Composition methods, Maxwell's Equations and Source Terms

J.G. Verwer
Centrum voor Wiskunde & Informatica
P.O. Box 94079, 1090 GB Amsterdam, The Netherlands
Jan.Verwer@cwi.nl

November 27, 2010

Abstract

This paper is devoted to high-order numerical time integration of first-order wave equation systems originating from spatial discretization of Maxwell's equations. The focus lies on the accuracy of high-order composition in the presence of source functions. Source functions are known to generate order reduction and this is most severe for high-order methods. For two methods based on two well-known fourth-order symmetric compositions, convergence results are given assuming simultaneous space-time grid refinement. Herewith physical sources and source functions emanating from Dirichlet boundary conditions are distinguished. Amongst others it is shown that the reduction can cost two orders. On the other hand, when a certain perturbation of a source function is used, the reduction is generally diminished by one order. In that case reduction is absent for physical sources and for Dirichlet sources the order is equal to at least three under stable simultaneous space-time grid refinement.

2000 Mathematics Subject Classification: Primary: 65L05, 65L20, 65M12, 65M20.
1998 ACM Computing Classification System: G.1.7, G.1.8.
Keywords and Phrases: Numerical integration, composition methods, Maxwell's equations, Source terms.
Note: Work carried out within theme MAC1.

1 Introduction

Common spatial discretization of the Maxwell equations from electromagnetism

$$\begin{aligned} \mu\,\partial_t H &= -\nabla\times E\,, \\ \varepsilon\,\partial_t E &= \nabla\times H - \sigma E - J_E\,, \end{aligned} \tag{1}$$

results in linear systems of ordinary differential equations of the type

$$\begin{pmatrix} M_u & 0 \\ 0 & M_v \end{pmatrix}\begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{pmatrix} 0 & -K \\ K^T & -D \end{pmatrix}\begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} f^u(t) \\ f^v(t) \end{pmatrix}. \tag{2}$$

The vectors $u = u(t)$ and $v = v(t)$ are the unknown vector (grid) functions approximating the values of the magnetic field H and electric field E on the space grid, respectively. The matrices K and K^T

emanate from the curl operator $\nabla\times$. The matrix D is associated with the dissipative conduction term $-\sigma E$ and the matrices M_u, M_v typically represent mass matrices such as arising with finite elements. They also contain the values of the coefficients μ and ε. Further, the vector functions $f^u(t)$ and $f^v(t)$ are time-dependent source terms. Normally $f^v(t)$ represents the given source current J_E on the grid, but $f^u(t)$ and $f^v(t)$ may also contain Dirichlet boundary data.

Hence the partitioned ODE system (2) is of considerable practical interest as it is generic for semi-discrete Maxwell equations. In this paper we discuss high-order numerical integration of (2) when considered as a semi-discrete system. In particular, we will assume that element wise

$$K \sim \frac{1}{h}, \quad h \to 0, \tag{3}$$

where h parameterizes the distance of the (possibly nonuniform) space grid and the dimensions of the arising matrices and vectors. Hence we assume that the dimension of (2) is variable (PDE setting) and we thus do not consider a single system of fixed dimension as in the ODE setting.

In the remainder we also assume that we have eliminated the mass matrices so that instead of (2) we proceed with the semi-discrete system

$$\begin{pmatrix} u' \\ v' \end{pmatrix} = \begin{pmatrix} 0 & -K \\ K^T & -D \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} f^u(t) \\ f^v(t) \end{pmatrix}. \tag{4}$$

This somewhat more convenient form is obtained from the mass-matrix form through a simple transformation, see [1], and the numerical integration methods we discuss can be implemented for either choice. In particular, results for (4) always carry over to (2) and vice versa. For convenience of notation and presentation, we will therefore proceed with (4). Herein the damping matrix D is symmetric, non-negative definite. Often K is not square so the lengths of u and v generally are different. Except for common sufficient differentiability of the source functions, no further conditions are imposed on (4).

Composition methods and partitioned systems like (4) form a perfect match, see e.g. [6] for a description of the composition technique. One of the most popular integration methods for Maxwell's equations, the second-order method (7), is a composition method, see e.g. [9] and [1]. Composition is an elegant and powerful technique. One can directly build high-order methods from known compositions from the literature. Composition methods are also known to be accurate. However, in the PDE setting of semi-discrete systems, the convergence order of such a method may be lower than the chosen composition order. Such a reduction of order emanates from source terms, even from physical ones, and this occurs for composition methods of order greater than two. We examine this for two methods based on two well-known fourth-order compositions from the literature.[1)]

In Section 2 we will review local error analysis results for the second-order method (7) since we need these further on. In this section we also propose to perturb one of the source functions in a manner that the second-order method no longer shows local order reduction. Whereas this is not relevant for the global error of the second-order method, it is for the global error of higher-order composition methods. We will discuss this in Section 3 for the two methods based on fourth-order composition. For these two methods we will prove that due to the perturbation the general PDE order increases by one. Specifically, if one of the source functions contains Dirichlet boundary data the order is a least two without the perturbation and at least three with the perturbation. On the other hand, if boundary data is absent in both, these numbers are three and four. For given source functions, these convergence orders depend on the sequence of u and v used in the composition method. We will numerically illustrate the PDE convergence results in the final Section 4.

[1)] When we write order without referring specifically to the PDE setting, we will always mean the ODE order which is determined by the composition order.

2 The second-order method

In this section we review the second-order method which forms the basis for the composition methods discussed further on in the paper. Let Φ_τ denote the integration method

$$\begin{aligned} \frac{u_{n+1}-u_n}{\tau} &= -Kv_{n+1} + f^u(t_{n+1})\,, \\ \frac{v_{n+1}-v_n}{\tau} &= K^T u_n - Dv_{n+1} + f^v(t_{n+1})\,, \end{aligned} \tag{5}$$

and Φ^*_τ its adjoint

$$\begin{aligned} \frac{u_{n+1}-u_n}{\tau} &= -Kv_n + f^u(t_n)\,, \\ \frac{v_{n+1}-v_n}{\tau} &= K^T u_{n+1} - Dv_n + f^v(t_n)\,. \end{aligned} \tag{6}$$

The composition $\Phi_{\tau/2} \circ \Phi^*_{\tau/2}$ then defines the second-order method

$$\begin{aligned} \frac{u_{n+1/2}-u_n}{\tau} &= -\tfrac{1}{2}Kv_n + \tfrac{1}{2} f^u(t_n)\,, \\ \frac{v_{n+1}-v_n}{\tau} &= K^T u_{n+1/2} - \tfrac{1}{2}D(v_n+v_{n+1}) + \tfrac{1}{2}(f^v(t_n) + f^v(t_{n+1}))\,, \\ \frac{u_{n+1}-u_{n+1/2}}{\tau} &= -\tfrac{1}{2}Kv_{n+1} + \tfrac{1}{2} f^u(t_{n+1})\,. \end{aligned} \tag{7}$$

This one-step method steps from (u_n, v_n) to (u_{n+1}, v_{n+1}) with step size τ. Here u_n denotes the approximation to the exact solution $u(t_n)$, etc., and $\tau = t_{n+1} - t_n$. The method is explicit in the wave terms and implicit in D (the trapezoidal rule). If D is block-diagonal with a small bandwidth, as it is for discontinuous Galerkin finite element and finite difference discretizations, this implicitness comes with little costs. For $n \geq 1$ the third-stage derivative computation can be copied to the first stage at the next time step. Per time step this method thus is very economical as it actually requires a single righthand side evaluation per time step (for zero D), while it is second-order consistent (a consequence of symmetry). Method (7) is well-known in the literature on geometric integration, see e.g. [6], in particular for zero D. With regard to time stepping it bears a close resemblance to the popular Yee-scheme [14] from electromagnetism and to Verlet's method from molecular dynamics [11]. For the Maxwell equations it has for example been studied in [9] and [1, 13].

Our error analysis concerns temporal convergence towards the true solutions of the underlying PDE problem restricted to the space grid. We denote these by $u_h(t)$ and $v_h(t)$ and observe that these exact grid functions are solutions of the semi-discrete system

$$\begin{aligned} u_h'(t) &= -Kv_h(t) + f^u(t) + \sigma_h^u(t)\,, \\ v_h'(t) &= K^T u_h(t) - Dv_h(t) + f^v(t) + \sigma_h^v(t)\,, \end{aligned} \tag{8}$$

where $\sigma_h^u(t)$ and $\sigma_h^v(t)$ represent local spatial errors. In [1, 13] the following theorem has been proved:

Theorem 2.1 *Let the source functions $f^u(t), f^v(t) \in C^2[0,T]$ on a given finite time interval $[0,T]$ and suppose a Lax-Richtmyer stable space-time grid refinement $\tau \sim h$, $h \to 0$. On the interval $[0,T]$ the approximations u_n, v_n of method (7) then converge with temporal order two to $u_h(t), v_h(t)$.*

This theorem thus says that the second-order method does not suffer from order reduction. This second-order result is special in that the local error may suffer from reduction, cf. (14), which cancels in the transition towards global error. Below we will review the local errors for $\tau \sim h, h \to 0$ since we will need these when the method is used as building block for the higher-order composition methods.[2)] For the full proof of the theorem explaining the fortunate cancelation we refer to [1, 13]. Details on stability properties and energy conservation can also be found in [1].

