ESAIM: M2AN 51 (2017) 209–223
DOI: 10.1051/m2an/2016061

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

OPTIMIZED WAVEFORM RELAXATION METHODS FOR RC CIRCUITS: DISCRETE CASE

Shu-Lin Wu[1] and Mohammad D. Al-Khaleel[2,3]

Abstract. The optimized waveform relaxation (OWR) methods, benefiting from intelligent information exchange between subsystems – the so-called *transmission conditions* (TCs), are recognized as efficient solvers for large scale circuits and get a lot of attention in recent years. The TCs contain a free parameter, namely α, which has a significant influence on the convergence rates. So far, the analysis of finding the best parameter is merely performed at the continuous level and such an analysis does not take into account the influence of temporal discretizations. In this paper, we show that the temporal discretizations do have an important effect on the OWR methods. Precisely, for the Backward–Euler method, compared to the parameter $\alpha^c_{\rm opt}$ from the continuous analysis, we show that the convergence rates can be further improved by using the one $\alpha^d_{\rm opt}$ analyzed at the discrete level, while for the Trapezoidal rule, it is better to use $\alpha^c_{\rm opt}$. This conclusion is confirmed by numerical results.

Mathematics Subject Classification. 65L12, 65L20, 65B99.

Received January 17, 2015. Revised August 28, 2016. Accepted September 8, 2016.

1. Introduction

In 2004, Gander and Ruehli proposed a new class of waveform relaxation (WR) methods [5], which are found very efficient for large scale circuit simulations. Instead of directly partitioning the coefficient matrix of the circuit equations – the basic feature of the *classical* WR methods [14, 15, 17], the new WR methods consist of direct circuit partitions and therefore the partitioning procedure is much simpler. The new methods are called *optimized* WR methods, where the *optimization* concerns are what we call the transmission conditions (TCs). The function of the TCs is to transmit information from each subcircuit to its connected neighbor subcircuits. Nowadays, it is well-understood that there is a close relevance between the OWR technique and the so-called Schwarz waveform relaxation methods for time-dependent PDEs, where the TCs in the OWR framework correspond to the coupling between subdomains in physical space. The classical WR methods correspond to the coupling of Dirichlet type between subdomains [6, 13] and therefore these methods often converge very slowly, because the TCs of Dirichlet type are inefficient. More efficient coupling can be obtained if more appropriate information, adapted to the physics of the underlying PDE problem, is exchanged [7,8,10]. Among these studies, the coupling of Robin type attracts a lot of attention in recent years, because the convergence rates of the resulting WR methods are much more satisfactory, compared to the classical WR methods.

Keywords and phrases. Waveform relaxation (WR), discretization, parameter optimization, RC circuits.

[1] Sichuan University of Science and Engineering, Zigong, Sichuan 643000, P.R. China. wushulin_ylp@163.com

[2] Department of Mathematics, Yarmouk University, 21163 Irbid, Jordan. khaleel@yu.edu.jo

[3] Department of Mathematics and Sciences, Khalifa University, 127788 Abu Dhabi, UAE. mohammad.alkhaleel@kustar.ac.ae

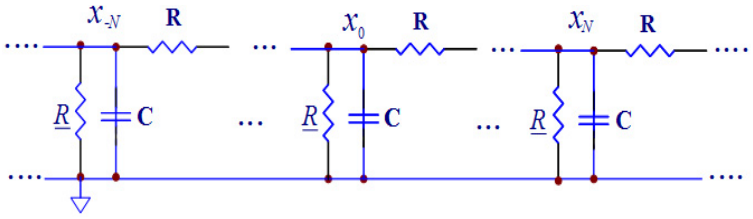

FIGURE 1. The infinite-size RC circuits.

The Robin TCs contain a free parameter, namely α, which has a significant effect on the convergence rate. The optimization procedure is investigated numerically in [2, 5, 9, 11], and theoretically in [1, 3, 4, 12]. All of these previous studies are performed at the continuous level and therefore the effect of temporal discretizations on the convergence behavior is not taken into account. Note that, in a real computation, we need to apply some time-integrator to the continuous OWR methods and the solutions are obtained through discrete OWR iterations. Hence, it is important to understand the influence of temporal discretizations on the convergence rate. Actually, as we will see in this paper, the temporal discretizations do have a remarkable influence. Precisely, for the Backward–Euler method, we find that the performance of the OWR methods can be further improved by using the parameter α^d_{opt} analyzed at the discrete level, compared to the one α^c_{opt} from the continuous analysis, while for the Trapezoidal rule it is better to use α^c_{opt}, instead of α^d_{opt}.

The layout of this paper is organized as follows: in Section 2, we introduce the model circuits and the OWR methods studied in this paper. In Section 3, we present the analysis of finding the best parameter α^d_{opt} at the discrete level. Section 4 presents the asymptotic dependence of the discrete OWR methods on Δt (the step size) and T (the length of time interval). Section 5 provides numerical results to validate the theoretical analysis and we finish this paper in Section 6 with conclusions.

