Remarks on the optimal convolution kernel for CSOR waveform relaxation

Min Hu[1,*], Ken Jackson[2,*], Jan Janssen[3,†] and Stefan Vandewalle[3,‡]

[1] *Comnetix Computer Systems Inc., 1440 Hurontario Street, Mississauga, Ontario, Canada L5G 3H4*
E-mail: `mhu@Comnetix.COM`

[2] *Department of Computer Science, University of Toronto, 10 King's College Road, Toronto, Ontario, Canada M5S 3G4*
E-mail: `krj@cs.toronto.edu`

[3] *Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200 A, B-3001 Heverlee, Belgium*
E-mail: `janj,stefan@cs.kuleuven.ac.be`

The convolution SOR waveform relaxation method is a numerical method for solving large-scale systems of ordinary differential equations on parallel computers. It is similar in spirit to the SOR acceleration method for solving linear systems of algebraic equations, but replaces the multiplication with an overrelaxation parameter by a convolution with a time-dependent overrelaxation function. Its convergence depends strongly on the particular choice of this function. In this paper, an analytic expression is presented for the optimal continuous-time convolution kernel and its relation to the optimal kernel for the discrete-time iteration is derived. We investigate whether this analytic expression can be used in actual computations. Also, the validity of the formulae that are currently used to determine the optimal continuous-time and discrete-time kernels is extended towards a larger class of ODE systems.

Keywords: convolution, iterative methods, parallel ODE solvers, successive overrelaxation, waveform relaxation

Subject classification: AMS(MOS) 65F10, 65L05

*This research was supported in part by the Natural Sciences and Engineering Research Council of Canada and the Information Technology Research Centre of Ontario.

†This research has been funded by the Research Fund K.U.Leuven (OT/94/16) and the Belgian National Fund for Scientific Research (N.F.W.O., project G.0235.96).

‡Postdoctoral Fellow of the Belgian National Fund for Scientific Research (N.F.W.O.).

1 Introduction

Waveform relaxation is an iterative method for numerically solving large-scale systems of ordinary differential equations (ODEs). The method is well suited for implementation on parallel computers, and high parallel efficiencies have been demonstrated for various applications, [6, 19, 22]. The convergence of the basic Jacobi and Gauss–Seidel waveform relaxation methods can be accelerated in several ways, such as by successive overrelaxation ([2, 3, 8, 14, 15, 19]), by Chebyshev iteration ([11, 20]), by Krylov subspace methods ([13]) and by multigrid techniques ([9, 10, 12, 21]).

Although waveform methods have been applied successfully to general, nonlinear, time-dependent coefficient problems, the convergence studies have concentrated on linear initial-value problems of the form

$$B\dot{u} + Au = f \ , \quad u(0) = u_0 \ , \tag{1}$$

with $B, A \in \mathbb{C}^{d\times d}$ and B nonsingular.

Recently, we have analysed the acceleration of the waveform method for (1) by successive overrelaxation (SOR) techniques, [8]. The first step of an SOR waveform relaxation algorithm consists of the computation of a Gauss–Seidel like iterate, $\hat{u}_i^{(\nu)}(t)$, which satisfies

$$\begin{aligned}\left(b_{ii}\frac{d}{dt} + a_{ii}\right)\hat{u}_i^{(\nu)}(t) \quad = \quad & -\sum_{j=1}^{i-1}\left(b_{ij}\frac{d}{dt} + a_{ij}\right)u_j^{(\nu)}(t) \\ & -\sum_{j=i+1}^{d_b}\left(b_{ij}\frac{d}{dt} + a_{ij}\right)u_j^{(\nu-1)}(t) + f_i(t) \ ,\end{aligned} \tag{2}$$

with $\hat{u}_i^{(\nu)}(0) = (u_0)_i$. We assume the matrices B and A to be partitioned into similar systems of $d_b \times d_b$ rectangular blocks b_{ij} and a_{ij} (in the pointwise case we have $d_b = d$, that is, b_{ij} and a_{ij} denote the matrix elements of B and A, respectively). In the second step, the old approximation $u_i^{(\nu-1)}(t)$ is updated to give the new iterate $u_i^{(\nu)}(t)$. In the standard SOR waveform scheme this involves the multiplication of the correction $\hat{u}_i^{(\nu)}(t) - u_i^{(\nu-1)}(t)$ by a scalar overrelaxation parameter ω, [8, eq. (2.2)]. In the convolution SOR (CSOR) waveform relaxation algorithm, the correction is convolved with a time-dependent kernel $\Omega(t)$,

$$u_i^{(\nu)}(t) = u_i^{(\nu-1)}(t) + \int_0^t \Omega(t-s)\left(\hat{u}_i^{(\nu)}(s) - u_i^{(\nu-1)}(s)\right) ds \ . \tag{3}$$

The success of the latter depends strongly on the particular choice of convolution kernel. The kernel that minimises the spectral radius of the corresponding operator, which will be referred to as the *optimal* kernel $\Omega_{opt}(t)$, can be determined for a certain class of ODE systems of the form (1).

In an implementation of the CSOR waveform relaxation method, the continuous-time algorithm is replaced by its discrete-time counterpart. This discrete-time CSOR algorithm is defined by applying a time-discretisation method to equation (2) and by replacing the convolution integral in (3) by a convolution sum using a discrete sequence Ω_τ. Again, the *optimal* convolution sequence $(\Omega_{opt})_\tau$ can be constructed for certain classes of problems and time discretisations.

With use of the optimal convolution kernel or sequence, the CSOR waveform relaxation method becomes vastly superior to the classical waveform methods. This superiority can be demonstrated quantitatively for the ODE system (1), obtained by finite-difference discretisation of the heat equation on a mesh with mesh-size h. We have shown (both theoretically and numerically) that for this problem CSOR attains an identical acceleration as the classical SOR method does for the linear system $Au = f$. This means that the asymptotic convergence factor of the CSOR waveform relaxation method behaves as $1 - O(h)$ for small h, while the spectral radii of the Jacobi, Gauss–Seidel and standard SOR waveform methods are all known to satisfy a formula of the form $1 - O(h^2)$, [8]. A very substantial improvement over the less sophisticated waveform schemes may therefore be expected.

In this paper, we will continue our exploration of the CSOR waveform relaxation method. The continuous-time and discrete-time techniques are recalled in §2, together with the Laplace- and Z-transform expressions of their respective optimal convolution kernels. In §3, we derive an explicit analytic expression of the optimal convolution kernel $\Omega_{opt}(t)$ for ODE systems of the form (1) with $B = I$. The connection between this optimal continuous-time kernel and the corresponding optimal discrete-time kernel is derived in §4. Whether this connection can be used in actual computations is investigated in §5, where the practical determination of a suitable discrete-time convolution kernel is discussed. Finally, the validity of the expressions for the Laplace transform and Z-transform of the optimal convolution kernels that are presented in §2, is extended towards a broader class of ODE systems in §6. Although most of our theoretical results are restricted to constant-coefficient linear problems, numerical evidence indicates a much wider applicability. Hence, we also comment on the robustness and applicability of the latter formulae.

So far, little experience has been gained with the CSOR method as a solver for practical real-life problems and many open questions on how to apply and implement the method in such cases remain unanswered. With this study we want to answer some of the questions that arise when the method is applied to model problems. We hope the insight gained from this study will prove to be useful when more difficult problems are addressed.

2 A review of CSOR waveform relaxation results

In this section, we will summarise some theoretical properties of the CSOR waveform relaxation method. For a more detailed study of the method, including proofs, references and a comparison with other waveform methods, we refer to [8].

2.1 The continuous-time case

The continuous-time CSOR waveform relaxation method, defined by (2) and (3), can be written formally as

$$u^{(\nu)}(t) = \mathcal{K}^{CSOR} u^{(\nu-1)}(t) + \varphi(t) \; ,$$

where $\mathcal{K}^{CSOR}$ is a linear operator consisting of a matrix multiplication and a Volterra convolution part. This operator is fully characterised by its symbol, $\mathbf{K}^{\mathbf{CSOR}}(z)$, which equals the Laplace transform of the operator's kernel. This symbol can be expressed in terms of $\tilde{\Omega}(z) = \mathcal{L}(\Omega(t))$, the Laplace transform of $\Omega(t)$, and the components of the matrix splittings $B = D_B - L_B - U_B$ and $A = D_A - L_A - U_A$, with D_B and D_A block diagonal, L_B and L_A block lower triangular and U_B and U_A block upper triangular matrices:

$$\mathbf{K}^{\mathbf{CSOR}}(z) = \left(z \left(\frac{1}{\tilde{\Omega}(z)} D_B - L_B \right) + \left(\frac{1}{\tilde{\Omega}(z)} D_A - L_A \right) \right)^{-1} \cdot \\ \left(z \left(\frac{1 - \tilde{\Omega}(z)}{\tilde{\Omega}(z)} D_B + U_B \right) + \left(\frac{1 - \tilde{\Omega}(z)}{\tilde{\Omega}(z)} D_A + U_A \right) \right) \; .$$

The spectral radius of $\mathcal{K}^{CSOR}$ is given below. A convolution kernel of the form

$$\Omega(t) = \omega \delta(t) + \omega_c(t) \; , \tag{4}$$

is assumed, with ω a scalar parameter and $\delta(t)$ the delta function.