2.1 Local error properties

We review the local error properties of method (7). To this end we first replace $f^v(t_n) + f^v(t_{n+1})$ by a perturbed source contribution $\tilde{f}^v(t_n) + \tilde{f}^v(t_{n+1})$ which will enable us to overcome the local order reduction. The precise definition will be given shortly. Simultaneously we eliminate the intermediate value $u_{n+1/2}$ from the second stage by substituting half of its expression obtained from the first and third stage. This yields the equivalent formulation

$$\begin{aligned}
\frac{u_{n+1}-u_n}{\tau} &= -\tfrac{1}{2}K(v_n+v_{n+1}) + \frac{1}{2}\left(f^u(t_n)+f^u(t_{n+1})\right), \\
\frac{v_{n+1}-v_n}{\tau} &= \tfrac{1}{2}K^T(u_n+u_{n+1}) - \tfrac{1}{2}D(v_n+v_{n+1}) + \tfrac{1}{2}\left(\tilde{f}^v(t_n)+\tilde{f}^v(t_{n+1})\right) \\
&\quad -\tfrac{1}{4}\tau K^T\left[-Kv_{n+1}+f^u(t_{n+1})\right] + \tfrac{1}{4}\tau K^T\left[-Kv_n+f^u(t_n)\right].
\end{aligned} \tag{9}$$

Substitution of $u_h(t_n)$ for u_n, etc., results in the defects δ_n^u and δ_n^v defined by

$$\begin{aligned}
\frac{u_h(t_{n+1})-u_h(t_n}{\tau} &= -\tfrac{1}{2}K\left(v_h(t_n)+v_h(t_{n+1})\right) + \tfrac{1}{2}\left(f^u(t_n)+f^u(t_{n+1})\right) + \delta_n^u, \\
\frac{v_h(t_{n+1})-v_h(t_n)}{\tau} &= \tfrac{1}{2}K^T\left(u_h(t_n)+u_h(t_{n+1})\right) - \tfrac{1}{2}D\left(v_h(t_n)+v_h(t_{n+1})\right) \\
&\quad + \tfrac{1}{2}\left(\tilde{f}^v(t_n)+\tilde{f}^v(t_{n+1})\right) - \tfrac{1}{4}\tau K^T\left[-Kv_h(t_{n+1})+f^u(t_{n+1})\right] \\
&\quad + \tfrac{1}{4}\tau K^T\left[-Kv_h(t_n)+f^u(t_n)\right] + \delta_n^v.
\end{aligned} \tag{10}$$

Using (8) we get

$$\begin{aligned}
\delta_n^u &= \frac{u_h(t_{n+1})-u_h(t_n}{\tau} - \frac{1}{2}\left(u_h'(t_n)+u_h'(t_{n+1})\right) + s_u^n, \\
\delta_n^v &= \frac{v_h(t_{n+1})-v_h(t_n)}{\tau} - \frac{1}{2}\left(v_h'(t_n)+v_h'(t_{n+1})\right) + \frac{1}{4}\tau K^T\left[u_h'(t_{n+1})-u_h'(t_n)\right] + s_n^v \\
&\quad + \tfrac{1}{2}\left(f^v(t_n)+f^v(t_{n+1})\right) - \tfrac{1}{2}\left(\tilde{f}^v(t_n)+\tilde{f}^v(t_{n+1})\right),
\end{aligned} \tag{11}$$

where

$$\begin{aligned}
s_n^u &= \tfrac{1}{2}\left(\sigma_h^u(t_n)+\sigma_h^u(t_{n+1})\right), \\
s_n^v &= \tfrac{1}{2}\left(\sigma_h^v(t_n)+\sigma_h^v(t_{n+1})\right) - \tfrac{1}{4}\tau K^T\left[\sigma_h^u(t_{n+1})-\sigma_h^u(t_n)\right],
\end{aligned} \tag{12}$$

denote the local spatial error contributions.

Because our focus lies on temporal accuracy, we will now omit s_n^u and s_n^v, that is, we simply put s_n^u and s_n^v to zero. This is not essential. Carrying the spatial errors along in further derivations yields

[2)] The notation $\tau \sim h, h \to 0$ is used throughout the paper and means that we consider a simultaneous space-time grid refinement, where the ratio between τ and h is determined by the common demand of Lax-Richtmyer stability.

no more insight in temporal accuracy. It would merely make our expressions more lengthy. We stress, however, that temporal accuracy will remain to be considered with respect to $u_h(t)$ and $v_h(t)$ for $\tau \sim h$, $h \to 0$. Henceforth $u_h(t)$ and $v_h(t)$ are supposed to be continuously differentiable as many times as the derivations require.

Let us first examine δ_n^u (for zero s_n^u) which is in fact the implicit trapezoidal rule defect. Expanding at the center point $t_{n+1/2}$ for $\tau \to 0$ yields the familiar expansion

$$\delta_n^u = -\frac{1}{12}\tau^2 u_h^{(3)} - \frac{1}{480}\tau^4 u_h^{(5)} + \cdots \,, \tag{13}$$

where $j = 2'$ means even values for j only and the derivatives are evaluated at $t = t_{n+1/2}$. The expansion contains only constants and (odd) solution derivatives which when appropriately measured (with the inner product norm) are bounded for $h \to 0$. So, if u_h is three times continuously differentiable, from Taylor's theorem with remainder we get $\delta_n^u = \mathscr{O}(\tau^2)$ with the order constant involved independent of τ and h.[3)]

Next we expand δ_n^v (for $s_n^v = 0$) at $t_{n+1/2}$, first without a source function perturbation, that is, with $\tilde{f}^v(t_n) + \tilde{f}^v(t_{n+1}) = f^v(t_n) + f^v(t_{n+1})$. We get

$$\delta_n^v = -\frac{1}{12}\tau^2 v_h^{(3)} - \frac{1}{480}\tau^4 v_h^{(5)} + \cdots + \tau K^T [\frac{1}{4}\tau u_h^{(2)} + \frac{1}{96}\tau^3 u_h^{(4)} + \cdots]\,. \tag{14}$$

Because of property (3) we have $\tau K^T = \mathscr{O}(1)$ for $\tau \sim h, h \to 0$. This means that in general the second part of the expansion is only $\mathscr{O}(\tau)$ and hence $\delta_n^v = \mathscr{O}(\tau)$ instead of $\mathscr{O}(\tau^2)$. Would

$$K^T u_h^{(2)}(t) = v_h^{(3)}(t) + D v_h^{(2)}(t) - \frac{d^2}{dt^2} f^v(t) - \frac{d^2}{dt^2}\sigma_h^{(v)}(t) = \mathscr{O}(1), \quad h \to 0\,, \tag{15}$$

then $\delta_n^v = \mathscr{O}(\tau^2)$ for $\tau \sim h, h \to 0$. This holds if

$$\frac{d^2}{dt^2} f^v(t) = \mathscr{O}(1), \quad h \to 0\,, \tag{16}$$

because the third derivative of $v_h(t)$ and the second derivatives of $D v_h(t)$ and $\sigma_h^{(v)}(t)$ are bounded. Condition (16) is true for physical sources $f^v(t)$ but generally not if $f^v(t)$ contains Dirichlet boundary data, since then part of its components behave as $\mathscr{O}(h^{-1}), h \to 0$, cf. property (3), and this generally also holds for the derivatives.

To overcome this possible cause of local order reduction[4)] we now define the perturbed source function contribution

$$\tilde{f}^v(t_n) + \tilde{f}^v(t_{n+1}) = f^v(t_n) + f^v(t_{n+1}) + \frac{1}{2}\tau\frac{d}{dt}\left(f^v(t_n) - f^v(t_{n+1})\right), \tag{17}$$

where we emphasize that the perturbation is defined for the sum. With this definition δ_n^v becomes

$$\begin{aligned}\delta_n^v = {} & \frac{v_h(t_{n+1}) - v_h(t_n)}{\tau} - \frac{1}{2}\left(v_h'(t_n) + v_h'(t_{n+1})\right) \\ & + \tfrac{1}{4}\tau\left(v_h''(t_{n+1} - v_h''(t_n)\right) + \tfrac{1}{4}\tau D\left(v_h(t_{n+1} - v_h(t_n)\right)\,.\end{aligned} \tag{18}$$

[3)] Unless noted otherwise, the symbol $\mathscr{O}(\cdot)$ will always be used with this meaning, that is, order constants exist which are independent of τ and h for $\tau \sim h \to 0$.

[4)] As proved in [1, 13], this local order reduction does not affect the 2nd-order convergence of method (7) for $\tau \sim h$, $h \to 0$.

Expanding in the same way as for δ_n^u gives

$$\delta_n^v = \tau^2 \left[\frac{1}{6} v_h^{(3)} + \frac{1}{4} D v_h^{(1)} \right] + \tau^4 \left[\frac{1}{120} v_h^{(5)} + \frac{1}{96} D v_h^{(3)} \right] + \cdots . \tag{19}$$

Like for u_h, if v_h is three times continuously differentiable, from Taylor's theorem with remainder we get $\delta_n^v = \mathscr{O}(\tau^2)$ for $\tau \sim h, h \to 0$ with the order constant involved independent of τ and h.