2. Model circuits and discrete OWR

To make our narrative clear and concise, similar to ([4,5] and [1], Chap. 3), we continue to use the RC circuits in infinite-size as our model:

The state equation of this model circuit is

$$\mathbf{x}'(t) = \begin{pmatrix} \ddots & \ddots & \ddots & & \\ & d & a & d & \\ & & d & a & d \\ & & & \ddots & \ddots & \ddots \end{pmatrix} \mathbf{x}(t) + \mathbf{f}(t), \tag{2.1a}$$

where $\mathbf{f}(t) = (\dots, f_{-1}(t), f_0(t), f_1(t), \dots)^\top$, and $\mathbf{x}(t) = (\dots, \mathbf{x}_{-1}(t), \mathbf{x}_0(t), \mathbf{x}_1(t), \dots)^\top$ denotes a set of nodal voltage values. The quantities a and d are determined by the circuit parameters, as,

$$d = \frac{1}{RC}, \quad a = -\left(2d + \frac{1}{\underline{R}C}\right). \tag{2.1b}$$

One can see that $d > 0$, $a < 0$ and $-a \geq 2d$, which makes our analysis in the following applicable to general RC circuits. Since the circuit is infinitely large, to have a well posed problem, we assume that all voltage values stay bounded as we move toward the infinite ends of the circuit.

To introduce the OWR methods, we divide the vector $\mathbf{x}(t)$ into two overlapping subvectors: $\tilde{\mathbf{x}}_1(t) = (\dots, \mathbf{x}_{-1}(t), \mathbf{x}_0(t), \mathbf{x}_1(t))^\top$ and $\tilde{\mathbf{x}}_2(t) = (\mathbf{x}_0(t), \mathbf{x}_1(t), \mathbf{x}_2(t), \dots)^\top$. Then, similar to [4, 5, 12] and ([1], Chap. 3),

we consider the following OWR method:

$$\begin{aligned}\left(\tilde{\mathbf{x}}_{1,m}^{k}\right)'(t)-\mathrm{d}\Delta_m\tilde{\mathbf{x}}_{1,m}^{k}(t)&=f_m(t)\ \text{for}\ m<1,\\ \left(\tilde{\mathbf{x}}_{2,m}^{k}\right)'(t)-\mathrm{d}\Delta_m\tilde{\mathbf{x}}_{2,m}^{k}(t)&=f_m(t)\ \text{for}\ m>0,\end{aligned}\tag{2.2a}$$

together with transmission conditions (TCs)

$$\alpha\tilde{\mathbf{x}}_{1,1}^{k}(t)-\tilde{\mathbf{x}}_{1,0}^{k}(t)=\alpha\tilde{\mathbf{x}}_{2,1}^{k-1}(t)-\tilde{\mathbf{x}}_{2,0}^{k-1}(t),\quad \alpha\tilde{\mathbf{x}}_{2,0}^{k}(t)-\tilde{\mathbf{x}}_{2,1}^{k}(t)=\alpha\tilde{\mathbf{x}}_{1,0}^{k-1}(t)-\tilde{\mathbf{x}}_{1,1}^{k-1}(t),\tag{2.2b}$$

where $k\geq 1$ is the iteration index, $\alpha\in\mathbb{R}$ is a free parameter and $\Delta_m\tilde{\mathbf{x}}_{j,m}=\tilde{\mathbf{x}}_{j,m-1}-2\zeta\tilde{\mathbf{x}}_{j,m}+\tilde{\mathbf{x}}_{j,m+1}$ with $\zeta=-a/(2d)\geq 1$ and $j=1,2$. Applying the linear θ-method to (2.2) gives

$$\begin{aligned}\frac{\tilde{\mathbf{x}}_{1,m}^{k}(n)-\tilde{\mathbf{x}}_{1,m}^{k}(n-1)}{\Delta t}-\mathrm{d}\Delta_m\left[\theta\tilde{\mathbf{x}}_{1,m}^{k}(n)+(1-\theta)\tilde{\mathbf{x}}_{1,m}^{k}(n-1)\right]&=\bar{f}_m(n),\ m<1,\\ \frac{\tilde{\mathbf{x}}_{2,m}^{k}(n)-\tilde{\mathbf{x}}_{2,m}^{k}(n-1)}{\Delta t}-\mathrm{d}\Delta_m\left[\theta\tilde{\mathbf{x}}_{2,m}^{k}(n)+(1-\theta)\tilde{\mathbf{x}}_{2,m}^{k}(n-1)\right]&=\bar{f}_m(n),\ m>0,\end{aligned}\tag{2.3a}$$

together with the following discrete TCs at $t=t_n$:

$$\alpha\tilde{\mathbf{x}}_{1,1}^{k}(n)-\tilde{\mathbf{x}}_{1,0}^{k}(n)=\alpha\tilde{\mathbf{x}}_{2,1}^{k-1}(n)-\tilde{\mathbf{x}}_{2,0}^{k-1}(n),\quad \alpha\tilde{\mathbf{x}}_{2,0}^{k}(n)-\tilde{\mathbf{x}}_{2,1}^{k}(n)=\alpha\tilde{\mathbf{x}}_{1,0}^{k-1}(n)-\tilde{\mathbf{x}}_{1,1}^{k-1}(n),\tag{2.3b}$$

where $n\geq 1$ and $\bar{f}_m(n)=\theta f_m(n)+(1-\theta)f_m(n-1)$. In (2.3a) and (2.3b), $\tilde{\mathbf{x}}_{j,m}^{k}(n)$ denotes the numerical approximation of $\tilde{\mathbf{x}}_{j,m}^{k}(t)$ at $t=t_n$. To analyze the discrete WR iteration (2.3), we use the discrete Laplace transform [18]. For any grid function $v=\{v_n\}_{n\geq 0}$ on a regular grid with time step Δt, the discrete Laplace transform is defined by:

$$\mathcal{L}(v)=\hat{v}(s)=\frac{\Delta t}{\sqrt{2\pi}}\sum_{n\geq 0}e^{-sn\Delta t}v_n\ \text{with}\ s=\sigma+i\omega\ \text{and}\ \pi/T\leq|\omega|\leq\pi/\Delta t,\tag{2.4}$$

where $\sigma>0$. The following lemma is useful for finding the best choice of the parameter α.

Lemma 2.1 (Minimizing-procedure)**.** *Let $J\geq 2$ be an integer and $g_j(x)$ be a continuous function, monotonically decreasing for $x\in[a,x_j^*]$ and increasing for $x\in[x_j^*,b]$, $j=1,2,\ldots,J$. Let $G_1(x)=g_1(x)$ and $X_1^*=x_1^*$. Define*

$$G_j(x)=\max\{G_{j-1}(x),g_j(x)\},\ X_j^*=\begin{cases}X_{j-1}^*, & \text{if}\ \ G_{j-1}(X_{j-1}^*)\geq g_j(X_{j-1}^*),\\ x_j^*, & \text{if}\ \ g_j(x_j^*)\geq G_{j-1}(x_j^*),\\ \tilde{x}_j^*, & \text{otherwise,}\end{cases}$$

where $j=2,3,\ldots,J$ and $\tilde{x}_j^*$ is the unique root of $G_{j-1}(x)=g_j(x)$ lying between X_{j-1}^* and x_j^*. Then, the quantity $X^*:=X_J^*$ is the unique local minimizer of $G(x):=\max_{1\leq j\leq J}\{g_j(x)\}$, i.e.*, $G(x)$ is monotonically decreasing for $x\in[a,X^*]$ and increasing for $x\in[X^*,b]$.*

Proof. For $j=2$, by assumption, we have $G_2(x)\geq\max\{g_1(x_1^*),g_2(x_2^*)\}$. If $g_1(x_1^*)\geq g_2(x_1^*)$, we have $G_2(x)\geq g_1(x_1^*)$ since $g_2(x_1^*)\geq g_2(x_2^*)$. On the other hand, the low bound $g_1(x_1^*)$ is reachable: $G_2(x_1^*)=\max\{g_1(x_1^*),g_2(x_1^*)\}=g_1(x_1^*)$. Hence, $X_2^*=x_1^*$. Similarly, for the case $g_2(x_2^*)\geq g_1(x_2^*)$, we shall have $X_2^*=x_2^*$. It remains to consider the third case, *i.e.*, the previous two cases do not hold. Without lose of generality, we assume $x_1^*<x_2^*$. Then, it is easy to understand that $G_2(x)$ is decreasing for $x\in[a,x_1^*]$ and increasing for $x\in[x_2^*,b]$. In the middle interval $x\in[x_1^*,x_2^*]$, we know that $g_1(x)$ is increasing and $g_2(x)$ is decreasing. This, together with $g_1(x_1^*)<g_2(x_1^*)$ and $g_2(x_2^*)<g_1(x_2^*)$, implies that $g_1(x)=g_2(x)$ has a unique root in the interval $[x_1^*,x_2^*]$ and that $G_2(x)=g_2(x)$ for $x\in[x_1^*,X^*]$ and $G_2(x)=g_1(x)$ for $x\in[X^*,x_2^*]$.

