar{V}$ with shorter 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 question. Instead, pointwise evaluation by convolution with the Green's function is performed [14] using importance sampling

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

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

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

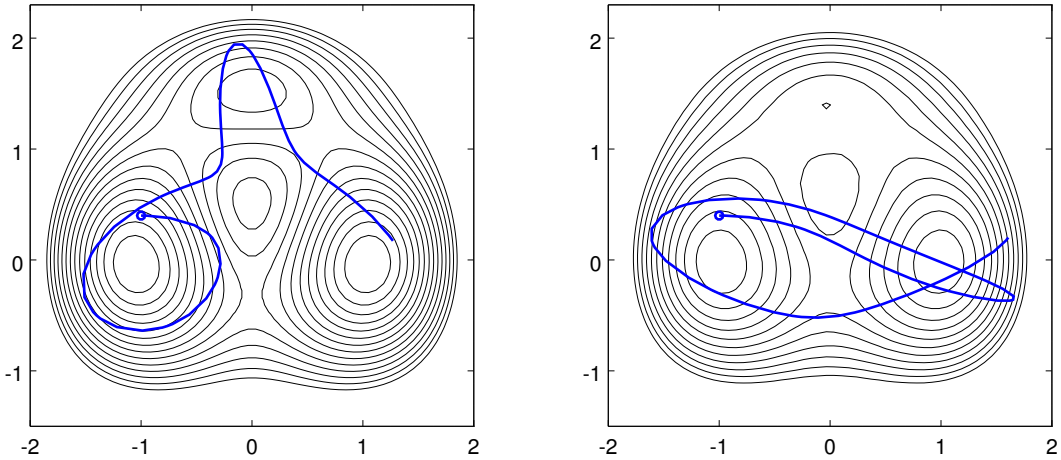

Figure 5: Potential and considered trajectory. *Left:* Original potential V from (4.2). *Right:* The smoothed potential $\bar{V}$ for $\lambda = 0.316$. The equipotential lines are at the same levels in both pictures.

as

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

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

When evaluating $\hat{V}$ with a requested local tolerance ϵ, the number of sampling points is doubled until $\|\sigma_m\| \leq \sqrt{m}\epsilon$. This defines a realization of $\hat{V}_\epsilon(x)$. 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.

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

$$\begin{aligned} V(x) =& 3\exp(-\|x - e_2\|^2) - 3\exp(-\|x - 5e_2\|^2) - 5\exp(-\|x - e_1\|^2) \\ & - 5\exp(-\|x + e_1\|^2) + (x_1^4 + (x_2 - 1/3)^4)/5, \quad \text{where } (e_i)_j = \delta_{ij}, \end{aligned} \tag{4.2}$$

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 Fig. 5 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.

Fig. 5 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 now easily 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

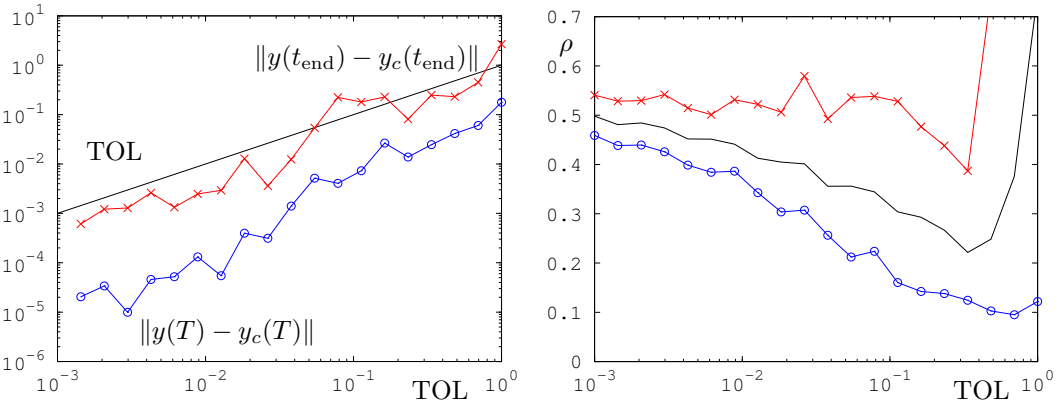

Figure 6: Numerical result averaged over 15 realizations. *Left:* Estimated error after the first time step of length T (circles) and at final time t_{end} (crosses) versus the requested step tolerance TOL. *Right:* Maximal, average, and minimal observed contraction factors ρ of the inexact SDC method in all time steps versus the requested step tolerance.

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 Fig. 6 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 (3.4) reflecting the worst case rather than the average case. Replacing the generously used triangle inequality by sharper bounds, however, would require to prescribe not only the magnitude of the evaluation error, but also restrict 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 Fig. 6, 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 local tolerances when evaluating right hand sides in SDC methods derived here allows significant savings in computational effort compared to a naive strategy. Effort reduction factors between 2 and 6 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, 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 to use larger local tolerances. On the other hand, the errors are highly correlated in several finite element computations. Consequently, the error in right hand side differences is small, which leads to different error propagation through the SDC iteration. Extending the approach to these settings is subject of further research.