2.2 The global error recursion

Let $\varepsilon_n^u = u_h(t_n) - u_n$ and $\varepsilon_n^v = v_h(t_n) - v_n$ denote the global errors. From (9) and the local error discussion we deduce the following global error recursion:

$$\begin{gathered} \begin{pmatrix} I & \frac{1}{2}\tau K \\ -\frac{1}{2}\tau K^T & I - \frac{1}{4}\tau^2 K^T K + \frac{1}{2}\tau D \end{pmatrix} \begin{pmatrix} \varepsilon_{n+1}^u \\ \varepsilon_{n+1}^v \end{pmatrix} = \\ \begin{pmatrix} I & -\frac{1}{2}\tau K \\ \frac{1}{2}\tau K^T & I - \frac{1}{4}\tau^2 K^T K - \frac{1}{2}\tau D \end{pmatrix} \begin{pmatrix} \varepsilon_n^u \\ \varepsilon_n^v \end{pmatrix} + \tau \begin{pmatrix} \delta_u^n \\ \delta_v^n \end{pmatrix}, \end{gathered} \tag{20}$$

and putting $\varepsilon_n = [(\varepsilon_n^u)^T, (\varepsilon_n^v)^T]^T$ and $\delta_n = [(\delta_n^u)^T, (\delta_n^v)^T]^T$, we arrive at the compact notation

$$\varepsilon_{n+1} = R\varepsilon_n + \tau\rho_n , \qquad R = R_L^{-1} R_R , \quad \rho_n = R_L^{-1} \delta_n , \tag{21}$$

with R_L and R_R the left and right block matrix, respectively. This recursion has the standard form featuring in the convergence analysis of one-step integration methods, see e.g. [7]. Assuming Lax-Richtmyer stability, whereby we include R_L inversely bounded for $\tau \sim h, h \to 0$, it transfers local errors to the global error by essentially adding all local errors. It reveals second-order convergence for $\tau \sim h, h \to 0$, if both δ_n^u and δ_n^v are $\mathscr{O}(\tau^2)$ for $\tau \sim h, h \to 0$.

2.3 Reversed u, v sequence

The sequence u, v in (7) may be reversed. For this second-order method this is not relevant. However, when used as a base method for higher-order composition, there may arise significant accuracy differences. This fully depends on the source terms, i.e., whether they contain Dirichlet boundary data or not. We will illustrate this in the numerical Section 4. Taking into account the sequence and the source function perturbation, this means that altogether four different second-order methods are distinguished, namely (7), its version with the perturbation (17),

$$\begin{aligned} \frac{u_{n+1/2} - u_n}{\tau} &= -\tfrac{1}{2} K v_n + \tfrac{1}{2} f^u(t_n) , \\ \frac{v_{n+1} - v_n}{\tau} &= K^T u_{n+1/2} - \tfrac{1}{2} D(v_n + v_{n+1}) + \tfrac{1}{2}(\tilde{f}^v(t_n) + \tilde{f}^v(t_{n+1})) , \\ \frac{u_{n+1} - u_{n+1/2}}{\tau} &= -\tfrac{1}{2} K v_{n+1} + \tfrac{1}{2} f^u(t_{n+1}) , \end{aligned} \tag{22}$$

its version with reversed sequence,

$$\begin{aligned}
\frac{v_{n+1/2}-v_n}{\tau} &= \tfrac{1}{2}K^T u_n - \tfrac{1}{2}Dv_n + \tfrac{1}{2}f^v(t_n)\,,\\
\frac{u_{n+1}-u_n}{\tau} &= -Kv_{n+1/2} + \tfrac{1}{2}\big(f^u(t_n)+f^u(t_{n+1})\big)\,,\\
\frac{v_{n+1}-v_{n+1/2}}{\tau} &= \tfrac{1}{2}K^T u_{n+1} - \tfrac{1}{2}Dv_{n+1} + \tfrac{1}{2}f^v(t_n+1)\,,
\end{aligned} \tag{23}$$

and its version with reversed sequence and the perturbation (17) applied to f^u,

$$\begin{aligned}
\frac{v_{n+1/2}-v_n}{\tau} &= \tfrac{1}{2}K^T u_n - \tfrac{1}{2}Dv_n + \tfrac{1}{2}f^v(t_n)\,,\\
\frac{u_{n+1}-u_n}{\tau} &= -Kv_{n+1/2} + \tfrac{1}{2}\big(\tilde{f}^u(t_n)+\tilde{f}^u(t_{n+1})\big)\,,\\
\frac{v_{n+1}-v_{n+1/2}}{\tau} &= \tfrac{1}{2}K^T u_{n+1} - \tfrac{1}{2}Dv_{n+1} + \tfrac{1}{2}f^v(t_n+1)\,.
\end{aligned} \tag{24}$$

For the analysis it is sufficient to only consider methods (7) and (22).

3 Symmetric composition methods

Our aimed methods are based on symmetric compositions

$$\Psi_\tau^{(4)} = \Psi_{\gamma_s\tau}^{(2)} \circ \cdots \circ \Psi_{\gamma_1\tau}^{(2)} \tag{25}$$

of composition order four ($\gamma_1+\cdots+\gamma_s=1$ and $\gamma_1^3+\cdots+\gamma_s^3=0$) where $\Psi_{\gamma_k\tau}^{(2)}$ represents one of the four methods from Section 2.3. Within this composition, the base method steps from $t_n+(\gamma_1+\cdots+\gamma_{k-1})\tau$ to $t_n+(\gamma_1+\cdots+\gamma_k)\tau$ for $k=1,\ldots,s$ spanning the interval $[t_n,t_{n+1}]$. For composition order four two compositions of interest have $s=3$ and $s=5$, respectively,

$$\gamma_1=\gamma_3=\frac{1}{2-2^{1/3}}\,,\quad \gamma_2=-\frac{2^{1/3}}{2-2^{1/3}}\,, \tag{26}$$

and

$$\gamma_1=\gamma_2=\gamma_4=\gamma_5=\frac{1}{4-4^{1/3}}\,,\quad \gamma_3=-\frac{4^{1/3}}{4-4^{1/3}}\,. \tag{27}$$

We have taken these parameters from [6], formulas (II.4.4) and (II.4.5), where for $s=3$ a reference is given to [2, 4, 10, 15] and for $s=5$ to [10].

A convergence proof for method (25) is given in [6]. This proof, however, does not take into account the Lipschitz constant of the ODE system which essentially means that for our case it is restricted to a fixed ODE system, whereas we wish to investigate the order for $\tau\sim h, h\to 0$. In Section 3.2 we will present a proof for the following counterpart of Theorem 2.1:

Theorem 3.1 *Let D be zero, $f^u(t), f^v(t)\in C^p[0,T]$, and suppose a Lax-Richtmyer stable space-time grid refinement $\tau\sim h$, $h\to 0$. On $[0,T]$ the approximations u_n, v_n of method (25) based on (22) and parameters (26) or (27) then converge to $u_h(t), v_h(t)$ with*
(i) at least order $p=3$,
(ii) order $p=4$, if in addition $K^Tu_h^{(3)}(t), Kv_h^{(3)}(t)=\mathcal{O}(1)$ for $h\to 0$.

We have taken $D = 0$ as this simplifies the analysis. With respect to order reduction this is not essential as order reduction is not related to conduction.

Theorem 3.1 states that on the whole problem class (4) with $f^u(t), f^v(t) \in C^3[0,T]$ order three is guaranteed. If both source functions are in $C^4[0,T]$ and the additional condition on the third derivatives is satisfied, the composition order four will hold. From

$$\begin{aligned} Kv_h^{(3)}(t) &= -u_h^{(4)}(t) + \frac{d^3}{dt^3} f^u(t) + \frac{d^3}{dt^3} \sigma_h^u(t)\,, \\ K^T u_h^{(3)}(t) &= v_h^{(4)}(t) + Dv_h^{(3)}(t) - \frac{d^3}{dt^3} f^v(t) - \frac{d^3}{dt^3} \sigma_h^v(t)\,, \end{aligned} \tag{28}$$

follows that this is true if for $h \to 0$ the source functions satisfy

$$\frac{d^3}{dt^3} f^u(t) = \mathscr{O}(1), \quad \frac{d^3}{dt^3} f^v(t) = \mathscr{O}(1)\,, \tag{29}$$

because the fourth derivatives of $u_h(t), v_h(t)$ and the third derivatives of $Dv_h(t), \sigma_h^u(t), \sigma_h^v(t)$ are bounded for $h \to 0$.

This boundedness condition applies to physical sources, but is violated by sources containing Dirichlet boundary data, since for these there will exist components which are $\mathscr{O}(h^{-1})$ for $h \to 0$. Hence with only physical sources we are guaranteed that there will be no order reduction. With Dirichlet boundary data we are guaranteed that we have order three, and we expect that normally order three will show up. However, for special solutions the order may lie between three and four, even if the (sufficient) condition of assertion (ii) will be violated.