In summary, $G_2(x)$ has a unique local minimizer X_2^* and $G_2(x)$ is decreasing for $x\in[a,X_2^*]$ and increasing for $x\in[X_2^*,b]$. Now, the function $G_2(x)$ has the similar property of $g_2(x)$ and therefore for $j=3$ the quantity X_3^* is the unique local minimizer of $G_3(x)$ and $G_3(x)$ is decreasing for $x\in[a,X_3^*]$ and increasing for $x\in[X_3^*,b]$. Repeating this process, we will arrive at X_J^*, the unique local minimizer of $G(x)$. □

3. Parameter optimizations

Denote by $\tilde{\mathbf{x}}_{j,m}(n)$ (with $j = 1, 2$) the converged solution of the OWR method. Then, the error sequences $\mathbf{e}^k_{j,m}(n) := \tilde{\mathbf{x}}^k_{j,m}(n) - \tilde{\mathbf{x}}_{j,m}(n)$ also satisfy (2.3a) and (2.3b), but with $\bar{f}_m(n) = 0$ and $\mathbf{e}^k_{j,m}(0) = 0$. Applying the discrete Laplace transform to the error equations, after some simple algebra, yields:

$$\hbar\hat{\mathbf{e}}^k_{j,m} - \gamma\Delta_m\hat{\mathbf{e}}^k_{j,m} = 0,\ j = 1, 2, \tag{3.1a}$$

$$\alpha\hat{\mathbf{e}}^k_{1,1} - \hat{\mathbf{e}}^k_{1,0} = \alpha\hat{\mathbf{e}}^{k-1}_{2,1} - \hat{\mathbf{e}}^{k-1}_{2,0},\ \alpha\hat{\mathbf{e}}^k_{2,0} - \hat{\mathbf{e}}^k_{2,1} = \alpha\hat{\mathbf{e}}^{k-1}_{1,0} - \hat{\mathbf{e}}^{k-1}_{1,1}, \tag{3.1b}$$

where

$$\gamma = \mathrm{d}\Delta t,\ z = e^{s\Delta t},\ \hbar = \frac{z-1}{\theta z + (1-\theta)}. \tag{3.1c}$$

It is well-known that the general form of the solution $\hat{\mathbf{e}}^k_{j,m}$ can be expressed as

$$\hat{\mathbf{e}}^k_{j,m}(s) = A^k_j\lambda^m_+ + B^k_j\lambda^m_-,\ j = 1, 2, \tag{3.2a}$$

where $\lambda_\pm$ is defined by:

$$\lambda_\pm = \frac{\hbar + 2\zeta\gamma \pm \sqrt{(\hbar + 2\zeta\gamma)^2 - 4\gamma^2}}{2\gamma}. \tag{3.2b}$$

Lemma 3.1. *Let $\theta \in \left[\frac{1}{2}, 1\right]$ and $c \geq 1$. Then, $\lambda_\pm$ are analytic for $s = \sigma + i\omega$ with $\sigma > 0$ and $|\omega| \in \left[\frac{\pi}{T}, \frac{\pi}{\Delta t}\right]$.*

Proof. The proof consists of two steps: $\hbar$ is analytic in the right half complex plane and $\Re(\hbar) \geq 0$. The second point ensures that the argument under the square root avoids the negative real axis.

Step 1: *$\hbar$ is analytic in the right half complex plane.* Since $z = e^{s\Delta t} = e^{\sigma\Delta t}(\cos(\omega\Delta t) + i\sin(\omega\Delta t))$ and $|\omega| \leq \frac{\pi}{\Delta t}$, we have $\Re(s) \in [-e^{\sigma\Delta t}, e^{\sigma\Delta t}]$. The argument $\hbar$ is a rational polynomial of z; hence, it suffices to prove that $\theta z + (1-\theta)$ does not have zeros for $\Re(s) \in [-e^{\sigma\Delta t}, e^{\sigma\Delta t}]$. This is true, because $\theta z + (1-\theta) = 0 \Leftrightarrow z = -\frac{1-\theta}{\theta} \in [-1, 0]$ for $\theta \in \left[\frac{1}{2}, 1\right]$, whereas $e^{\sigma\Delta t} > 1$ for $\sigma > 0$.

Step 2: $\Re(\hbar) \geq 0$. Let $h_0 = \theta e^{\sigma\Delta t}\cos(\omega\Delta t) + 1 - \theta$ and $h_1 = \theta e^{\sigma\Delta t}\sin(\omega\Delta t)$. Then, we have

$$\hbar = \frac{1}{\theta}\left[1 - \frac{h_0 - ih_1}{h_0^2 + h_1^2}\right], \tag{3.3}$$

and this implies $\Re(\hbar) \geq 0$. □

Let $s = \tilde{r}e^{i\tilde{\theta}}$ with $\tilde{\theta} \in (-\frac{\pi}{2}, \frac{\pi}{2})$. Then, from (3.3) we have $\lim\limits_{\tilde{r}\to+\infty}\hbar(\theta, z) = \frac{1}{\theta}$. Hence, for all $\tilde{\theta} \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$ it holds that $\lim_{\tilde{r}\to\infty}\lambda_\pm = \lambda^*_\pm$, where

$$\lambda^*_\pm := \frac{\frac{1}{\theta} + 2\gamma\zeta \pm \sqrt{\left(\frac{1}{\theta} + 2\gamma\zeta\right)^2 - 4\gamma^2}}{2\gamma}. \tag{3.4}$$