Theorem 1 [8, Thm. 3.4]
Assume all eigenvalues of $D_B^{-1} D_A$ have positive real parts, and let $\Omega(t)$ be of the form (4) with $\omega_c(t) \in L_1(0, \infty)$. Then, $\mathcal{K}^{CSOR}$ is a bounded operator in $L_p(0, \infty)$, $1 \leq p \leq \infty$, and

$$\rho\left(\mathcal{K}^{CSOR}\right) = \sup_{\mathrm{Re}(z) \geq 0} \rho\left(\mathbf{K}^{\mathbf{CSOR}}(z)\right) = \sup_{\xi \in \mathbb{R}} \rho\left(\mathbf{K}^{\mathbf{CSOR}}(i\xi)\right) \; . \tag{5}$$

Remark 1
In Theorem 1, $\tilde{\Omega}(z)$ is required to be the Laplace transform of a function of the form (4) with $\omega_c(t) \in L_1(0, \infty)$. A sufficient (but not necessary) condition to

satisfy this requirement is that $\widetilde{\Omega}(z)$ is a bounded and analytic function in an open domain containing the closed right half of the complex plane.

The Laplace-transform expression of the optimal convolution kernel $\Omega_{opt}(t)$ depends on the location of the eigenvalues of the Jacobi symbol

$$\mathbf{K}^{\mathbf{JAC}}(z) = (zD_B + D_A)^{-1} \left(z \left(L_B + U_B \right) + \left(L_A + U_A \right) \right) \ . \tag{6}$$

Lemma 2 [8, Lemma 3.6]
Assume the matrices B and A are such that $zB + A$ is a block-consistently ordered matrix with nonsingular diagonal blocks. Assume the spectrum of $\mathbf{K}^{\mathbf{JAC}}(z)$ lies on a line segment $[-\mu_1(z), \mu_1(z)]$ with $\mu_1(z) \in \mathbb{C} \setminus \{(-\infty, -1] \cup [1, \infty)\}$. The spectral radius of $\mathbf{K}^{\mathbf{CSOR}}(z)$ is then minimised for a fixed z by the unique optimum $\widetilde{\Omega}_{opt}(z)$, given by

$$\widetilde{\Omega}_{opt}(z) = \frac{2}{1 + \sqrt{1 - \mu_1^2(z)}} \ , \tag{7}$$

where $\sqrt{\cdot}$ denotes the root with the positive real part. In particular,

$$\rho \left(\mathbf{K}^{\mathbf{CSOR},opt}(z) \right) = \left| \widetilde{\Omega}_{opt}(z) - 1 \right| < 1 \ . \tag{8}$$

2.2 The discrete-time case

In an actual implementation, the continuous-time method is replaced by a discrete-time method. As in [8], we will only deal with (irreducible, consistent, zero-stable) linear multistep formulae for time discretisation. We recall the general multistep formula for solving $\dot{y}(t) = f(t, y)$,

$$\frac{1}{\tau} \sum_{l=0}^{k} \alpha_l y[n+l] = \sum_{l=0}^{k} \beta_l f[n+l] \ .$$

Here, α_l and β_l are real constants, τ is the step size and $y[n]$ denotes the approximation of $y(t)$ at $t = n\tau$. We will assume that k starting values $y[0], y[1], \ldots, y[k-1]$ are available. The characteristic polynomials of the linear multistep method are given by $a(z) = \sum_{l=0}^{k} \alpha_l z^l$ and $b(z) = \sum_{l=0}^{k} \beta_l z^l$, and the stability region is denoted by S. We will also need the notion of a *strictly stable* multistep method, which is such that 1 is the only (simple) root of $a(z)$ on the unit circle, [5].

The first step of the discrete-time CSOR waveform relaxation algorithm is obtained by discretising (2) using a linear multistep method. The second step approximates the convolution integral in (3) by a convolution sum with kernel $\Omega_\tau = \{\Omega[n]\}_{n=0}^{N-1}$, where N denotes the (possibly infinite) number of time steps,

$$u_i^{(\nu)}[n] = u_i^{(\nu-1)}[n] + \sum_{l=0}^{n} \Omega[n-l] \left(\hat{u}_i^{(\nu)}[l] - u_i^{(\nu-1)}[l] \right) \ . \tag{9}$$

The discrete-time CSOR waveform relaxation operator $\mathcal{K}_\tau^{CSOR}$ is defined by rewriting the discrete-time version of (2) and (9) as $u_\tau^{(\nu)} = \mathcal{K}_\tau^{CSOR} u_\tau^{(\nu-1)} + \varphi_\tau$. Since $\mathcal{K}_\tau^{CSOR}$ is a discrete convolution operator, its discrete-time symbol $\mathbf{K}_\tau^{\mathbf{CSOR}}(z)$ is obtained by Z-transformation of the operator's kernel. More precisely,

$$\mathbf{K}_\tau^{\mathbf{CSOR}}(z) = \left(\frac{1}{\tau}\frac{a(z)}{b(z)}\left(\frac{1}{\widetilde{\Omega}_\tau(z)}D_B - L_B\right) + \left(\frac{1}{\widetilde{\Omega}_\tau(z)}D_A - L_A\right)\right)^{-1} \cdot$$
$$\left(\frac{1}{\tau}\frac{a(z)}{b(z)}\left(\frac{1-\widetilde{\Omega}_\tau(z)}{\widetilde{\Omega}_\tau(z)}D_B + U_B\right) + \left(\frac{1-\widetilde{\Omega}_\tau(z)}{\widetilde{\Omega}_\tau(z)}D_A + U_A\right)\right) ,$$

with $\widetilde{\Omega}_\tau(z) = \mathcal{Z}(\Omega_\tau)$ the Z-transform of the sequence Ω_τ.

The discrete-time equivalent to Theorem 1 is given next.

Theorem 3 [8, Thm. 4.4]
Assume $\sigma(-\tau D_B^{-1} D_A) \subset \mathrm{int} S$ and $\Omega_\tau \in l_1(\infty)$. Then, $\mathcal{K}_\tau^{CSOR}$ is a bounded operator in $l_p(\infty)$, $1 \le p \le \infty$, and

$$\rho\left(\mathcal{K}_\tau^{CSOR}\right) = \max_{|z|\ge 1} \rho\left(\mathbf{K}_\tau^{\mathbf{CSOR}}(z)\right) = \max_{|z|=1} \rho\left(\mathbf{K}_\tau^{\mathbf{CSOR}}(z)\right) . \tag{10}$$

Remark 2
In Theorem 3, $\widetilde{\Omega}_\tau(z)$ is required to be the Z-transform of an l_1-kernel Ω_τ. A sufficient (but not necessary) condition to satisfy this requirement is that $\widetilde{\Omega}_\tau(z)$ is a bounded and analytic function in an open domain containing $\{z \in \mathbb{C} \mid |z| \ge 1\}$.

Finally, we recall the discrete-time version of Lemma 2, which gives an expression of the Z-transform of the optimal convolution sequence $(\Omega_{opt})_\tau$. It depends on the eigenvalue distribution of the discrete-time Jacobi symbol, which is related to its continuous-time equivalent (6) by [10, eq. (4.10)],

$$\mathbf{K}_\tau^{\mathbf{JAC}}(z) = \mathbf{K}^{\mathbf{JAC}}\left(\frac{1}{\tau}\frac{a(z)}{b(z)}\right) . \tag{11}$$

Lemma 4 [8, Lemma 4.5]
Assume the matrices B and A are such that $\tau^{-1}a(z)/b(z)B + A$ is a block-consistently ordered matrix with nonsingular diagonal blocks. Assume the spectrum of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ lies on a line segment $[-(\mu_1)_\tau(z), (\mu_1)_\tau(z)]$ with $(\mu_1)_\tau(z) \in \mathbb{C} \setminus \{(-\infty,-1] \cup [1,\infty)\}$. The spectral radius of $\mathbf{K}_\tau^{\mathbf{CSOR}}(z)$ is then minimised for a fixed z by the unique optimum $(\widetilde{\Omega}_{opt})_\tau(z)$, given by

$$(\widetilde{\Omega}_{opt})_\tau(z) = \frac{2}{1+\sqrt{1-(\mu_1)_\tau^2(z)}} , \tag{12}$$

where $\sqrt{\cdot}$ *denotes the root with the positive real part. In particular,*

$$\rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(z)\right) = \left|(\tilde{\Omega}_{opt})_\tau(z) - 1\right| < 1 \ . \tag{13}$$

3 An explicit expression for the optimal convolution kernel

For systems of the form (1) that satisfy the assumptions of Lemma 2, the optimal convolution kernel $\Omega_{opt}(t)$ is obtained by inverse Laplace transforming the resulting expression for $\tilde{\Omega}_{opt}(z)$. For ODE systems (1) with $B = I$ however, we have the following explicit expression for the optimal kernel in the case of the point relaxation method.