We owe the good convergence results of Theorem 3.1 to the perturbed source function contribution (17). Generally, by using (17) the reduction is diminished with one order. The following theorem, where the composition is based on the original method (7), clarifies this:

Theorem 3.2 *Let D be zero, $f^u(t), f^v(t) \in C^p[0,T]$, and suppose a Lax-Richtmyer stable space-time grid refinement $\tau \sim h, h \to 0$. On $[0,T]$ the approximations u_n, v_n of method (25) based on (7) and parameters (26) or (27) then converge to $u_h(t), v_h(t)$ with*
(i) at least order $p = 2$,
(ii) at least order $p = 3$, if in addition $K^T u_h^{(2)}(t) = \mathscr{O}(1)$ for $h \to 0$,
(iii) order four $p = 4$, if in addition $K^T u_h^{(3)}(t)$, $Kv_h^{(3)}(t)$, $KK^T u_h^{(2)}(t) = \mathscr{O}(1)$ for $h \to 0$.

Similar as above, from (8) follows that $K^T u_h^{(2)}(t) = \mathscr{O}(1)$ if

$$\frac{d^2}{dt^2} f^v(t) = \mathscr{O}(1), \quad h \to 0\,, \tag{30}$$

while the additional conditions for order four are satisfied if (29) holds and if

$$K\frac{d^2}{dt^2} f^v(t) = \mathscr{O}(1), \quad h \to 0\,. \tag{31}$$

In particular this latter condition is restrictive and implies that even with only physical sources order four for $\tau \sim h, h \to 0$ will rarely occur. However, as observed above, for special solutions the order reduction may be less, even if the (sufficient) conditions of assertion (ii) and (iii) will be violated.

When comparing Theorems 3.1 and 3.2, it is obvious that the perturbed source function contribution (17) should be used as a default option. In the numerical Section 4 we will illustrate this, both for the base methods (7) and (22), assumed in these theorems, as well as for their reversed versions (23) and (24). Finally, because the proof of Theorem 3.2 goes similar as that of Theorem 3.1, we refrain from presenting it here so as to avoid duplication.

3.1 Step-by-step stability

Before proving the above convergence theorems we recall the stability analysis as this is based on material also needed for the proofs. Consider the semi-discrete system (4). Assume $u \in \mathbb{R}^m, v \in \mathbb{R}^n$ with $n \geqslant m$ (the reversed case can be treated likewise) and thus $K \in \mathbb{R}^{m \times n}$ and $D \in \mathbb{R}^{n \times n}$. Let $w \in \mathbb{R}^{n+m}$ denote the solution vector composed by u, v. A natural norm for establishing stability is the inner-product norm $\|w\|^2 = \langle u, u \rangle + \langle v, v \rangle$. As D is symmetric positive semi-definite, and for zero D the matrix of the system is skew-symmetric, for the homogeneous part of (4) follows

$$\frac{d}{dt}\|w\|^2 = -2\langle Dv, v \rangle \leqslant 0\,, \tag{32}$$

showing stability in the inner product norm.

For numerical stability analysis we suppose that the conduction matrix D is constant diagonal, $D = \alpha I$ say. This holds if in (1) the conductivity coefficient σ and the permittivity coefficient ε are constant scalars and allows the use of the singular value decomposition $K = U \Lambda V^T$ where $U \in \mathbb{R}^{m \times m}$ and $V \in \mathbb{R}^{n \times n}$ are orthogonal and Λ is a diagonal $m \times n$ matrix with nonnegative diagonal entries λ_1, $\ldots$, λ_m satisfying

$$\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_r > \lambda_{r+1} = \cdots = \lambda_m = 0\,. \tag{33}$$

Here $r \leqslant m$ is the (row) rank of K and the λ_i are the singular values of K (the square roots of the eigenvalues of KK^T). The transformed variables and source terms

$$\overline{u}(t) = U^T u(t)\,, \quad \overline{v}(t) = V^T v(t)\,, \quad \overline{f}^u(t) = U^T f^u(t)\,, \quad \overline{f}^v(t) = V^T f^v(t)\,, \tag{34}$$

satisfy the ODE system

$$\begin{pmatrix} \overline{u}' \\ \overline{v}' \end{pmatrix} = \begin{pmatrix} 0 & -\Lambda \\ \Lambda^T & -\alpha I \end{pmatrix} \begin{pmatrix} \overline{u} \\ \overline{v} \end{pmatrix} + \begin{pmatrix} \overline{f}^u(t) \\ \overline{f}^v(t) \end{pmatrix}. \tag{35}$$

Because the matrix transformation induced by (34) is a similarity transformation, the matrices of systems (4) and (35) have the same eigenvalues. Further, $\|u\|_2^2 + \|v\|_2^2 = \|\overline{u}\|_2^2 + \|\overline{v}\|_2^2$ due to the orthogonality of U and V. Thus, if $D = \alpha I$ applies, the stability of any time integration method may be studied for the homogeneous part of (35), provided also the method is invariant under the transformations leading to (35). This holds for the methods considered in this paper.

Since the matrix Λ is diagonal, system (35) decouples into r two-by-two systems

$$\begin{pmatrix} \hat{u}' \\ \hat{v}' \end{pmatrix} = \begin{pmatrix} 0 & -\lambda \\ \lambda & -\alpha \end{pmatrix} \begin{pmatrix} \hat{u} \\ \hat{v} \end{pmatrix} + \begin{pmatrix} \hat{f}^u(t) \\ \hat{f}^v(t) \end{pmatrix}, \qquad \lambda = \lambda_k > 0\,, \quad k = 1, \ldots, r\,, \tag{36}$$

$m-r$ scalar equations $\hat{u}' = \hat{f}^u(t)$, and $n-r$ scalar equations $\hat{v}' = -\alpha\hat{v} + \hat{f}^v(t)$.[5] This the canonical form for semi-discrete Maxwell equation systems with $D = \alpha I$. Both with regard to stability, consistency and convergence analysis, numerical methods which are invariant under the used transformation can be examined for this canonical form. Herewith the $m-r$ scalar equations $\hat{u}' = \hat{f}^u(t)$, and $n-r$ scalar equations $\hat{v}' = -\alpha\hat{v} + \hat{f}^v(t)$ are trivial. What matters are the r two-by-two systems (36) of which the homogeneous form provides a useful test model for stability.

[5] We have used the singular value decomposition also in [1, 12] and note that the description of the decoupling given in [1] contains an error.

When applied to the homogeneous form of (36), the composition method based on (7) or (22) yields

$$\begin{pmatrix} \hat{u}_{n+1} \\ \hat{v}_{n+1} \end{pmatrix} = \prod_{k=s}^{1} \frac{1}{1+\frac{1}{2}\gamma_k z_\alpha} \begin{pmatrix} 1+\frac{1}{2}\gamma_k z_\alpha - \frac{1}{2}\gamma_k^2 z_\lambda^2 & -\gamma_k z_\lambda + \frac{1}{4}\gamma_k^3 z_\lambda^3 \\ \gamma_k z_\lambda & 1-\frac{1}{2}\gamma_k z_\alpha - \frac{1}{2}\gamma_k^2 z_\lambda^2 \end{pmatrix} \begin{pmatrix} \hat{u}_n \\ \hat{v}_n \end{pmatrix}, \qquad (37)$$

where $z_\alpha = \tau\alpha, z_\lambda = \tau\lambda$. We define stability through the common root condition: at (z_α, z_λ) the two roots of the characteristic equation of the amplification matrix lie on the unit disc and are different when both lie on the unit circle. We recall that for method (7) and its three counterparts from Section 2.3 holds that for $\alpha = 0$ the root condition is satisfied if and only if $z_\lambda < 2$, while for $\alpha > 0$ the root condition is satisfied if and only if $z_\lambda \leq 2$ [1]. Hence there is no step size restriction on the conduction coefficient α.

For the composition methods defined by the parameter sets (26) and (27) we also distinguish between $\alpha = 0$ and $\alpha > 0$. For $\alpha = 0$ the stability interval is the largest interval $(0, z_\lambda)$ along which the root condition holds. Along this interval both roots lie on the unit circle. A numerical search has resulted in $(0, \frac{1}{2}\pi]$ for $s = 3$ and $(0, e]$ for $s = 5$, where $\frac{1}{2}\pi$ and e are accurate lower bounds. For $\alpha > 0$ we have computed, with a numerical search, the stability regions

$$\mathscr{S} = \{(z_\alpha, z_\lambda) : \ z_\alpha, z_\lambda \geq 0 \ \text{ and both roots have modulus} < 1\}, \qquad (38)$$

where we impose the slightly stricter condition < 1, see Figure 1. Both regions contain a hole along the z_α-axis due to the negative time step (see (26) and (27)) which imposes a step size restriction determined by the conduction coefficient α. Further, for $s = 5$ the region is larger due to smaller coefficients γ_k. Taking into account the workload (five sub steps or stages compared to three), the advantage of a larger $\mathscr{S}$ still exists. This advantage is negligible for $\alpha = 0$ (compare the scaled lengths $e/5 \approx 0.54$ and $\frac{1}{2}\pi/3 \approx 0.52$). Finally, when stability is more important than accuracy and the workload is taken into account, it is clear that the methods for $s = 3$ and $s = 5$ cannot compete with the second-order methods. This holds in particular if conduction terms limit the step size.