This, together with the maximum principle for complex analytic functions, gives

$$\max_{\Re(s)\geq 0}|\lambda_\pm| = \max\left\{\max_{\Re(s)=0}|\lambda_\pm|,\quad \lambda^*_\pm\right\}. \tag{3.5}$$

Let $c = \cos(\omega\Delta t)$. Then, for $s = i\omega$ it holds that

$$\hbar = \hbar_R(c) + i\hbar_I(c),\quad \lambda_+ = \Lambda_R(c) + i\Lambda_I(c), \tag{3.6}$$

where $\hbar_{R,I}$ and $\Lambda_{R,I}$ are defined by:

$$\begin{aligned}
&\hbar_R(c) = \frac{1}{\theta}\left[1 - \frac{\theta c + 1 - \theta}{\theta^2 + (1-\theta)^2 + 2\theta(1-\theta)c}\right],\ \hbar_I(c) = \frac{\sqrt{1-c^2}}{\theta^2 + (1-\theta)^2 + 2\theta(1-\theta)c},\\
&H_R(c) = [\hbar_R(c) + 2\gamma\zeta]^2 - \hbar_I^2(c) - 4\gamma^2,\ H_I(c) = 2\hbar_I(c)\,[\hbar_R(c) + 2\gamma\zeta],\\
&S_R(c) = \frac{\sqrt{H_R^2(c) + H_I^2(c)} + H_R(c)}{2},\ S_I(c) = \frac{\sqrt{H_R^2(c) + H_I^2(c)} - H_R(c)}{2},\\
&\Lambda_R(c) = \frac{\hbar_R(c) + 2\gamma\zeta + \sqrt{S_R(c)}}{2\gamma},\quad \Lambda_I(c) = \frac{\hbar_I(c) + \sqrt{S_I(c)}}{2\gamma}.
\end{aligned} \tag{3.7}$$

From (3.6), we have

$$\min_{\Re(s)\geq 0}|\lambda_+| = \frac{1}{\max\limits_{\Re(s)\geq 0}|\lambda_-|} = \min\left\{\sqrt{\Lambda_R^2(c) + \Lambda_I^2(c)}, \lambda_+^*\right\}, \tag{3.8}$$

where we have used $\lambda_-\lambda_+ \equiv 1$. For $\theta \in [\frac{1}{2}, 1]$, it is easy to verify $\hbar_R(c) \geq 0$ for $c \in [-1, 1]$. Hence, $\Lambda_R(c) \geq 1$ for $c \in [-1, 1]$ and this implies $|\lambda_+| \geq 1$ and $|\lambda_-| = |1/\lambda_+| \leq 1$.

Now, by using the boundedness of the error functions $\mathbf{e}_j^k$ at infinity, the constants A_j^k and B_j^k in (2.2a) can be determined as $B_1^k = 0$ for the first subsystem and $A_2^k = 0$ for the second one. Hence,

$$\hat{\mathbf{e}}_{1,m}^k(s) = A_1^k\lambda_+^m,\quad \hat{\mathbf{e}}_{2,m}^k(s) = B_2^k\lambda_-^m. \tag{3.9}$$

This, together with the boundary conditions (3.1b), gives the following recurrence relations:

$$A_1^k\,(\alpha\lambda_+ - 1) = B_2^{k-1}\,(\alpha\lambda_- - 1)\,,\ B_2^k\,(\alpha - \lambda_-) = A_1^{k-1}\,(\alpha - \lambda_+)\,. \tag{3.10}$$

Define

$$\hat{\rho}_{\text{opt}}^d = \frac{(\alpha - \lambda_+)^2}{(\alpha\lambda_+ - 1)^2},\ \rho_{\text{opt}}^d(\theta, \alpha) = \max_{\Re(s)\geq 0, |\Im(s)|\in[\pi/T, \pi/\Delta t]}\left|\hat{\rho}_{\text{opt}}^d\right| \tag{3.11}$$

(the subscript 'd' denotes '*discrete*'). We call $\hat{\rho}_{\text{opt}}^d$ and ρ_{opt}^d the convergence factors of the OWR methods in the frequency and time domain, respectively. Then, we have $A_1^k = \hat{\rho}_{\text{opt}}^d A_1^{k-2}$ and $B_2^k = \hat{\rho}_{\text{opt}}^d B_2^{k-2}$. Mathematically, we want $\rho_{\text{opt}}^d(\theta, \alpha) \ll 1$, which leads to the following min-max problem

$$\min_{\alpha\in\mathbb{R}}\max_{\Re(s)\geq 0, \frac{\pi}{T}\leq|\Im(s)|\leq\frac{\pi}{\Delta t}}\left|\frac{\alpha - \lambda_+}{\alpha\lambda_+ - 1}\right|^2. \tag{3.12}$$