Theorem 5
Consider (1) with $B = I$. Assume A is a consistently ordered matrix with constant positive diagonal $D_A = d_a I$ ($d_a > 0$) and the eigenvalues of $\mathbf{K}^{\mathbf{JAC}}(0)$ are real with $\mu_1 = \rho(\mathbf{K}^{\mathbf{JAC}}(0)) < 1$. Then, $\Omega_{opt}(t) = \delta(t) + (\omega_c)_{opt}(t)$ with $(\omega_c)_{opt}(t) \in L_1(0,\infty)$. In particular,

$$(\omega_c)_{opt}(t) = 2e^{-d_a t} I_2(\mu_1 d_a t)/t \ . \tag{14}$$

with $I_2(\cdot)$ the second-order modified Bessel function of the first kind.

Proof
From (6), we derive, under the assumptions of the theorem, that

$$\sigma\left(\mathbf{K}^{\mathbf{JAC}}(z)\right) = \frac{d_a}{z+d_a}\sigma\left(\mathbf{K}^{\mathbf{JAC}}(0)\right) \quad \text{and} \quad \mu_1(z) = \frac{d_a}{z+d_a}\mu_1 \ , \tag{15}$$

which implies that the eigenvalues of $\mathbf{K}^{\mathbf{JAC}}(z)$ with $\mathrm{Re}(z) \geq 0$ lie on a line segment $[-\mu_1(z), \mu_1(z)]$ with $\mu_1(z) \in \mathbb{C}\backslash\{(-\infty,-1]\cup[1,\infty)\}$. As a result, formula (7), which equals

$$\tilde{\Omega}_{opt}(z) = \frac{2}{1+\sqrt{1-\left(\frac{d_a}{z+d_a}\mu_1\right)^2}} \ , \tag{16}$$

is a bounded and analytic function for $\mathrm{Re}(z) \geq 0$. Remark 1 then yields that $\Omega_{opt}(t)$ is of the form (4) with $\omega_c(t) \in L_1(0,\infty)$, while $\omega = \lim_{z\to\infty}\tilde{\Omega}_{opt}(z) = 1$ by [8, Prop. 3.2].

The correctness of the analytic expression for $\Omega_{opt}(t)$ or $(\omega_c)_{opt}(t)$ can be checked as an elementary exercise by using the Laplace-transform pairs $\mathcal{L}(\delta(t)) = 1$ and

$$\mathcal{L}\left(2e^{-(a+b)t/2} I_2(\frac{a-b}{2}t)/t\right) = \frac{(a-b)^2}{(\sqrt{z+a}+\sqrt{z+b})^4} \ ,$$

the latter of which can be found in [1, eq. (29.3.53)]. □

The shape of the function $(\omega_c)_{opt}(t)$ is characterised by the properties given in the following lemma.

Lemma 6
Under the assumptions of Theorem 5, $(\omega_c)_{opt}(t)$, given by (14), satisfies the following properties:

1. $(\omega_c)_{opt}(0) = 0$
2. $0 \le (\omega_c)_{opt}(t) < \mu_1 d_a e^{-(1-\mu_1)d_a t}$
3. $\frac{\mu_1^2}{4e} d_a < \max_{t \ge 0} (\omega_c)_{opt}(t) < \mu_1 d_a$
4. $\int_0^\infty (\omega_c)_{opt}(t)\, dt = \frac{\mu_1^2}{(1+\sqrt{1-\mu_1^2})^2}$.

Proof
A series expression for the modified Bessel function $I_2(t)$ can be found in [1, eq. (9.6.10)]. It reads

$$I_2(t) = \left(\frac{t}{2}\right)^2 \sum_{k=0}^{\infty} \frac{(\frac{t^2}{4})^k}{k!(k+2)!} ,$$

from which we derive

$$(\omega_c)_{opt}(t) = \mu_1 d_a\, e^{-d_a t} \sum_{k=0}^{\infty} \frac{(\mu_1 d_a)^{2k+1}}{2^{2k+1} k!(k+2)!} t^{2k+1} . \tag{17}$$

Property 1 and the positivity of $(\omega_c)_{opt}(t)$ result. Formula (17) can be written as

$$(\omega_c)_{opt}(t) = \mu_1 d_a\, e^{-d_a t} \sum_{k=0}^{\infty} \frac{(2k+1)!}{2^{2k+1} k!(k+2)!} \frac{(\mu_1 d_a t)^{2k+1}}{(2k+1)!} .$$

Since the coefficients $(2k+1)!/(2^{2k+1}k!(k+2)!)$ are (strictly) smaller than 1 for $k \ge 0$, we have

$$(\omega_c)_{opt}(t) < \mu_1 d_a\, e^{-d_a t}\, e^{\mu_1 d_a t} = \mu_1 d_a\, e^{-(1-\mu_1)d_a t} , \tag{18}$$

which proves the upper bound in Property 2. We now truncate series (17) after the first term to get

$$\mu_1 d_a\, e^{-d_a t} \frac{\mu_1 d_a}{4} t \le (\omega_c)_{opt}(t) , \tag{19}$$

with equality only for $t = 0$. Calculation of the maxima of the upper and lower bounds (18) and (19) over $t \ge 0$ leads to Property 3. Finally, Property 4 follows immediately from [8, Prop. 3.2]. □

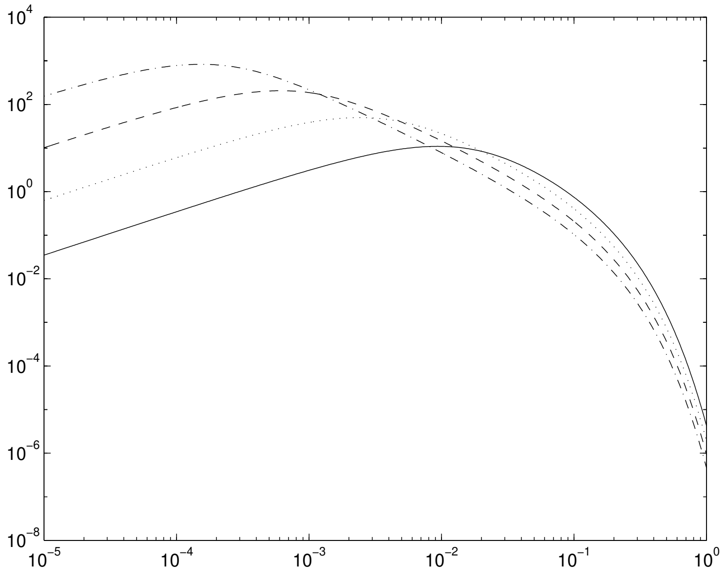

Figure 1: $(\omega_c)_{opt}(t)$ vs. t for the one-dimensional heat equation (20) with finite-difference discretisation on a mesh with mesh size $h = 1/8$ (solid), $h = 1/16$ (dotted), $h = 1/32$ (dashed) and $h = 1/64$ (dash-dotted).

When the system of ODEs is derived by semi discretisation of a parabolic partial differential equation, μ_1 is often close to one. The characteristics of the optimal kernel are then largely determined by the parameter d_a, whose value is often rapidly increasing with decreasing mesh spacing. In that case, $(\omega_c)_{opt}(t)$ is a positive function, which starts from 0 at $t = 0$ and has an area that is bounded by 1. Its maximum is proportional to d_a, hence it is large for small h, while the function decreases exponentially for sufficiently large t. As an example, we will illustrate these implications of Lemma 6 for the one-dimensional heat equation on the unit interval,

$$\frac{\partial \mathbf{u}(x,t)}{\partial t} - \frac{\partial^2 \mathbf{u}(x,t)}{\partial x^2} = 0 \ , \quad x \in [0,1] \ , \quad t > 0 \ , \tag{20}$$

discretised using finite differences on a mesh $\{x_i = ih \mid 0 \leq i \leq 1/h\}$ with mesh size h. The resulting ODE system (1), with $B = I$, $d_a = 2/h^2$ and $\mu_1 = \cos(\pi h)$, satisfies the conditions of Theorem 5. Figure 1 shows a logarithmic plot of $(\omega_c)_{opt}(t)$ for $t \in [10^{-5}, 1]$ and for several values of the mesh size h. Note that its maximum increases and is attained at a smaller t-value for decreasing h, while, for sufficiently large t, the value of the optimal kernel rapidly approaches 0. Consequently, we may expect the use of a truncated kernel $\Omega(t)$, defined by

$\Omega_{opt}(t)$ for $t \le T$ and by 0 for $t > T$ for some large enough T, to lead to nearly optimal convergence results.