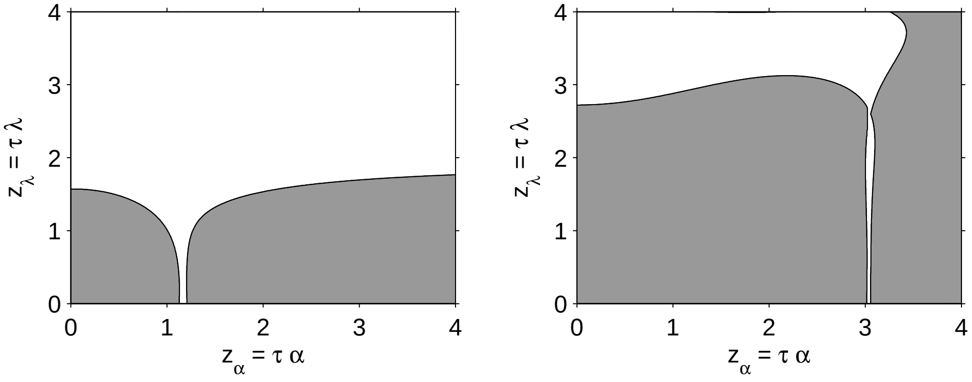

Figure 1: The stability regions $\mathscr{S}$ of the two composition methods. At the left for $s = 3$, at the right for $s = 5$.

3.2 Proof of Theorem 3.1

We will give the proof for $s=3$. From the derivations and results gathered for $s=3$ one can readily see that the case $s=5$ goes in precisely the same way.

3.2.1 Preliminaries

Consider the global error recursion (20). Let R_k denote the amplification operator R introduced in (21) with τ replaced by $\gamma_k\tau$ and define $R_{k,L}$ as the counterpart of R_L. With the material of Sections 2.1 and 2.2 one then readily derives for the composition method (25) based on (22) the global error recursion

$$\varepsilon_{n+1} = R_3R_2R_1\varepsilon_n + R_3R_2R_{1,L}^{-1}\delta_n^{(1)} + R_3R_{2,L}^{-1}\delta_n^{(2)} + R_{3,L}^{-1}\delta_n^{(3)}, \tag{39}$$

where

$$\delta_n^{(k)} = \begin{pmatrix} -\frac{1}{12}\gamma_k^3\tau^3u_h^{(3)}(s_k) - \frac{1}{480}\gamma_k^5\tau^5u_h^{(5)}(s_k) + \cdots \\ \frac{1}{6}\gamma_k^3\tau^3v_h^{(3)}(s_k) + \frac{1}{120}\gamma_k^5\tau^5v_h^{(5)}(s_k) + \cdots \end{pmatrix}, \tag{40}$$

and $s_k = t_n + (\gamma_1 + \cdots + \gamma_{k-1} + \frac{1}{2}\gamma_k)\tau$ denotes the center point for the k-th sub step. Note that we here have included the step size factor $\gamma_k\tau$ into the defect expressions. For zero D we can express $R_{k,L}^{-1}$ and R_k as

$$\begin{aligned} R_{k,L}^{-1} &= \begin{pmatrix} I - \frac{1}{4}\gamma_k^2\tau^2KK^T & -\frac{1}{2}\gamma_k\tau K \\ \frac{1}{2}\gamma_k\tau K^T & I \end{pmatrix}, \\ R_k &= \begin{pmatrix} I - \frac{1}{2}\gamma_k^2\tau^2KK^T & -\gamma_k\tau K + \frac{1}{4}\gamma_k^3\tau^3KK^TK \\ \gamma_k\tau K^T & I - \frac{1}{2}\gamma_k^2\tau^2K^TK \end{pmatrix}, \end{aligned} \tag{41}$$

and since $\tau K = \mathscr{O}(1)$ for $\tau \sim h, h \to 0$ due to (3), this also holds for these two matrices and any combination thereof.

We write (39) as

$$\varepsilon_{n+1} = \mathscr{R}\varepsilon_n + \rho_n\,, \quad \mathscr{R} = R_3R_2R_1, \quad \rho_n = R_3R_2R_{1,L}^{-1}\delta_n^{(1)} + R_3R_{2,L}^{-1}\delta_n^{(2)} + R_{3,L}^{-1}\delta_n^{(3)}\,, \tag{42}$$

and introduce the following Ansatz ([7], Lemma II.2.3): ρ_n can be written as

$$\rho_n = (I - \mathscr{R})\xi_n + \eta_n\,, \tag{43}$$

with ξ_n and η_n local error quantities satisfying

$$\xi_n = \mathscr{O}(\tau^p), \quad \xi_{n+1} - \xi_n = \mathscr{O}(\tau^{p+1}) \quad \text{and} \quad \eta_n = \mathscr{O}(\tau^{p+1})\,. \tag{44}$$

If this holds, then $\tilde{\varepsilon}_n = \varepsilon_n - \xi_n$ satisfies the recurrence

$$\tilde{\varepsilon}_{n+1} = \mathscr{R}\tilde{\varepsilon}_n - (\xi_{n+1} - \xi_n) + \eta_n\,, \tag{45}$$

with an $\mathscr{O}(\tau^{p+1})$ local error. Assuming Lax-Richtmeyer stability then gives in the standard way $\mathscr{O}(\tau^p)$ for $\tilde{\varepsilon}_n$ and hence for ε_n. The importance of the Ansatz is thus that the global order can be proven to be equal to the order of ξ_n, which is a local quantity. Consequently, the proof of

Theorem 3.1 is complete if for $\tau \sim h, h \to 0$ the Ansatz applies with $p = 3$ for assertion (i) and with $p = 4$ for assertion (ii).

For examining (43) we use the singular value decomposition of Section 3.1. This means that within the expressions (41) one may read Λ for K and Λ^T for K^T and then, following the decoupling into the r two-by-two systems (36), decouple also $R_{k,L}^{-1}$ and R_k in r two-by-two matrices[6]

$$\hat{R}_{k,L}^{-1} = \begin{pmatrix} 1-\frac{1}{4}\gamma_k^2 z^2 & -\frac{1}{2}\gamma_k z \\ \frac{1}{2}\gamma_k z & 1 \end{pmatrix}, \quad \hat{R}_k = \begin{pmatrix} 1-\frac{1}{2}\gamma_k^2 z^2 & -\gamma_k z + \frac{1}{4}\gamma_k^3 z^3 \\ \gamma_k z & 1-\frac{1}{2}\gamma_k^2 z^2 \end{pmatrix}, \quad z = \tau\lambda\,. \tag{46}$$

Hence (43) is replaced by r two-by-two systems

$$\hat{\rho}_n = (I - \hat{\mathscr{R}})\hat{\xi}_n + \hat{\eta}_n\,, \tag{47}$$

where $\hat{\rho}_n$ is the transformed counterpart of ρ_n, etc. In accordance with the limit transition $\tau \sim h, h \to 0$ and $\tau K = \mathscr{O}(1)$ we will consider z uniformly in an interval $[0, z_{max}]$ with $(0, z_{max}] \subset$ the stability interval of the numerical method as defined in Section 3.1. Note that this implies Lax-Richtmyer stability. The end point z_{max} will be defined below.

3.2.2 Assertion (i)

If $f^u, f^v \in C^3[0,T]$, then $u_h, v_h \in C^4[0,T]$. For $\tau \sim h, h \to 0$, Taylor's theorem with remainder then allows us to replace (40) by

$$\delta_n^{(k)} = \gamma_k^3 \begin{pmatrix} -\frac{1}{12}\tau^3 u_h^{(3)} + \mathscr{O}(\tau^4) \\ \frac{1}{6}\tau^3 v_h^{(3)} + \mathscr{O}(\tau^4) \end{pmatrix}, \tag{48}$$

where the third derivatives may be taken at any $t \in [t_n, t_{n+1}]$ independent of k. Hence we can express the local error ρ_n as

$$\rho_n = \mathscr{L} w_n + \mathscr{O}(\tau^4), \quad w_n = \begin{pmatrix} -\frac{1}{12}\tau^3 u_h^{(3)} \\ \frac{1}{6}\tau^3 v_h^{(3)} \end{pmatrix}, \quad \mathscr{L} = \gamma_1^3 R_3 R_2 R_{1,L}^{-1} + \gamma_2^3 R_3 R_{2,L}^{-1} + \gamma_3^3 R_{3,L}^{-1}\,, \tag{49}$$

The local error is of order three. For proving convergence order three in the standard way we need a local error of order four. To circumvent this we now employ the Ansatz with $p = 3$. Trivially, for η_n we may choose the $\mathscr{O}(\tau^4)$ term in (49) and there remains to deal with the relation $\mathscr{L} w_n = (I - \mathscr{R})\xi_n$. For this purpose we proceed with the transformed counterpart

$$\hat{\mathscr{L}} \hat{w}_n = (I - \hat{\mathscr{R}})\hat{\xi}_n\,. \tag{50}$$

Let us write

$$\begin{aligned} \hat{R}_{k,L}^{-1} &= I + z\hat{A}_k, \quad \hat{A}_k = \begin{pmatrix} -\frac{1}{4}\gamma_k^2 z & -\frac{1}{2}\gamma_k \\ \frac{1}{2}\gamma_k & 0 \end{pmatrix}, \\ \hat{R}_k &= I + z\hat{B}_k, \quad \hat{B}_k = \begin{pmatrix} -\frac{1}{2}\gamma_k^2 z & -\gamma_k + \frac{1}{4}\gamma_k^3 z^2 \\ \gamma_k & -\frac{1}{2}\gamma_k^2 z \end{pmatrix}, \end{aligned} \tag{51}$$