We have proved $|\lambda_+| \geq 1$ for all $\theta \in [\frac{1}{2}, 1]$. Then, if $|\alpha| \leq 1$, we have $|\alpha - \lambda_+|^2 - |\alpha\lambda_+ - 1|^2 = \left(|\lambda_+|^2 - 1\right)(1 - \alpha^2) \geq 0$, which implies $\rho_{\text{opt}}^d(\theta, \alpha) > 1$. Hence, we only need to consider $|\alpha| > 1$ in (3.12). Moreover, if $\alpha < -1$,

$$\left|\frac{-\alpha - \lambda_+}{-\alpha\lambda_+ - 1}\right|^2 - \left|\frac{\alpha - \lambda_+}{\alpha\lambda_+ - 1}\right|^2 = \frac{4\alpha\Re(\lambda_+)(1 - \alpha^2)(1 - |\lambda_+|^2)}{|\alpha^2\lambda_+^2 - 1|^2} \leq 0,$$

which implies $\rho_{\text{opt}}^d(\theta, -\alpha) > \rho_{\text{opt}}^d(\theta, \alpha)$ for $\alpha > 1$. In summary, we can assume $\alpha > 1$ in (3.12).

Lemma 3.2. *Assume $\alpha > 1$ and $\theta \in \left[\frac{1}{2}, 1\right]$. Then,*

$$\rho_{\text{opt}}^d(\theta, \alpha) = \max\left\{\max_{\Re(s)=0}\left|\hat{\rho}_{\text{opt}}^d\right|, \left(\frac{\alpha - \lambda_+^*}{\alpha\lambda_+^* - 1}\right)^2\right\},$$

where λ_+^* is given by (3.4).

Proof. We have already proved that the argument λ_+ is analytic and satisfies $|\lambda_+| \geq 1$ in the right half complex plane. Then, we know that $\frac{\alpha-\lambda_+}{\alpha\lambda_+-1}$ is also analytic in the right half complex plane, since $\alpha > 1$ and $|\lambda_+| \geq 1$ guarantees that the denominator $\alpha\lambda_+ - 1$ will never be zero. Hence, the claim follows by applying the maximum principle for complex analytic functions. □

To analyze the min-max problem (3.12), we define the following notations:

$$N(\alpha, c) = (\alpha - \Lambda_R(c))^2 + \Lambda_I^2(c),\ D(\alpha, c) = [\alpha\Lambda_R(c) - 1]^2 + \alpha^2\Lambda_I^2(c), \tag{3.13}$$

where $c = \cos(\omega\Delta t)$ and $\Lambda_{R,I}(c)$ are defined by (3.7). For $\Re(s) = 0$, by using Lemma 3.2 we have

$$\left|\hat{\rho}_{\text{opt}}^d\right| = \mathcal{T}_\theta(\alpha, c) := \frac{N(\alpha, c)}{D(\alpha, c)}. \tag{3.14}$$

The aforementioned analysis implies that the min-max problem (3.12) is equivalent to

$$\min_{\alpha>1} \max \left\{ \max_{c\in[-1,c^*]} \mathcal{T}_\theta(\alpha, c),\ \left(\frac{\alpha - \lambda_+^*}{\alpha\lambda_+^* - 1}\right)^2 \right\}, \tag{3.15}$$

where $c^* = \cos\left(\frac{\Delta t}{T}\pi\right)$. In what follows, we focus on the analysis of finding the solution of (3.15).

3.1. The Trapezoidal rule: $\theta = \frac{1}{2}$

For $\theta = \frac{1}{2}$, it holds that $\hbar_R \equiv 0$ and $\hbar_I(c) = 2\sqrt{\frac{1-c}{1+c}}$. Define $y = \sqrt{\frac{1-c}{1+c}}$ and $R(y) = \gamma^2(\zeta^2 - 1) - y^2$. Then, we have $\Lambda_R = \frac{2\gamma\zeta + \sqrt{R_+(y)}}{2\gamma}$, $\Lambda_I = \frac{2y + \sqrt{R_-(y)}}{2\gamma}$ and

$$S_R = R_+(y) := 2\left(\sqrt{R^2(y) + 4\gamma^2\zeta^2 y^2} + R(y)\right),\ S_I = R_-(y) := 2\left(\sqrt{R^2(y) + 4\gamma^2\zeta^2 y^2} - R(y)\right).$$