4 The relation between the continuous-time and discrete-time optimal kernels

By inserting (4) into (3) we find that

$$u_i^{(\nu)}(t) = u_i^{(\nu-1)}(t)+\omega\left(\hat{u}_i^{(\nu)}(t) - u_i^{(\nu-1)}(t)\right) + \int_0^t \omega_c(t-s)\left(\hat{u}_i^{(\nu)}(s) - u_i^{(\nu-1)}(s)\right) ds\ . \tag{21}$$

The corresponding step in the discrete-time iteration is derived as follows. We set $\Omega_\tau = \omega\delta_\tau + (\omega_c)_\tau$ with $\delta_\tau = \{1, 0, 0, \ldots\}$ the discrete delta function and insert this expression into (9). This yields

$$u_i^{(\nu)}[n] = u_i^{(\nu-1)}[n] + \omega\left(\hat{u}_i^{(\nu)}[n] - u_i^{(\nu-1)}[n]\right) + \sum_{l=0}^{n}\omega_c[n-l]\left(\hat{u}_i^{(\nu)}[l] - u_i^{(\nu-1)}[l]\right)\ . \tag{22}$$

In this section we will relate the optimal continuous-time and discrete-time convolution kernels. Comparing (21) to (22) already suggests that $\omega_c[n]$ should be such that it approximates $\tau\omega_c(n\tau)$ for small τ. In that case the discrete convolution sum approximates the continuous convolution integral as a simple numerical integration rule. This intuition is confirmed and cast into a more precise mathematical form in the following theorem.

Theorem 7
Consider (1) with $B = I$. Assume A is a consistently ordered matrix with constant positive diagonal $D_A = d_a I$ ($d_a > 0$), the eigenvalues of $\mathbf{K}^{\mathbf{JAC}}(0)$ are real with $\mu_1 = \rho(\mathbf{K}^{\mathbf{JAC}}(0)) < 1$, and the linear multistep method is strictly stable. Then, the continuous-time optimal kernel $\Omega_{opt}(t) = \delta(t) + (\omega_c)_{opt}(t)$ and its discrete-time equivalent $(\Omega_{opt})_\tau = \delta_\tau + ((\omega_c)_{opt})_\tau$ are related by

$$\lim_{\substack{\tau\to 0\\ (t=n\tau)}} \frac{((\omega_c)_{opt})_\tau[n]}{\tau} = (\omega_c)_{opt}(t)\ ,\quad t \ge 0\ . \tag{23}$$

Note that we have used a subscript τ in the notation of the optimal discrete kernel to emphasise that the function depends on the value of the time increment. An equivalent but somewhat less intuitive form of (23) is obtained by replacing τ by t/n:

$$\lim_{n\to\infty} \frac{n((\omega_c)_{opt})_{\frac{t}{n}}[n]}{t} = (\omega_c)_{opt}(t)\ ,\quad t > 0\ . \tag{24}$$

Proof

Under the assumptions of the theorem, Lemma 2 holds for $\mathrm{Re}(z) \geq 0$ with $\mu_1(z)$ given in (15). The function $\widetilde{\Omega}_{opt}(z)$, given by (16), is bounded and analytic for $\mathrm{Re}(z) \geq 0$. Consequently, by the inverse Laplace-transform formula, we have

$$(\omega_c)_{opt}(t) = \frac{1}{2\pi i} \int_{-i\infty}^{i\infty} e^{zt} \left(\widetilde{\Omega}_{opt}(z) - 1 \right) dz \ . \tag{25}$$

A similar expression will be derived for the discrete-time kernel by using the inverse Z-transform formula. To apply Lemma 4, we have to ensure that

$$(\mu_1)_\tau(z) \in \mathbb{C} \setminus \{(-\infty, -1] \cup [1, \infty)\} \ , \quad \forall \ |z| \geq 1 \ , \tag{26}$$

with $(\mu_1)_\tau(z)$ calculated from (11) and (15), i.e.,

$$(\mu_1)_\tau(z) = \frac{d_a}{\frac{1}{\tau}\frac{a(z)}{b(z)} + d_a} \, \mu_1 \ . \tag{27}$$

Because of the strict stability of the multistep method at least a small disk of the form $\{\eta : |\eta + d| \leq d\}$ with $d > 0$ is contained in the stability region S, [5, p. 259]. Consequently, we have for small enough τ that $\{\eta : |\eta + d_a| \leq d_a\} \subset \tau^{-1} S$. Since, by definition of stability region, $\{\tau^{-1} a(z)/b(z) \mid |z| \geq 1\} = \bar{\mathbb{C}} \setminus \tau^{-1}\mathrm{int}S$, we immediately obtain $|\tau^{-1} a(z)/b(z) + d_a| \geq d_a$ for $|z| \geq 1$. For these values of z, (27) yields $|(\mu_1)_\tau(z)| < 1$, and, hence, (26). From this we may conclude that for small enough τ, the conditions of Lemma 4 are satisfied for all z on or outside the unit disk. Therefore, for any such z the optimal $(\widetilde{\Omega}_{opt})_\tau(z)$ is given by the combination of (12) and (27) . This function is bounded and analytic for $|z| \geq 1$; by using the inverse Z-transform formula, [4, p. 262], we arrive at the expression

$$((\omega_c)_{opt})_\tau[n] = \frac{1}{2\pi i} \oint_{|z|=1} z^{n-1} \left((\widetilde{\Omega}_{opt})_\tau(z) - 1 \right) dz \ . \tag{28}$$

As we have derived the conditions for existence of the optimal kernels, we can now prove the correctness of (23). We start by considering the case of $t = 0$. In that case we can use Property 1 of Lemma 6. Hence, we need to show that

$$\lim_{\tau \to 0} \frac{((\omega_c)_{opt})_\tau[0]}{\tau} = (\omega_c)_{opt}(0) = 0 \ . \tag{29}$$

By the initial-value theorem for the Z-transform, [16, Eq. (7.35)], we find

$$((\omega_c)_{opt})_\tau[0] = \lim_{z\to\infty} \left((\widetilde{\Omega}_{opt})_\tau(z) - 1 \right) = \frac{2}{1 + \sqrt{1 - \left(\lim_{z\to\infty} (\mu_1)_\tau(z) \right)^2}} - 1 \ .$$

The limit in this expression can be calculated from (27),

$$\lim_{z\to\infty}(\mu_1)_\tau(z) = \lim_{z\to\infty}\frac{d_a}{\frac{1}{\tau}\frac{a(z)}{b(z)}+d_a}\mu_1 = \frac{d_a\mu_1}{\frac{1}{\tau}\frac{\alpha_k}{\beta_k}+d_a} \ .$$

Equality (29) follows by a straightforward limit calculation.

Next, we will prove (23) for $t > 0$. By a change of variables $z = i\theta$ in (25) and and $z = e^{i\tau\theta}$ in (28), we obtain respectively

$$(\omega_c)_{opt}(t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\theta}\left(\widetilde{\Omega}_{opt}(i\theta) - 1\right) d\theta \tag{30}$$

and

$$((\omega_c)_{opt})_\tau[n] = \frac{1}{2\pi}\tau\int_{-\frac{\pi}{\tau}}^{\frac{\pi}{\tau}} e^{in\tau\theta}\left((\widetilde{\Omega}_{opt})_\tau(e^{i\tau\theta}) - 1\right) d\theta \ . \tag{31}$$

Consider now a fixed $t > 0$ with $t = n\tau$. Switching to the notation of (24), expression (31) can be transformed into

$$\frac{n\left((\omega_c)_{opt}\right)_{\frac{t}{n}}[n]}{t} = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it\theta}\left((\widetilde{\Omega}_{opt})_{\frac{t}{n}}(e^{i\frac{t}{n}\theta}) - 1\right)\chi_{[-\frac{n}{t}\pi,\frac{n}{t}\pi]}(\theta)\, d\theta \ , \tag{32}$$

where the characteristic function $\chi_{[-\frac{n}{t}\pi,\frac{n}{t}\pi]}(\theta)$ equals 1 for $\theta \in [-\frac{n}{t}\pi, \frac{n}{t}\pi]$ and 0 elsewhere. As before, (32) holds only if τ is small enough. For a fixed t this is equivalent to requiring n to be large enough, say $n \geq N$.

The limit relation (24) follows immediately from (30) and (32) by the dominated convergence theorem, [17, Thm. I.16], if we can prove the pointwise convergence

$$\lim_{n\to\infty} e^{it\theta}\left((\widetilde{\Omega}_{opt})_{\frac{t}{n}}(e^{i\frac{t}{n}\theta}) - 1\right)\chi_{[-\frac{n}{t}\pi,\frac{n}{t}\pi]}(\theta) = e^{it\theta}\left(\widetilde{\Omega}_{opt}(i\theta) - 1\right) \tag{33}$$

and the uniform, n-independent bound

$$\left|e^{it\theta}\left((\widetilde{\Omega}_{opt})_{\frac{t}{n}}(e^{i\frac{t}{n}\theta}) - 1\right)\chi_{[-\frac{n}{t}\pi,\frac{n}{t}\pi]}(\theta)\right| \leq g(\theta) \ , \quad n \geq N \ , \tag{34}$$

with $g(\theta) \in L_1(-\infty, \infty)$.