[6] The scalar equations associated with zero singular values play a trivial role. Note that instead of z_λ we here write z.

and substitute into $\hat{\mathscr{R}}$ and $\hat{\mathscr{L}}$. Using the third-order condition $\gamma_1^3 + \gamma_2^3 + \gamma_3^3 = 0$ we then can extract one factor z from (50), that is, we can write

$$\hat{\mathscr{L}} = z\hat{\mathscr{C}}, \quad I - \hat{\mathscr{R}} = z\hat{\mathscr{D}}\,, \tag{52}$$

where the two-by-two matrix $\hat{\mathscr{C}}$ collects remaining $\mathscr{O}(1)$ terms. If $\hat{\mathscr{D}}^{-1}$ exists and is bounded uniformly in $[0, z_{max}]$, then

$$\hat{\xi}_n = \hat{\mathscr{D}}^{-1}\hat{\mathscr{C}}\hat{w}_n = \mathscr{O}(\tau^3), \quad \hat{\xi}_{n+1} - \hat{\xi}_n = \mathscr{O}(\tau^4)\,, \tag{53}$$

and $\hat{\xi}_n$ satisfies (50).

Consequently, we are done if $\hat{\mathscr{D}}^{-1}$ exists and is bounded uniformly in $[0, z_{max}]$. There holds

$$\hat{\mathscr{D}} = -\hat{\mathscr{D}}_0 - z\left(\hat{B}_3\hat{B}_2 + \hat{B}_3\hat{B}_1 + \hat{B}_2\hat{B}_1\right) - z^2\hat{B}_3\hat{B}_2\hat{B}_1\,, \tag{54}$$

where, using $\gamma_1 + \gamma_2 + \gamma_3 = 1$ and $\gamma_1^3 + \gamma_2^3 + \gamma_3^3 = 0$,

$$\hat{\mathscr{D}}_0 = \sum_{k=1}^{3} \hat{B}_k = \begin{pmatrix} -\frac{1}{2}z\sum_{k=1}^{3}\gamma_k^2 & -1 \\ 1 & -\frac{1}{2}z\sum_{k=1}^{3}\gamma_k^2 \end{pmatrix}. \tag{55}$$

Hence $\hat{\mathscr{D}}^{-1}$ exists in a neighborhood of $z = 0$ which proves the existence of a $z_{\max} > 0$. Obviously, we wish to maximize $z_{\max}$. For $z > 0$ follows that $\hat{\mathscr{D}}^{-1} = z(I - \hat{\mathscr{R}})^{-1}$ exists if both eigenvalues of $\hat{\mathscr{R}}$ are unequal one. This is true inside the whole stability interval, where the eigenvalues lie on the unit circle, but for $z \rightarrow$ the right end point of the stability interval the eigenvalues coincide in one. Necessarily we thus have $z_{max} <$ than the right endpoint. With $z_{max} = \pi/2$ we can conclude that $\hat{\mathscr{D}}^{-1}$ exists and is bounded uniformly in $[0, z_{max}]$, because $\pi/2$ is smaller than the true end point.[7]
This completes our proof of assertion (i).

3.2.3 Assertion (ii)

If $f^u, f^v \in C^4[0,T]$, then $u_h, v_h \in C^5[0,T]$. For $\tau \sim h, h \rightarrow 0$, Taylor's theorem with remainder then allows us to replace (40) by

$$\delta_n^{(k)} = \gamma_k^3 \begin{pmatrix} -\frac{1}{12}\tau^3 u_h^{(3)} - \frac{1}{12}\left(s_k - t_{n+1/2}\right)\tau^4 u_h^{4)} + (\mathscr{O}(\tau^5) \\ \frac{1}{6}\tau^3 v_h^{(3)} + \frac{1}{6}\left(s_k - t_{n+1/2}\right)\tau^4 v_h^{(4)} + \mathscr{O}(\tau^5) \end{pmatrix}, \tag{56}$$

where the derivatives are taken at $t_{n+1/2} = s_k - (\gamma_1 + \cdots + \gamma_{k-1} + \frac{1}{2}\gamma_k - \frac{1}{2})\tau$. Note that due to symmetry $s_2 = t_{n+1/2}$ and $s_3 - s_2 = s_2 - s_1$. As a consequence

$$\delta_n^{(1)} + \delta_n^{(2)} + \delta_n^{(3)} = \mathscr{O}(\tau^5)\,. \tag{57}$$

Alternatively, the $\mathscr{O}(\tau^5)$ result can also be concluded from the quadrature order four, since (57) is the local error for zero K for which the composition method reduces to a 4th-order quadrature rule.

[7] For $z \rightarrow z_{max}$, $\|\hat{\mathscr{D}}^{-1}\|_2$ monotonically increases but remains close to one (the value at $z = 0$) on the greatest part of the interval. For example, at the values $(0.50, 0.75, 0.90, 1.00) \cdot \pi/2$ the norm equals, approximately, $1.08, 1.48, 3.09, 189.9$.

Proceeding with transformed variables we thus can express the local error as $\hat{\rho}_n = z\,\hat{\beta}_n + \mathcal{O}(\tau^5)$,

$$\begin{aligned}\hat{\beta}_n = {} & [\hat{B}_3 + \hat{B}_2 + \hat{A}_1 + z\left(\hat{B}_3\hat{B}_2 + \hat{B}_3\hat{A}_1 + \hat{B}_2\hat{A}_1\right) + z^2\hat{B}_3\hat{B}_2\hat{A}_1]\,\hat{\delta}_n^{(1)} + \\ & [\hat{B}_3 + \hat{A}_2 + z\hat{B}_3\hat{A}_2]\,\hat{\delta}_n^{(2)} + \hat{A}_3\,\hat{\delta}_n^{(3)}\,,\end{aligned} \tag{58}$$

and our task is now to check the Ansatz rule (47) for $p = 4$. Obviously we assign $\hat{\eta}_n$ to the $\mathcal{O}(\tau^5)$ term and we are done if in the interval $[0, z_{max}]$ we can solve $\hat{\xi}_n$ with order $p = 4$ from

$$(I - \hat{\mathscr{R}})\hat{\xi}_n = z\,\hat{\beta}_n\,, \tag{59}$$

or, equivalently, from $\hat{\mathscr{D}}\hat{\xi}_n = \hat{\beta}_n$, see (53) and the discussion thereafter on the existence and uniform boundedness of $\hat{\mathscr{D}}^{-1}$. Hence what remains to show is that $\hat{\beta}_n = \mathcal{O}(\tau^4)$.

From (56) follows

$$\begin{aligned}\hat{\beta}_n = {} & \gamma_1^3\,[\hat{B}_3 + \hat{B}_2 + \hat{A}_1 + z\left(\hat{B}_3\hat{B}_2 + \hat{B}_3\hat{A}_1 + \hat{B}_2\hat{A}_1\right) + z^2\hat{B}_3\hat{B}_2\hat{A}_1]\,\hat{w}_n + \\ & \gamma_2^3\,[\hat{B}_3 + \hat{A}_2 + z\hat{B}_3\hat{A}_2]\,\hat{w}_n + \gamma_3^3\hat{A}_3\,\hat{w}_n + \mathcal{O}(\tau^4) \\ = {} & \left(\gamma_1^3\,[\hat{B}_3 + \hat{B}_2 + \hat{A}_1] + \gamma_2^3\,[\hat{B}_3 + \hat{A}_2] + \gamma_3^3\hat{A}_3\right)\,\hat{w}_n + \hat{\mathscr{T}}\cdot z\,\hat{w}_n + \mathcal{O}(\tau^4)\,,\end{aligned} \tag{60}$$

where $\hat{\mathscr{T}}$ collects remaining $\mathcal{O}(1)$ terms and

$$\hat{w}_n = \begin{pmatrix} -\frac{1}{12}\tau^3\,\hat{u}_h^{(3)}(t_{n+1/2}) \\ \frac{1}{6}\tau^3\,\hat{v}_h^{(3)}(t_{n+1/2}) \end{pmatrix}. \tag{61}$$

Recall that the $\hat{A}_k, \hat{B}_k$ and their combinations are $\mathcal{O}(1)$ since $z \in [0, z_{max}]$ with z_{max} finite.