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

References

[1] P. Alfeld. Fixed Point Iteration with Inexact Function Values. *Mathematics of Computation*, 38(157):87–98, 1982.

[2] M.P. Allen and D.J. Tildesley. *Computer simulation of liquids*. Oxford University Press, 1989.

[3] P. Amodio and L. Brugnano. A note on the efficient implementation of implicit methods for ODEs. *J. Comp. Appl. Math.*, 87:1–9, 1997.

[4] J. Barnes and P. Hut. A hierarchical o(n log n) force-calculation algorithm. *Nature*, 324(4):446–449, 1986.

[5] P. Birken. Termination criteria for inexact fixed-point schemes. *Numer. Lin. Algebra Appl.*, 22(4):702–716, 2015.

[6] J. Carrier, L. Greengard, and V. Rokhlin. A fast adaptive multipole algorithm for particle simulations. *SIAM J. Sci. Stat. Comput.*, 9(4):669–686, 1988.

[7] G.J. Cooper and J.C. Butcher. An iteration scheme for implicit Runge-Kutta methods. *IMA J Numer. Anal.*, 3:127–140, 1983.

[8] G.J. Cooper and R. Vignesvaran. A scheme for the implementation of implicit Runge-Kutta methods. *Computing*, 45, 1990.

[9] P. Deuflhard and F.A. Bornemann. *Scientific Computing with Ordinary Differential Equations*. Texts in Applied Mathematics, vol. 42. Springer, New York, 2nd edition, 2002.

[10] F.P.E. Dunne and D.R. Hayhurst. Efficient cycle jumping techniques for the modeling of materials and structures under cyclic mechanical and thermal loading. *Europ. J. Mech. A: Solids*, 13:639–660, 1994.

[11] A. Dutt, L. Greengard, and V. Rokhlin. Spectral deferred correction methods for ordinary differential equations. *BIT*, 40(2):241–266, 2000.

[12] B.V. Faleichik. Analytic iterative processes and numerical algorithms for stiff problems. *Comp. Meth. Appl. Math.*, 8(2):116–129, 2008.

[13] K. Frischmuth and D. Langemann. Numerical calculation of wear in mechanical systems. *Math. Comp. Sim.*, 81:2688–2701, 2011.

[14] E. Gallicchio, S.A. Egorov, and B.J. Berne. On the application of numerical analytic continuation methods for the study of quantum mechanical vibrational relaxation processes. *J. Chem. Phys.*, 109(18):7745–7755, 1998.

[15] S. Güttel and J.W. Pearson. A rational deferred correction approach to PDE-constrained optimization. Preprint, U Kent, 2016.

[16] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations. I: Nonstiff Problems*, volume 8 of *Springer Series in Computational Mathematics*. Springer, 2nd edition, 1993.

[17] J. Huang, J. Jia, and M.L. Minion. Accelerating the convergence of spectral deferred correction methods. *J. Comp. Phys.*, 214(2):633–656, 2006.

[18] L.O. Jay and T. Braconnier. A parallelizable preconditioner for the iterative solution of implicit runge–kutta-type methods. *J. Comp. Appl. Math.*, 111(1–2):63–76, 1999.

[19] J. Nocedal and S.J. Wright. *Numerical Optimization*. Springer, 1999.

[20] J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several Variables*. Academic Press, 1970.

[21] Ch. Schütte, A. Nielsen, and M. Weber. Markov state models and molecular alchemy. *Molecular Physics*, 113(1):69–78, 2015.

[22] R. Speck, D. Ruprecht, M. Minion, M. Emmett, and R. Krause. *Domain Decomposition Methods in Science and Engineering XXII*, chapter Inexact Spectral Deferred Corrections, pages 389–396. Springer, 2016.

[23] W. Tutschke. *Solution of Initial Value Problems in Classes of Generalized Analytic Functions*. Springer-Verlag, 1989.

[24] P.J. van der Houwen and J.J.B. de Swart. Triangularly implicit iteration methods for ODE-IVP solvers. *SIAM J. Sci. Comput.*, 18(1):41–55, 1997.

[25] M. Weiser. Faster SDC convergence on non-equidistant grids by DIRK sweeps. *BIT Numerical Mathematics*, 55(4):1219–1241, 2015.

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

A Uniqueness of work minimizer

Here we prove that for fixed local tolerance ϵ_0, the continuous relaxation of the work model with respect to the iteration count J is quasi-convex and thus has a unique minimizer.

Theorem A.1. *Let*

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

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

Proof. The derivative of W is

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

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

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

We obtain $q(J_{\min}) = (J_{\min}+1)d\rho^{J_{\min}}\log\rho < 0$ and $q(J) \to \mathrm{TOL}/\delta > 0$ 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 at exactly one zero and consequenctly W exactly one extremum. The sign of W' changes from negative to positive there, such that W has exactly one local minimizer. □