Define

$$P(y) = \frac{2\gamma\zeta + \sqrt{R_+(y)}}{2\gamma},\ Q(y) = \frac{2y + \sqrt{R_-(y)}}{2\gamma},$$
$$\mathcal{P}(\alpha, y) = (\alpha - P)^2 + Q^2,\ \mathcal{Q}(\alpha, y) = (\alpha P - 1)^2 + \alpha^2 Q^2,\ \mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y) = \frac{\mathcal{P}(\alpha, y)}{\mathcal{Q}(\alpha, y)}. \tag{3.16}$$

Then, the min-max problem (3.15) is equivalent to

$$\min_{\alpha>1} \max \left\{ \max_{y\geq y_{\min}} \mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y),\ \left(\frac{\alpha - \lambda_+^*}{\alpha\lambda_+^* - 1}\right)^2 \right\}, \tag{3.17}$$

where $y_{\min} = \sqrt{\frac{1-c^*}{1+c^*}}$ and $c^* = \cos\left(\frac{\Delta t}{T}\pi\right)$.

Lemma 3.3. *Assume $\alpha > 1$. Then we have*

$$\max_{y\geq y_{\min}} \mathcal{T}_{\theta=\frac{1}{2}} = \max\left\{\mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y_{\min}), \alpha^{-2}\right\}.$$

Proof. To prove this claim, we need the following calculations:

$$\mathcal{P}_y := \frac{\partial\mathcal{P}}{\partial y} = 2[P - \alpha]P' + 2QQ',\ \mathcal{Q}\mathcal{P}_y = 2\left[(\alpha P - 1)^2 + \alpha^2 Q^2\right][(P - \alpha)P' + QQ'],$$
$$\mathcal{Q}_y := \frac{\partial\mathcal{Q}}{\partial y} = 2\alpha(\alpha P - 1)P' + 2\alpha^2 QQ',\ \mathcal{P}\mathcal{Q}_y = 2\alpha\left[(P - \alpha)^2 + Q^2\right][(\alpha P - 1)P' + \alpha QQ'],$$

where $P' = P'(y)$ and $Q' = Q'(y)$ are given by:

$$P' = \frac{y}{\gamma\sqrt{R_+(y)}}\left(\frac{(\zeta^2+1)\gamma^2+y^2}{\sqrt{R^2(y)+4\gamma^2\zeta^2y^2}}-1\right),\ Q' = \frac{1}{\gamma}+\frac{y}{\gamma\sqrt{R_-(y)}}\left(\frac{(\zeta^2+1)\gamma^2+y^2}{\sqrt{R^2(y)+4\gamma^2\zeta^2y^2}}+1\right).$$

Clearly, $Q' > 0$ for $y > 0$. Moreover, $[(\zeta^2+1)\gamma^2+y^2]^2 - R^2(y) - 4\gamma^2\zeta^2y^2 = 4\gamma^4\zeta^2 > 0$ and this gives $P' > 0$ for $y > 0$. Since $\alpha > 1$ and $P' > 0$, we have the following deduction:

$$-\alpha^2 < -1 \Leftrightarrow \alpha(P-\alpha) < \alpha P - 1 \Leftrightarrow \frac{(P-\alpha)P'+QQ'}{(\alpha P-1)P'+\alpha QQ'} < \frac{1}{\alpha} \Leftrightarrow \frac{(P-\alpha)P'+QQ'}{\alpha(\alpha P-1)P'+\alpha^2QQ'} < \frac{1}{\alpha^2}. \tag{3.18}$$

We now suppose that $\mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y)$ has a local extremum located at $y = y^* > 0$. Then,

$$\frac{\partial \mathcal{T}_{\theta=\frac{1}{2}}(\alpha,y)}{\partial y}\Big|_{y=y^*} = \frac{\mathcal{Q}\mathcal{P}_y(\alpha,y^*) - \mathcal{P}\mathcal{Q}_y(\alpha,y^*)}{\mathcal{Q}^2(\alpha,y^*)} = 0.$$

Since $\alpha > 1$, $y^* > 0$ and $P(y) \geq 1$, it holds that $\mathcal{Q}(\alpha, y^*), \mathcal{Q}_y(\alpha, y^*) \neq 0$. Therefore, we have

$$\mathcal{T}_{\theta=\frac{1}{2}}(\alpha,y^*) = \frac{\mathcal{P}(\alpha,y^*)}{\mathcal{Q}(\alpha,y^*)} = \frac{\mathcal{P}_y(\alpha,y^*)}{\mathcal{Q}_y(\alpha,y^*)},$$