The equality in (33) follows from the consistency of the linear multistep method. Indeed, from $a(1) = 0$ and $a'(1) = b(1)$, we derive

$$\lim_{n\to\infty}\frac{n}{t}\frac{a(e^{i\frac{t}{n}\theta})}{b(e^{i\frac{t}{n}\theta})} = \lim_{n\to\infty}\frac{a(e^{i\frac{t}{n}\theta})}{\frac{t}{n}b(e^{i\frac{t}{n}\theta})} = i\theta \ ,$$

and thus,

$$\lim_{n\to\infty}(\widetilde{\Omega}_{opt})_{\frac{t}{n}}(e^{i\frac{t}{n}\theta}) = \lim_{n\to\infty}\widetilde{\Omega}_{opt}\left(\frac{n}{t}\frac{a(e^{i\frac{t}{n}\theta})}{b(e^{i\frac{t}{n}\theta})}\right) = \widetilde{\Omega}_{opt}(i\theta) \ .$$

In order to prove condition (34) we will construct a function $g(\theta)$ explicitly. Because of the strict stability requirement, 1 is the only root of $a(z)$ on the unit circle. Since it is also the only root of the rational function $a(z)/b(z)$ on the unit circle and since this root is simple, there exists a finite positive constant M such that

$$\left|\frac{a(e^{i\theta})}{b(e^{i\theta})}\right| \geq \frac{|\theta|}{M}\ , \quad \theta \in [-\pi,\pi] \quad \text{or} \quad \left|\frac{n}{t}\frac{a(e^{i\frac{t}{n}\theta})}{b(e^{i\frac{t}{n}\theta})}\right| \geq \frac{|\theta|}{M}\ , \quad \theta \in [-\frac{n}{t}\pi, \frac{n}{t}\pi]\ . \quad (35)$$

To bound the left-hand side of (34) we note that

$$\left|(\widetilde{\Omega}_{opt})_\tau(z) - 1\right| = \frac{|(\mu_1)_\tau(z)|^2}{\left|1+\sqrt{1-(\mu_1)^2_\tau(z)}\right|^2}\ .$$

Since $\sqrt{\cdot}$ denotes the root with the positive real part, (27) yields

$$\left|(\widetilde{\Omega}_{opt})_{\frac{t}{n}}(e^{i\frac{t}{n}\theta}) - 1\right| \leq \left|(\mu_1)^2_{\frac{t}{n}}(e^{i\frac{t}{n}\theta})\right| = \frac{(\mu_1 d_a)^2}{\left|\frac{n}{t}\frac{a(e^{i\frac{t}{n}\theta})}{b(e^{i\frac{t}{n}\theta})} + d_a\right|^2} \leq \frac{(\mu_1 d_a)^2}{\left|\left|\frac{n}{t}\frac{a(e^{i\frac{t}{n}\theta})}{b(e^{i\frac{t}{n}\theta})}\right| - d_a\right|^2}\ .$$

We can now use (35) to construct the following bound, valid for $|\theta| > M d_a$,

$$\left|e^{it\theta}\left((\widetilde{\Omega}_{opt})_{\frac{t}{n}}(e^{i\frac{t}{n}\theta}) - 1\right)\chi_{[-\frac{n}{t}\pi,\frac{n}{t}\pi]}(\theta)\right| \leq \frac{(\mu_1 d_a)^2}{\left|\frac{|\theta|}{M} - d_a\right|^2}\ .$$

This bound holds even if $\theta \notin [-\frac{n}{t}\pi, \frac{n}{t}\pi]$ because of the presence of the χ-function. Note finally from (13) that the left-hand side of (34) is always bounded by 1. The proof is then completed by setting $g(\theta)$ to be the $L_1(-\infty,\infty)$-function

$$g(\theta) = \begin{cases} 1 & \theta \in [-L, L] \\ \frac{(\mu_1 d_a)^2}{\left(\frac{|\theta|}{M} - d_a\right)^2} & \theta \notin [-L, L] \end{cases}\ ,$$

with $L > M d_a$. □

Remark 3

The strict stability condition is a very natural condition. In [5, p. 272] we find that it is satisfied by any multistep method of practical interest with nonempty int S. *However, methods that do not satisfy the condition on the stability region do exist – Nyström methods, for example, [5, p. 262]. For such methods* $(\widetilde{\Omega}_{opt})_\tau(z)$ *from (12) is not analytic for* $|z| \geq 1$, *and the inverse Z-transform calculation is not feasible.*

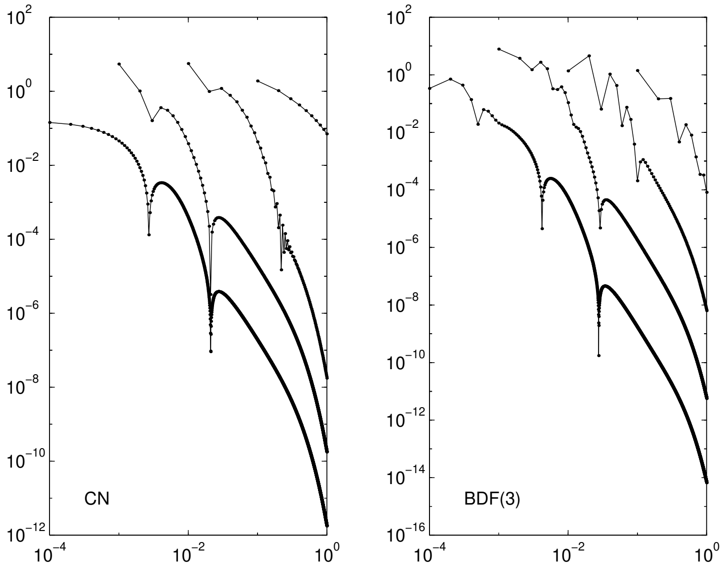

Figure 2: Absolute value of (36) vs. $t = n\tau$ for the one-dimensional heat equation (20) with finite-difference discretisation on a mesh with mesh size $h = 1/16$ and the CN and BDF(3) method with (from top to bottom) $\tau = 1/10$, $\tau = 1/100$, $\tau = 1/1000$ and $\tau = 1/10000$.

Remark 4

For strictly stable multistep methods the optimal discrete kernel was only proved to exist for small enough τ. This condition on τ was required in the proof of (26), i.e., to guarantee the analyticity of (12). For $A(\alpha)$-stable methods, however, condition (26) is satisfied irrespective of the size of the time increment. This can be explained by noting that in this case $(-\infty, 0) \subset \text{int}S$, which implies that $\tau^{-1}a(z)/b(z) + d_a$ with $|z| \geq 1$ is either complex or real with absolute value larger than or equal to d_a. Hence, for these methods the optimal kernel exists for any τ if the other assumptions of the theorem are satisfied.

We will illustrate Theorem 7 for model problem (20), discretised using finite differences with $h = 1/16$. To show the convergence of the discrete-time kernel to the continuous-time one with decreasing time increment, we have plotted the absolute value of the difference

$$\frac{((\omega_c)_{opt})_\tau[n]}{\tau} - (\omega_c)_{opt}(n\tau) \tag{36}$$

in Figure 2 for several values of τ and $t = n\tau \in [10^{-4}, 1]$. We used the Crank–

Table 1: $\log\left(\frac{((\omega_c)_{opt})_\tau[0]}{\tau}\right)$ for the one-dimensional heat equation (20) with finite-difference discretisation on a mesh with mesh size $h = 1/16$.

τ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
CN	0.702	1.231	1.008	0.176	-0.805	-1.803
BDF(3)	0.710	1.261	1.069	0.249	-0.729	-1.727

Nicolson (CN) method and the third-order backward differentiation (BDF(3)) formula for time discretisation and approximated the discrete kernel from (12) by an inverse Z-transform algorithm based on the use of FFT's, as explained in [8, §5.4]. The downward peaks are due to the zero crossing of (36).

To illustrate the convergence at $t = 0$, we report values of $\log\left(\frac{((\omega_c)_{opt})_\tau[0]}{\tau}\right)$ in Table 1. Note that the convergence to the limiting value $-\infty$ is very slow.

5 Practical determination of a suitable convolution kernel

In this section, we comment on the practical determination of a suitable convolution sequence Ω_τ for the discrete-time CSOR waveform relaxation algorithm.

We will first consider ODE systems of the form $\dot{u} + Au = f$ for which the assumptions of Theorem 5 are satisfied. For such problems, we have an explicit expression for the optimal continuous-time kernel $\Omega_{opt}(t)$, which is completely determined by the scalar μ_1 and the diagonal value d_a. Unfortunately, a similar expression does not seem to exist for the optimal discrete-time sequence $(\Omega_{opt})_\tau$, as the required inverse Z-transform appears in general to be too complex to be performed analytically.

One might, at first, try to employ the continuous-time kernel in the discrete-time computations. This idea is inspired by the existence of the limit relation (23). More precisely, one could set $\Omega_\tau = \delta_\tau + (\omega_c)_\tau$, and select

$$\omega_c[n] = \tau(\omega_c)_{opt}(n\tau) \ , \quad n = 0, \ldots, N-1 \ . \tag{37}$$

Experimental convergence factors for the model problem obtained with this discrete kernel and the CN time-discretisation method are given in Table 2. They are unsatisfactory, except when very small time steps τ are used. Another attempt at using the continuous-time kernel could be based on the observation that one is not really interested in determining the value of the kernel but in computing an integral. In particular, the discrete convolution sum in (22) can be regarded as a numerical approximation by quadrature of the convolution integral in (21).