At this stage we invoke the additional condition $K^T u_h^{(3)}(t), K v_h^{(3)}(t) = \mathcal{O}(1), h \to 0$ made for assertion (ii). For the transformed variables this implies, for $h \to 0$,

$$\lambda\,\hat{u}_h^{(3)}(t),\ \lambda\,\hat{v}_h^{(3)}(t) = \mathcal{O}(1)\,, \tag{62}$$

for any component pair $\hat{u}_h, \hat{v}_h$ and occurring singular value λ of K. This provides us with an additional factor τ such that $\hat{\mathscr{T}}\cdot z\,\hat{w}_n = \mathcal{O}(\tau^4)$ and likewise we can simplify expression (60) to

$$\hat{\beta}_n = \left(\gamma_1^3\,[\hat{B}_3 + \hat{B}_2 + \hat{A}_1] + \gamma_2^3\,[\hat{B}_3 + \hat{A}_2] + \gamma_3^3\hat{A}_3\right)\,\hat{w}_n + \mathcal{O}(\tau^4)\,. \tag{63}$$

Continuing this we find

$$\hat{\beta}_n = \begin{pmatrix} 0 & -\gamma \\ \gamma & 0 \end{pmatrix}\hat{w}_n + \mathcal{O}(\tau^4)\,, \quad \gamma = \gamma_1^3(\gamma_3 + \gamma_2 + \frac{1}{2}\gamma_1) + \gamma_2^3(\gamma_3 + \frac{1}{2}\gamma_2) + \gamma_3^3\cdot\frac{1}{2}\gamma_3\,, \tag{64}$$

and since $\gamma = 0$ we have proved that $\hat{\beta}_n = \mathcal{O}(\tau^4)$ which completes the proof of assertion (ii).

4 Numerical illustration

In this section we illustrate the results of Theorems 3.1, 3.2 for the parameter sets (26), (27).

4.1 The test model class

Let μ, ε, σ in (1) be scalar. Writing $E = (E^x, E^y, E^z)$, etc., in three dimensions we then have

$$\begin{aligned}
\mu\,\frac{\partial H^x}{\partial t} &= \frac{\partial E^y}{\partial z} - \frac{\partial E^z}{\partial y}\,, \qquad & \varepsilon\,\frac{\partial E^x}{\partial t} &= \frac{\partial H^z}{\partial y} - \frac{\partial H^y}{\partial z} - \sigma E^x - J_E^x\,,\\
\mu\,\frac{\partial H^y}{\partial t} &= \frac{\partial E^z}{\partial x} - \frac{\partial E^x}{\partial z}\,, \qquad & \varepsilon\,\frac{\partial E^y}{\partial t} &= \frac{\partial H^x}{\partial z} - \frac{\partial H^z}{\partial x} - \sigma E^y - J_E^y\,,\\
\mu\,\frac{\partial H^z}{\partial t} &= \frac{\partial E^x}{\partial y} - \frac{\partial E^y}{\partial x}\,, \qquad & \varepsilon\,\frac{\partial E^z}{\partial t} &= \frac{\partial H^y}{\partial x} - \frac{\partial H^x}{\partial y} - \sigma E^z - J_E^z\,.
\end{aligned} \tag{65}$$

From this 3D model we derive the 2D (transversal magnetic) model with components H^x, H^z, E^y:

$$\begin{aligned}
\frac{\partial H^x}{\partial t} &= \frac{\partial E^y}{\partial z}\,,\\
\frac{\partial H^z}{\partial t} &= -\frac{\partial E^y}{\partial x}\,,\\
\frac{\partial E^y}{\partial t} &= \frac{\partial H^x}{\partial z} - \frac{\partial H^z}{\partial x} - J_E^y\,,
\end{aligned} \tag{66}$$

where we have put $\mu = \varepsilon = 1$ and $\sigma = 0$. As space domain we take the unit square $0 < x, z < 1$. We suppose initial conditions for E^y, H^x, H^z and Dirichlet boundary conditions for E^y only, which is natural since E^y satisfies the second-order wave equation

$$\frac{\partial^2 E^y}{\partial t^2} = \frac{\partial^2 E^y}{\partial^2 x} + \frac{\partial^2 E^y}{\partial^2 z} - \frac{\partial J_E^y}{\partial t}\,, \tag{67}$$

and uniquely determines H^x, H^z.

4.2 Spatial discretization

For spatial discretization we use a uniform grid with grid size $h = 1/m$, staggering, and 2nd-order central-difference discretization. Let $x_i = ih, x_{i+1/2} = (i + 1/2)h$, etc. Then, E^y is approximated at (x_i, z_j) for $i, j = 1(1)m-1$, H^x at $(x_i, z_{j+1/2})$ for $i = 1(1)m-1$ and $j = 0(1)m-1$, and H^z at $(x_{i+1/2}, z_j)$ for $i = 0(1)m-1$ and $j = 1(1)m-1$. This spatial discretization yields a semi-discrete system that fits in format (4) with u of length $2m(m-1)$ and v of length $(m-1)(m-1)$, see [13] for details. Note that the staggering accommodates our boundary condition, because due to the staggering H^x and H^z are not required at the domain boundary, with the benefit that always $f^v(t) = \mathscr{O}(1)$, whereas either $f^u(t) = \mathscr{O}(h^{-1})$ or $f^u(t) = 0$, depending on whether a time-dependent Dirichlet condition is chosen for component E^y or not. Hence, with our staggering, starting with component u is profitable because then always $f^v(t) = \mathscr{O}(1)$. For illustration purposes, however, we will use all four base methods mentioned in Section 2.3, including the reversed sequence methods.

For this spatial discretization, the maximum singular value λ_1 from (33) equals $2\sqrt{2}/h$. This leads for method (25) to the step size restrictions

$$\tau \le \tau_c = \begin{cases} \dfrac{\pi}{4\sqrt{2}}h \;\approx\; 0.555 \cdot h\,, \quad & s = 3\,,\\ \dfrac{e}{2\sqrt{2}}h \;\approx\; 0.961 \cdot h\,, \quad & s = 5\,. \end{cases} \tag{68}$$

In the tests we will use the critical step size values τ_c. However, to account for the different numbers of stages in the convergence plots, accuracy will be plotted against the total numbers of stages.

4.3 Two test solutions

4.3.1 Test solution one

As a first test solution we impose the artificial functions

$$\begin{aligned} E^y &= e^t(x-a)(x-b)z(1-z)\,, \\ H^x &= e^t(x-a)(x-b)(1-2z)\,, \\ H^z &= -e^t(2x-a-b)z(1-z)\,, \end{aligned} \tag{69}$$

with $(a,b) = (0,1)$ or $(a,b) = (0.5,0.5)$. With $(a,b) = (0,1)$ we have E^y zero at the boundary, and thus $f^u(t) = 0$. For $(a,b) = (0.5,0.5)$ we have E^y nonzero at the $x = 0,1$ boundary and thus $f^u(t) = \mathscr{O}(h^{-1})$. For both choices the source function $f^v(t) = \mathscr{O}(1)$ and nonzero as determined by J_E^y. In space the solution is quadratic and hence we have a zero spatial error. As integration interval we have used $[0,1]$. Convergence plots are given in Figure 2, where, for a sequence of decreasing values of h, the maximum norm global error of u_N, v_N for $N\tau = 1$ is plotted against the total number of stages Ns. As step size the critical value $\tau = \tau_c$ given in (68) is used.

In the left plot (the zero boundary case) the o-and $*$-marker refer to $s = 3$ and $s = 5$, respectively. The solid lines refer to the base scheme (22) using the perturbation for the source function $f^v(t)$. These solid lines confirm assertion (ii) of Theorem 3.1 on order four (the parallel lower dashed line has slope four). The dash-dotted lines refer to the base scheme (7) not using the perturbation. These dash-dotted lines confirm assertion (ii) of Theorem 3.2 on order three (the parallel upper dashed line has slope three). Note that the composition scheme yields smaller errors for $s = 5$ than for $s = 3$. This was expected due to the smaller γ_k-parameters and the nearly equal scaled critical step sizes τ_c/s.

In the right plot (the nonzero boundary case) we give results for $s = 5$ only. Because we have a nonzero Dirichlet boundary condition, we expect to obtain maximal convergence order three. The results confirm this. The three solid lines with the $*, \square, \diamond$-markers all three represent a third-order convergence result (the parallel lower dashed line has slope three). The $*$-marker corresponds with the base method (22) using the perturbation, and the $\square$-marker with the base method (7) without the perturbation. The fact that both methods lead to order three is in line with assertion (i) from Theorem 3.1 and assertion (ii) from Theorem 3.2. In other words, with nonzero Dirichlet boundary values contained in $f^u(t)$ the perturbation has no effect on the convergence order. However, this changes if the sequence in treating u, v is reversed. The $\diamond$-marker corresponds with method (24) where the reversed sequence v, u is used, with in addition the perturbation applied to $f^u(t)$. This order-three result is in line with assertion (i) of Theorem 3.1 and is clearly the most accurate one. The $+$-marker along the dash-dotted line corresponds with method (23), that is also with reversed sequence but without the perturbation. In line with assertion (i) of Theorem 3.2 this case reveals only order two (the dashed upper line has slope two). So this case illustrates that we can loose two orders if we consider convergence in the PDE sense compared to the order in the ODE sense.

To sum up, albeit contrived, the current test solution confirms the order reduction predicted by Theorem 3.1 and 3.2. On the other hand, when using the source term perturbation as in methods (22) and (24), the obtained accuracies are high. In this regard we expect that in general the smaller $s = 5$-parameters will be competitive with the $s = 3$-parameters.