Table 2: Averaged convergence factors for the one-dimensional heat equation (20) with finite-difference discretisation and the CN method, using (37) and (in parentheses) (38)–(39) to approximate the optimal kernel or convolution integral.

τ, h	1/8	1/16	1/32	1/64
1/100	0.543 (0.448)	0.885 (0.690)	0.982 (0.961)	0.996 (0.995)
1/500	0.461 (0.448)	0.745 (0.671)	0.952 (0.851)	0.994 (0.984)
1/1000	0.455 (0.452)	0.701 (0.667)	0.913 (0.837)	0.991 (0.948)

Table 3: Averaged convergence factors for the one-dimensional heat equation (20) with finite-difference discretisation and the CN method, using an inverse Z-transform technique to approximate the optimal kernel.

τ, h	1/8	1/16	1/32	1/64
1/100	0.441	0.676	0.820	0.907
1/500	0.441	0.676	0.816	0.902
1/1000	0.441	0.676	0.816	0.902

Hence, instead of using the first order quadrature rule that one gets when one uses (37), one could try to compute that integral more accurately by using an integration rule of higher order. In a second experiment, we used the composite midpoint integration rule,

$$\sum_{l=0}^{n-1} \omega_c[n-l] \cdot \left(\frac{\hat{u}_i^{(\nu)}[l] + \hat{u}_i^{(\nu)}[l+1]}{2} - \frac{u_i^{(\nu-1)}[l] + u_i^{(\nu-1)}[l+1]}{2} \right) , \tag{38}$$

where the fractions denote linearly interpolated approximations of $\hat{u}_i^{(\nu)}((l+1/2)\tau)$ and $u_i^{(\nu-1)}((l+1/2)\tau)$ respectively, and

$$\omega_c[n] = \tau(\omega_c)_{opt}((n-1/2)\tau) , \quad n = 1, \ldots, N-1 . \tag{39}$$

The corresponding convergence factors given in Table 2 in parentheses are somewhat better than the ones obtained by using (37), but overall they do not convince. Other numerical integration rules lead to similar conclusions. This follows from the discussion in the previous section and from the observation that the optimal discrete kernel for a particular problem and time-discretisation method can be very different from the optimal continuous kernel unless τ is very small.

Hence we will now consider methods that derive the optimal discrete kernel directly, based on the analytical expression of its Z-transform, which is given by the combination of (12) and (27),

$$(\widetilde{\Omega}_{opt})_\tau(z) = \frac{2}{1+\sqrt{1-\left(\frac{d_a}{\frac{1}{\tau}\frac{a(z)}{b(z)}+d_a}\mu_1\right)^2}} \ .$$

The inverse Z-transform can be computed symbolically by a series expansion of this expression in terms of powers of z^{-1} from which elements of the sequence $\Omega_{opt}[n]$ can easily be derived. A more practical procedure is to use a numerical inverse Z-transform technique. Note that $\Omega_{opt}[n]$ equals the n-th Fourier coefficient of the 2π-periodic function $(\widetilde{\Omega}_{opt})_\tau(e^{-it})$. Thus, the value $\Omega_{opt}[n]$ can be approximated by the n-th element of the discrete Fourier transform of the sequence $\left\{(\widetilde{\Omega}_{opt})_\tau\left(e^{-i2\pi k/M}\right)\right\}_{k=0}^{M-1}$ for some $M \geq N$ (M is usually selected to be larger than N to anticipate certain aliasing effects), [8, §5.4]. The approach is illustrated for the model problem with CN time discretisation in Table 3. We observe that the experimental convergence factors are independent of the time increment τ. More precisely, they are almost identical to the optimal continuous-time spectral radii $\rho(\mathcal{K}^{CSOR,opt})$, which are proved to behave as $1-2\pi h$ for small h, [8, §5.1].

Next, we consider the case of problems for which $(\mu_1)_\tau(z)$ is not known analytically. For such systems, we cannot compute the analytic expression of $(\widetilde{\Omega}_{opt})_\tau(z)$ explicitly. The numerical inverse Z-transform method is in theory still applicable, but it will be very time-consuming, since it now requires the computation of the value of $(\mu_1)_\tau(z)$ for M equidistant points around the unit circle. For use in this situation, an automatic procedure was developed by Reichelt et al. for computing an analytic approximation to $(\widetilde{\Omega}_{opt})_\tau(z)$, [18, §5.6] and [19, §6]. The largest-magnitude eigenvalue $(\mu_1)_\tau(z)$ for some specific values of z (e.g. $z = 1, -1, \infty, \ldots$) are estimated by subspace iteration or the implicitly restarted Arnoldi method, and inserted in (12). Next, these computed values of $(\widetilde{\Omega}_{opt})_\tau(z)$ are fitted by a ratio of low-order polynomials in z^{-1}:

$$(\widetilde{\Omega}_{opt})_\tau(z) \approx \frac{\sum_{j=0}^{K} b_j z^{-j}}{\sum_{j=0}^{L} a_j z^{-j}} = \sum_{j=1}^{L} \frac{c_j}{1-r_j z^{-1}} \ .$$

The inverse Z-transform of this approximation yields

$$\Omega_{opt}[n] \approx \sum_{j=1}^{L} c_j r_j^n \ .$$

It is as yet unclear how to select the specific values of z and what other conditions are to be imposed on the rational approximation (e.g. as to the degree of numerator and denominator, the number of poles, and the pole placement). The procedure is illustrated for certain nonlinear semi-conductor device problems in [19], and is shown to lead to very satisfactory results, even for systems of ODEs that do not satisfy the assumptions of Theorem 4. We will further comment on this in the next section, where we discuss the robustness of formulae (7) and (12).

6 An extension of the optimal CSOR theory

So far, the applicability of Lemmas 2 and 4 is restricted to problems whose Jacobi symbols have collinear spectra. In this section we formulate analogous results for more general problems. We will limit the discussion to the discrete-time case. The continuous-time case can be treated similarly.

Lemma 4 was proved by applying a classical SOR result for complex matrices to the linear system $(\tau^{-1}a(z)/b(z)B+A)u = f$. It was noted in [8] that the CSOR symbol $\mathbf{K}_\tau^{\mathbf{CSOR}}(z)$ represents the SOR iteration matrix for the latter system, with $\widetilde{\Omega}_\tau(z)$ acting as the complex overrelaxation parameter. Since the coefficient matrix of the linear system is assumed to be block-consistently ordered, the eigenvalues of the SOR iteration matrix, $\lambda_\tau(z)$, are related to the eigenvalues $\mu_\tau(z)$ of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ by the Young-relation, [23, Thm. 14-3.4],

$$\lambda_\tau(z) + \widetilde{\Omega}_\tau(z) - 1 = \sqrt{\lambda_\tau(z)}\widetilde{\Omega}_\tau(z)\mu_\tau(z) \ . \tag{40}$$

This implies that the spectral radius $\rho(\mathbf{K}_\tau^{\mathbf{CSOR}}(z))$ for a given $\widetilde{\Omega}_\tau(z)$ equals

$$\max_{\mu_\tau(z)\in\mathbf{K}_\tau^{\mathbf{JAC}}(z)} \left\{ |\lambda_\tau(z)| : \lambda_\tau(z) + \widetilde{\Omega}_\tau(z) - 1 = \sqrt{\lambda_\tau(z)}\widetilde{\Omega}_\tau(z)\mu_\tau(z) \right\} \ . \tag{41}$$

When the eigenvalues of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ are on a line segment in the complex plane, classical SOR theory provides a simple expression for the $\widetilde{\Omega}_\tau(z)$ that minimises (41). This optimal value is denoted by $(\widetilde{\Omega}_{opt})_\tau(z)$ and given by (12). If the collinearity assumption is not satisfied, however, one cannot find an optimal $\widetilde{\Omega}_\tau(z)$ easily, and a more *complex* SOR theory may have to be used. Such a theory was recently developed by Hu, Jackson and Zhu, [7]. They assume the eigenvalues of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ to lie in a region $R(p_\tau(z), q_\tau(z), \phi_\tau(z))$, the closed interior of an ellipse centred around the origin. This ellipse is given by

$$E(p_\tau(z), q_\tau(z), \phi_\tau(z)) = \left\{ \mu : \mu = e^{i\phi_\tau(z)} \left(p_\tau(z)\cos(\theta) + iq_\tau(z)\sin(\theta) \right) \right\} \ ,$$

with semi-axes $p_\tau(z)$ and $q_\tau(z)$ that satisfy $p_\tau(z) \geq q_\tau(z) \geq 0$, angle $\phi_\tau(z)$ with $-\pi/2 \leq \phi_\tau(z) \leq \pi/2$, and θ varying between 0 and 2π. This is illustrated graphically in Figure 3. Obviously, the spectral radius $\rho(\mathbf{K}_\tau^{\mathbf{CSOR}}(z))$, given by (41), is bounded from above by the value $r_\tau(z)$ which is defined in terms of the current

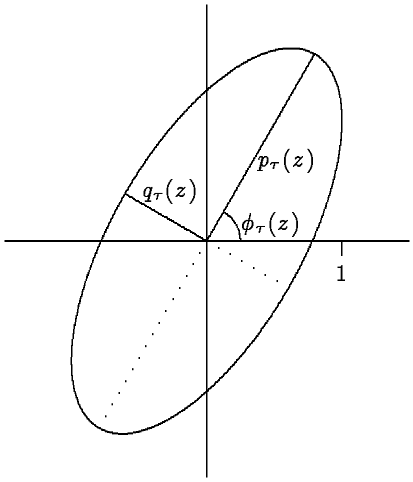

Figure 3: An ellipse $E(p_\tau(z), q_\tau(z), \phi_\tau(z))$.