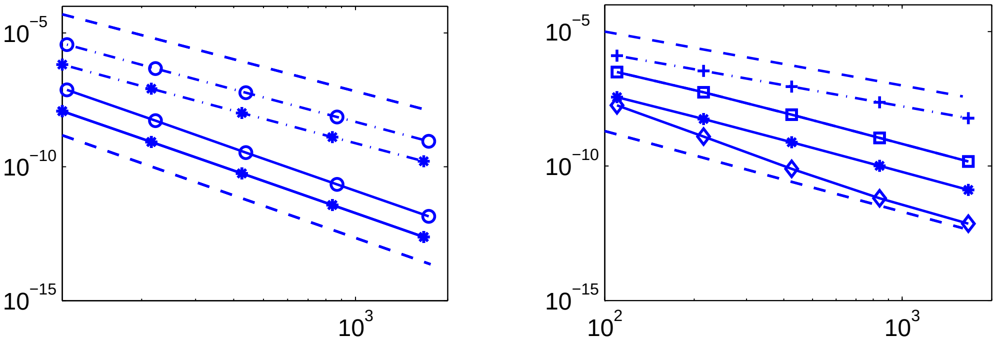

Figure 2: Convergence plots for test solution one. See Section 4.3.1 for explanations.

4.3.2 Test solution two

The second test solution is the eigenmode

$$\begin{aligned} H^x &= \frac{k_z}{\sqrt{k_x^2+k_z^2}} \sin(k_x\pi x + s_x\pi/2)\cos(k_z\pi z)\sin\left(\sqrt{k_x^2+k_z^2}\,\pi t\right), \\ H^z &= \frac{-k_x}{\sqrt{k_x^2+k_z^2}} \cos(k_x\pi x + s_x\pi/2)\sin(k_z\pi z)\sin\left(\sqrt{k_x^2+k_z^2}\,\pi t\right), \qquad (70) \\ E^y &= \sin(k_x\pi x + s_x\pi/2)\sin(k_z\pi z)\cos\left(\sqrt{k_x^2+k_z^2}\,\pi t\right), \end{aligned}$$

where we fix $k_x = k_z = 2$ and take as an option $s_x = 0$ or $s_x = 1$ in order to impose, respectively, a zero and nonzero Dirichlet boundary condition for component E^y. So for $s_x = 0$ we have $f^u(t) = 0$, whereas for $s_x = 1$ the source function $f^u(t) = \mathscr{O}(h^{-1})$. Further, both options result in $f^v(t) = 0$ as $J_E^y = 0$.

While we discuss temporal order p up to four, the chosen spatial discretization yields only 2nd-order convergence for the spatial error, see e.g. [8]. In the tests we have therefore applied standard Richardson extrapolation in space to the E^y-approximations to lift the spatial order to four for error measuring at the output time. Let $v_{N;2h}$ denote the v_n obtained at the output time $T = N\tau = 1$ with grid size $2h$. Similarly, let $v_{N;h\to 2h}$ denote the v_n obtained with grid size h and restricted to the $2h$-grid. Then, at the output time we measure the PDE error for E^y at the $2h$-grid by [8)]

$$v_{2h}(T) - \left(\frac{4}{3}v_{N;h\to 2h} - \frac{1}{3}v_{N;2h}\right) = \mathscr{O}(\tau^p) + \mathscr{O}(h^4). \qquad (71)$$

Convergence plots are given in Figure 3, where, for a sequence of decreasing values of h, the maximum norm of this PDE error for $N\tau = 1$ is plotted against the total number of stages Ns. As step size again the critical value $\tau = \tau_c$ given in (68) is used.

In the left plot the o-and $*$-marker refer to $s = 3$ and $s = 5$, respectively. The solid and dash-dotted lines refer, respectively, to the zero- and nonzero boundary case. For the zero case, where we

[8)] Because we extrapolate only at the output time, the integration methods are not changed. This would be the case with extrapolation after every step. Extrapolation at the output time only serves our purpose of testing here. We do not advocate it over long time intervals for wave equations without damping. See also [1] and [5] for comments on this issue regarding extrapolation in time. Higher spatial orders are better achieved with spatial discretization techniques such as based on the discontinuous Galerkin method, see e.g. [3] and references therein.

have no source terms, we see a straight order four (the parallel dashed line has slope four) with again more accurate results for $s = 5$. According to Theorem 3.1, we expected to see order three for the nonzero case because then $f^u(t) = \mathscr{O}(h^{-1})$. However, while the errors zigzag slightly as shown by the two dash-dotted lines lying between the two solid ones, overall we see order four. We probably owe this to fortunate error cancelation emanating from the oscillatory nature of the solution.

In the right plot the o-and $*$-marker again refer to $s = 3$ and $s = 5$, respectively. Here we treat only the nonzero boundary case and reverse the u, v sequence. The solid lines refer to (24) with the source term perturbation (17) now applied to $f^u(t)$. The dash-dotted lines refer to (23) which does not employ this perturbation. Without the perturbation we find order two (the upper dashed line has slope two) in accordance with assertion (i) of Theorem 3.2, while again the $s = 5$ method is notably more accurate. With the perturbation we expected to see order three in accordance with assertion (i) of Theorem 3.1. The order turns out to lie between three and four (the lower dashed line has slope three). Like in the left plot, we probably owe this to fortunate error cancelation emanating from the oscillatory nature of the solution. Note that with the perturbation, $s = 3$ and $s = 5$ now yield the same accuracy (the two solid lines nearly coincide).

On the other hand, similar as for test solution one, we can conclude that the idea of perturbing the source function works out very well. We therefore anticipate that for many Maxwell applications the composition method (25) based on method (22) or method (24) provides an efficient integration method when high accuracy is in demand. In particular the parameter set (27) for $s = 5$ due to [10] is then most attractive.

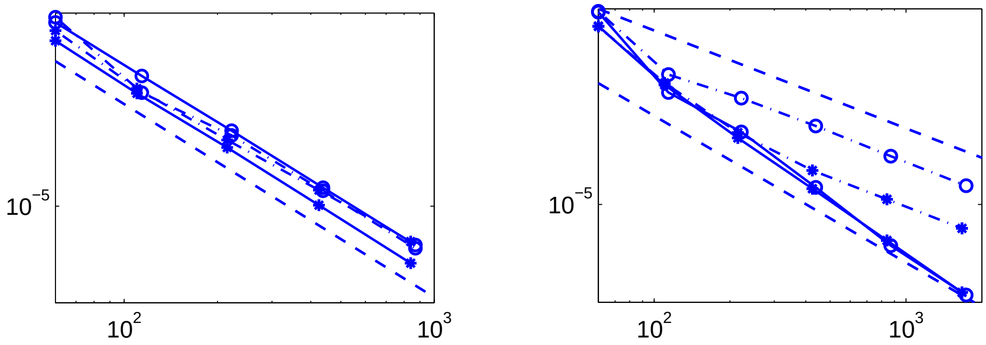

Figure 3: Convergence plots for test solution two. See Section 4.3.2 for explanations.

References

[1] Botchev, M.A., Verwer, J.G.: Numerical integration of damped Maxwell equations. SIAM J. Sci. Comput. 31, 1322–1346 (2009)

[2] Creutz, M., Gocksch, A.: Higher-order Monte Carlo algorithms. Phys. Rev. Lett. 63, 9–12 (1989)

[3] Fahs, H., Lanteri, S.: A high-order non-conforming discontinuous Galerkin method for time-domain electromagnetics. J. Comput. Appl. Math. 234, 1088–1096 (2010)

[4] Forest, E.: Canonical integrators as tracking codes. AIP Conference Proceedings 184, 1106–1136 (1989)

[5] Fornberg, B., Zuev, J., Lee, J.: Stability and accuracy of time-extrapolated ADI-FDTD methods for solving wave equations. J. Comp. Appl. Math. 200, 178–192 (2007)

[6] Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Springer Series in Computational Mathematics, Vol. 31 (second edition), Springer (2006)

[7] Hundsdorfer, W., Verwer, J.G.: Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations. Springer Series in Computational Mathematics, Vol. 33, Springer (2003)

[8] Monk, P., Süli, S.: A convergence analysis of Yee's scheme on nonuniform grids. SIAM J. Numer. Anal. 31, 393–412 (1994)

[9] Rodrigue, G., White, D.: A vector finite element time-domain method for solving Maxwell's equations on unstructured hexahedral grids. SIAM J. Sci. Comput. 23, 683–706 (2001)

[10] Suzuki, M.: Fractal decomposition of exponential operators with applications to many-body theories and Monte-Carlo simulations. Phys. Lett. A 146, 319–323 (1990)

[11] Verlet, L.: Computer ”experiments” on classical fluids. I. Thermodynamical properties of Lennart-Jones molecules. Physical Review 159, 98–103 (1967)

[12] Verwer, J.G., Botchev, M.A.: Unconditionally stable integration of Maxwell's equations. Linear Algebra and its Applications 431, 300–317 (2009)

[13] Verwer, J.G.: Component splitting for semi-discrete Maxwell equations. BIT Numer Math, DOI 10.1007/s10543-010-0296-y (2010)

[14] Yee, K.S.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. IEEE Trans. Antennas Propag. 14, 302–307 (1966)

[15] Yoshida, H.: Construction of higher order symplectic integrators. Phys. Lett. A 150 262–268 (1990)