$\widetilde{\Omega}_\tau(z)$ as

$$\max_{\mu_\tau(z)\in R(p_\tau(z),q_\tau(z),\phi_\tau(z))} \left\{ |\lambda_\tau(z)| : \lambda_\tau(z) + \widetilde{\Omega}_\tau(z) - 1 = \sqrt{\lambda_\tau(z)}\widetilde{\Omega}_\tau(z)\mu_\tau(z) \right\} . \tag{42}$$

In [7], Hu, Jackson and Zhu determine a value $(\widetilde{\Omega}_{ellipse})_\tau(z)$ which minimises this upper bound *for a given ellipse*. Based on their result, [7, Thm. 1], we can immediately formulate the following lemma.

Lemma 8
Assume the matrices B and A are such that $\tau^{-1}a(z)/b(z)B + A$ is a block-consistently ordered matrix with nonsingular diagonal blocks. Assume the spectrum of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ lies in the closed interior of the ellipse $E(p_\tau(z), q_\tau(z), \phi_\tau(z))$, which does not contain the point 1. Define

$$(\widetilde{\Omega}_{ellipse})_\tau(z) = \frac{2}{1 + \sqrt{1 - (p_\tau^2(z) - q_\tau^2(z))e^{i2\phi_\tau(z)}}} , \tag{43}$$

where $\sqrt{\cdot}$ denotes the root with the positive real part. We then have

$$r_\tau^{ellipse}(z) = \left(|(\widetilde{\Omega}_{ellipse})_\tau(z)| \frac{p_\tau(z) + q_\tau(z)}{2} \right)^2 \tag{44}$$

and

$$\rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(z)\right) \le \rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},ellipse}(z)\right) \le r_\tau^{ellipse}(z) < 1 . \tag{45}$$

As before, the supscripts *opt* and *ellipse* are added to $\rho(\mathbf{K}_\tau^{\mathbf{CSOR}}(z))$ or r_τ to indicate the fact that the expressions of (41) or (42) are evaluated by using $(\tilde{\Omega}_{opt})_\tau(z)$ and $(\tilde{\Omega}_{ellipse})_\tau(z)$, respectively.

The following remarkable result from [7, §3] shows that there actually exists an ellipse for which the bound in (45) is attained. Uniqueness, however, of this *optimal* ellipse is not guaranteed by the theory in the above reference.

Lemma 9
There exists an optimal ellipse surrounding the spectrum of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ *for which* $(\tilde{\Omega}_{ellipse})_\tau(z) = (\tilde{\Omega}_{opt})_\tau(z)$, *and*

$$\rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(z)\right) = \rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},ellipse}(z)\right) = r_\tau^{ellipse}(z) < 1 \ . \tag{46}$$

In order to use Lemma 8 to compute the optimal convolution sequence $(\Omega_{opt})_\tau$ by one of the methods described in §5, one would have to determine the optimal ellipse containing $\sigma(\mathbf{K}_\tau^{\mathbf{JAC}}(z))$ for several values of z. A solution to the problem of finding this ellipse does exist when the eigenvalues of the Jacobi symbol $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ lie on a line segment $[-(\mu_1)_\tau(z), (\mu_1)_\tau(z)]$. In that case Lemma 4 shows that the optimal ellipse is degenerated and corresponds to the line segment linking the extremal eigenvalues. In particular, the parameters defining this ellipse are found by setting $(\mu_1)_\tau(z) = p_\tau(z)e^{i\phi_\tau(z)}$ and $q_\tau(z) = 0$. We do not know how to find such an optimal ellipse when the eigenvalues of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ are not collinear. Although an example has been given in [7, §4], even the problem of finding a *good* ellipse (which surrounds the spectrum of the Jacobi symbol and for which the associated bound is relatively sharp) may prove to be a formidable task. In practice, one therefore tries to determine a suitable convolution sequence without calculating these (nearly) optimal ellipses for all needed values of z. We suggest two possible strategies to this end.

A first attempt, for ODE systems with $B = I$ and $D_A = d_a I$, could start from the knowledge of a (nearly) optimal ellipse surrounding $\sigma(\mathbf{K}_\tau^{\mathbf{JAC}}(0))$. From formulae (11) and (15), it is clear that the spectrum of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ is obtained by rotating and scaling the spectrum of $\mathbf{K}_\tau^{\mathbf{JAC}}(0)$, corresponding to the multiplication with $d_a/(\tau^{-1}a(z)/b(z) + d_a)$. A similar operation applied to the ellipse surrounding the latter spectrum is then expected to lead to a good ellipse for the current value of z. This approach is however not applicable when $B \neq I$ or $D_A \neq d_a I$. Therefore, the numerical experiments in [8, §6] and [19] were performed with still another convolution sequence. In those references, the convolution kernel was computed by using the formula in the right-hand side of (12). This formula only requires the computation of a single eigenvalue $(\mu_1)_\tau(z)$ (which is the largest one in magnitude) for every z that appears in the particular inverse Z-transform method used. As the spectra $\sigma(\mathbf{K}_\tau^{\mathbf{JAC}}(z))$ are not collinear, the resulting kernel is not guaranteed to be the optimal one, or even a good one. Nevertheless, numerical evidence showed that this procedure yields excellent convergence rates for the problems considered. This observation led us in [8] to talk about the *robustness* of the CSOR waveform relaxation method.

With the generalised CSOR theory, this robustness can now be explained in an intuitive manner as follows. When the eigenvalues of the Jacobi symbol are not too far from being on a line, any reasonable ellipse – most probably also the optimal one – will be very elongated with $p_\tau(z)e^{i\phi_\tau(z)} \approx (\mu_1)_\tau(z)$ and $q_\tau(z)$ small. Thus the right-hand side of (12), which we denote further on by $(\widetilde{\Omega}_{appr})_\tau(z)$, is a good approximation to the optimal $(\widetilde{\Omega}_{opt})_\tau(z)$, computed by (43). If we set

$$(\widetilde{\Omega}_{appr})_\tau(z) = (\widetilde{\Omega}_{opt})_\tau(z) + \delta(z) \ ,$$

then it is easy to derive from (40), e.g. by doing a series expansion with Mathematica, that

$$\rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(z)\right) - \rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},appr}(z)\right) = O(\delta(z)) \ , \quad \delta(z) \to 0 \ .$$

The overall spectral radius of the CSOR iteration is found by a maximisation procedure over the unit circle, see Theorem 3 and in particular formula (10). Hence it is especially important for $(\widetilde{\Omega}_{appr})_\tau(z)$ to be close to $(\widetilde{\Omega}_{opt})_\tau(z)$ near the values of z for which $\rho(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(z))$ is large. In our experiments, this always appeared to be near the value $z = 1$. Fortunately, this is exactly where the eigenvalues of the Jacobi symbol are collinear or nearly collinear.

The above discussion will be illustrated by means of the *model problem* of [8, §6], that is, the two-dimensional heat equation, discretised on a regular, triangular mesh $\{(x_i = ih, y_j = jh) \mid 0 \le i, j \le \frac{1}{h}\}$ using linear finite elements. The resulting ODE system (1) was solved using linewise CSOR waveform relaxation, and the CN method with $\tau = 1/100$ was used for time discretisation. We computed ellipses surrounding $\sigma(\mathbf{K}_\tau^{\mathbf{JAC}}(z))$ for $h = 1/8$ and several values of $z = e^{i\theta}$ on the unit circle, as illustrated in Figure 4. These ellipses were obtained by choosing

$$\begin{cases} p_\tau(z) &= \quad |(\mu_1)_\tau(z)| \\ \phi_\tau(z) &= \quad \mathrm{Arg}((\mu_1)_\tau(z)) \end{cases} ,$$

and by determining $q_\tau(z)$ as the smallest value for which all eigenvalues of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ lie in the closed interior of the resulting ellipse. There is no firm guarantee that these ellipses are truly optimal. Yet, numerical experiments evaluating formula (41) with overrelaxation parameter from (43) for various *neighbouring* ellipses did never lead to a smaller value of the spectral radius. Hence, it seems reasonable to assume we have found an (at least locally) nearly optimal $\widetilde{\Omega}_\tau(z)$.

In the upper picture of Figure 5, we plotted $|\delta(z)|$, the modulus of the difference between the approximating $(\widetilde{\Omega}_{appr})_\tau(z)$ and the one we assume to be the

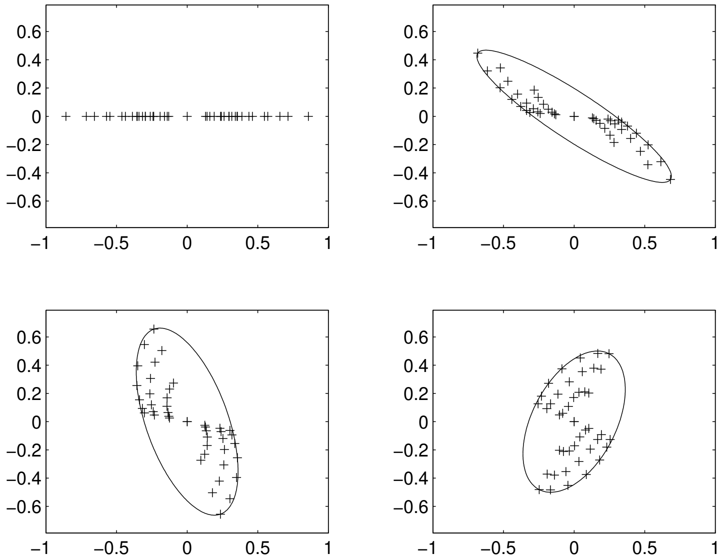

Figure 4: Eigenvalues ('+') of $\mathbf{K}_\tau^{\mathbf{JAC}}(z)$ and the optimal ellipses for several values of $z = e^{i\theta}$ for the model problem with $h = 1/8$. The respective pictures for $\theta = 0,\ 3\pi/12,\ 6\pi/12$ and $9\pi/12$ are ordered from left to right, top to bottom.

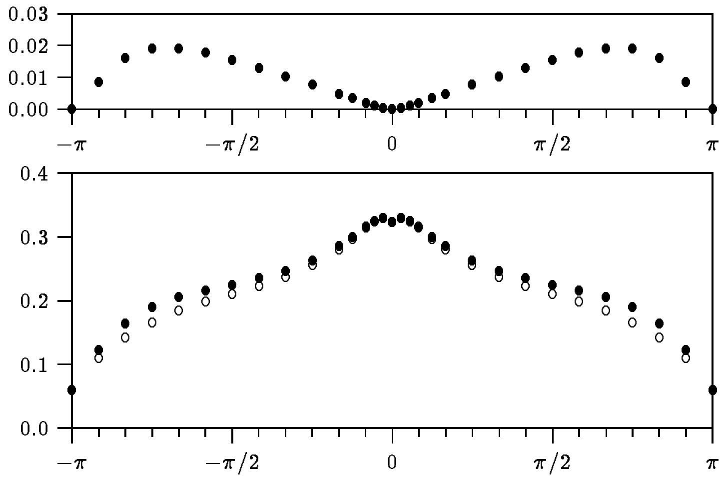

Figure 5: $|\delta(z)|$ (upper picture), $\rho(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(z))$ (lower picture, '∘') and $\rho(\mathbf{K}_\tau^{\mathbf{CSOR},appr}(z))$ (lower picture, '•') for several values of $z = e^{i\theta}$ for the model problem with $h = 1/8$.

Table 4: Averaged convergence factors for the model problem with linear finite-element discretisation and the CN method with $\tau = 1/100$, using an inverse Z-transform technique based on the right-hand side of (12) to calculate a suitable kernel.

h	1/8	1/16	1/32	1/64
$1-2\sqrt{2}\pi h$	-	0.445	0.722	0.861
convergence factors	0.320	0.569	0.757	0.870

optimal one for several values of $z = e^{i\theta}$ on the unit circle. The difference between the corresponding spectral radii $\rho(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(z))$ and $\rho(\mathbf{K}_\tau^{\mathbf{CSOR},appr}(z))$ is depicted in the lower picture of Figure 5. The collinearity of the spectrum of the Jacobi symbol implies that $\delta(e^{i\theta}) = 0$, and hence, $\rho(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(e^{i\theta})) = \rho(\mathbf{K}_\tau^{\mathbf{CSOR},appr}(e^{i\theta}))$ for $\theta = 0$ and $\theta = \pm\pi$. By noting that the maximum of the latter spectral radius over the unit circle is attained for θ in the neighbourhood of 0, we then derive that

$$\rho\left(\mathcal{K}_\tau^{CSOR,appr}\right) \approx \rho\left(\mathcal{K}_\tau^{CSOR,opt}\right) \approx \rho\left(\mathbf{K}_\tau^{\mathbf{CSOR},opt}(e^{i0})\right) .$$

The rightmost spectral radius of this expression, which equals the spectral radius of the optimal linewise SOR method for the linear system $Au = f$, is known to behave as $1-2\sqrt{2}\pi h$ for small enough h. Consequently, the linewise CSOR waveform relaxation method with the approximating kernel from (12) should demonstrate similar convergence results. This observation is confirmed by the numerical experiments in [8, §6], the resulting averaged convergence factors of which are recalled in Table 4.

Acknowledgements

The authors would like to thank Andrew Lumsdaine and Mark W. Reichelt for many interesting discussions on the topic of this paper.

References

[1] M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications, New York, 1970.

[2] A. Bellen, Z. Jackiewicz, and M. Zennaro. Contractivity of waveform relaxation Runge-Kutta iterations and related limit methods for dissipative systems in the maximum norm.

SIAM J. Numer. Anal., 31(2):499–523, April 1994.

[3] A. Bellen and M. Zennaro. The use of Runge-Kutta formulae in waveform relaxation methods. *Appl. Numer. Math.*, 11:95–114, 1993.

[4] R. N. Bracewell. *The Fourier Transform and its Applications.* McGraw-Hill Kogakusha, Ltd., Tokyo, 2nd edition, 1978.

[5] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*, volume 14 of *Springer Series in Computational Mathematics.* Springer-Verlag, Berlin, 1991.

[6] G. Horton, S. Vandewalle, and P. Worley. An algorithm with polylog parallel complexity for solving parabolic partial differential equations. *SIAM J. Sci. Comput.*, 16(3):531–541, May 1995.

[7] M. Hu, K. Jackson, and B. Zhu. Complex optimal SOR parameters and convergence regions. Department of Computer Science, University of Toronto, Canada, Working Notes, 1995.

[8] J. Janssen and S. Vandewalle. On SOR waveform relaxation methods. Technical Report CRPC-95-4, Center for Research on Parallel Computation, California Institute of Technology, Pasadena, California, U.S.A., October 1995. (accepted for publication in SIAM J. Numer. Anal.)

[9] J. Janssen and S. Vandewalle. Multigrid waveform relaxation on spatial finite-element meshes: The continuous-time case. *SIAM J. Numer. Anal.*, 33(2):456–474, April 1996.

[10] J. Janssen and S. Vandewalle. Multigrid waveform relaxation on spatial finite-element meshes: The discrete-time case. *SIAM J. Sci. Comput.*, 17(1):133–155, January 1996.

[11] C. Lubich. Chebyshev acceleration of Picard-Lindelöf iteration. *BIT*, 32:535–538, 1992.

[12] C. Lubich and A. Ostermann. Multi-grid dynamic iteration for parabolic equations. *BIT*, 27:216–234, 1987.

[13] A. Lumsdaine. *Theoretical and Practical Aspects of Parallel Numerical Algorithms for Initial Values Problems, with Applications.* Ph.D.-thesis, Deptartment of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A., Januari 1992.

[14] U. Miekkala and O. Nevanlinna. Convergence of dynamic iteration methods for initial value problems. *SIAM J. Sci. Statist. Comput.*, 8(4):459–482, July 1987.

[15] U. Miekkala and O. Nevanlinna. Sets of convergence and stability regions. *BIT*, 27:554–584, 1987.

[16] A. D. Poularakis and S. Seely. *Elements of Signals and Systems.* PWS-Kent Series in Electrical Engineering. PWS-Kent Publishing Company, Boston, 1988.

[17] S. Reed and B. Simon. *Functional Analysis*, volume 1 of *Methods of Modern Mathematical Physics.* Academic Press, New York, 1972.

[18] M. W. Reichelt. *Accelerated Waveform Relaxation Techniques for the Parallel Transient Simulation of Semiconductor Devices.* Ph.D.-thesis, Deptartment of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, Massachusetts, U.S.A., June 1993.

[19] M. W. Reichelt, J. K. White, and J. Allen. Optimal convolution SOR acceleration of waveform relaxation with application to parallel simulation of semiconductor devices. *SIAM J. Sci. Comput.*, 16(5):1137–1158, September 1995.

[20] R. Skeel. Waveform iteration and the shifted Picard splitting. *SIAM J. Sci. Stat. Comput.*, 10(4):756–776, July 1989.

[21] S. Vandewalle. *Parallel Multigrid Waveform Relaxation for Parabolic Problems.* B.G. Teubner, Stuttgart, 1993.

[22] S. Vandewalle and E. Van de Velde. Space-time concurrent multigrid waveform relaxation. *Annals of Numer. Math.*, 1:347–363, 1994.

[23] D. M. Young. *Iterative Solution of Large Linear Systems.* Academic Press, New York, 1971.