i.e., at $y = y^*$ we have $\mathcal{T}_{\theta=\frac{1}{2}}(\alpha,y^*) = \frac{[P(y^*)-\alpha]P'(y^*)+Q(y^*)Q'(y^*)}{\alpha[\alpha P(y^*)-1]P'(y^*)+\alpha^2Q(y^*)Q'(y^*)}$. Using (3.18), we have $\mathcal{T}_{\theta=\frac{1}{2}}(\alpha,y^*) < \frac{1}{\alpha^2}$. On the other hand, for any $\alpha > 1$ it holds that $\lim\limits_{y\to+\infty}\frac{P(y)}{y} = 0$ and $\lim\limits_{y\to+\infty}\frac{Q(y)}{y} = \frac{2}{\gamma}$. Therefore,

$$\lim_{y\to+\infty}\mathcal{T}_{\theta=\frac{1}{2}}(\alpha,y) = \lim_{y\to+\infty}\frac{[\alpha-P(y)]^2+Q^2(y)}{[\alpha P(y)-1]^2+\alpha^2Q^2(y)} = \lim_{y\to+\infty}\frac{\left(\frac{\alpha-P(y)}{y}\right)^2+\left(\frac{Q(y)}{y}\right)^2}{\left(\frac{\alpha P(y)-1}{y}\right)^2+\alpha^2\left(\frac{Q(y)}{y}\right)^2} = \frac{1}{\alpha^2}.$$

In summary: (1) any local maximum of $\mathcal{T}_{\theta=\frac{1}{2}}$ can not exceed $\frac{1}{\alpha^2}$; (2) $\frac{1}{\alpha^2}$ can be reached at $y = +\infty$. □

Define

$$\widetilde{\mathcal{T}}_{\theta=\frac{1}{2}}(\alpha) = \max\left\{\mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y_{\min}), \left(\frac{\alpha-\lambda_+^*}{\alpha\lambda_+^*-1}\right)^2\right\}. \tag{3.19a}$$

Then, using Lemma 3.3 we know that the min-max problem (3.17) is equivalent to

$$\min_{\alpha>1}\max\left\{\widetilde{\mathcal{T}}_{\theta=\frac{1}{2}}(\alpha),\ \alpha^{-2}\right\}. \tag{3.19b}$$

Theorem 3.4 ($\theta = \frac{1}{2}$)**.** *Let* $y_{\min} = \sqrt{\frac{1-\cos(\Delta t\pi/T)}{1+\cos(\Delta t\pi/T)}}$. *Then, the best performance of the discrete OWR method* (2.3) *with* $\theta = \frac{1}{2}$ (i.e., *the Trapezoidal rule*) *is obtained for* $\alpha = \alpha_{\rm opt}^d$, *where* $\alpha_{\rm opt}^d$, *the solution of* (3.19b), *is given by:*

$$\alpha_{\rm opt}^d = \begin{cases}\alpha^*, & \text{if } \widetilde{\mathcal{T}}_{\theta=\frac{1}{2}}(\alpha^*) \geq \frac{1}{\alpha_*^2},\\ \alpha_0^*, & \text{otherwise,}\end{cases} \tag{3.20}$$

where α_0^* *is the unique root of* $\widetilde{\mathcal{T}}_{\theta=\frac{1}{2}}(\alpha) = \frac{1}{\alpha^2}$ *and is given by*

$$\alpha^* = \begin{cases}\alpha_1, & \text{if } \mathcal{T}_{\theta=\frac{1}{2}}(\alpha_1, y_{\min}) \geq \left(\frac{\alpha_1-\lambda_+^*}{\alpha_1\lambda_+^*-1}\right)^2,\\ \alpha_0, & \text{otherwise.}\end{cases}$$

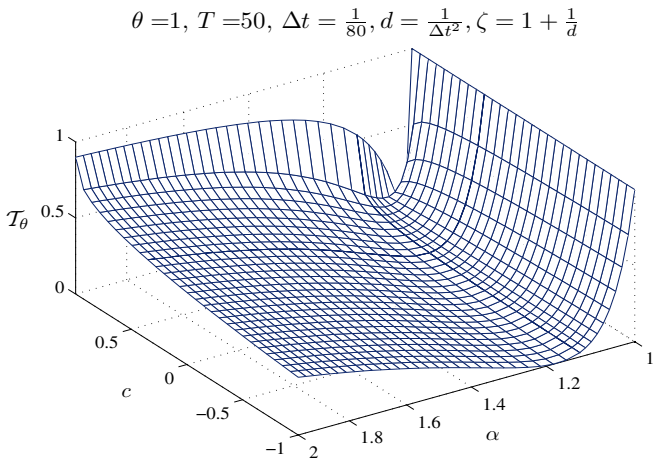

FIGURE 2. The profile of $\mathcal{T}_\theta$ as a function of α and c.

The arguments α_0 and α_1 are defined by

$$\mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y_{\min}) = \left(\frac{\alpha-\lambda_+^*}{\alpha\lambda_+^*-1}\right)^2 (\Rightarrow \text{root is } \alpha_0),\ \ \alpha_1 = \frac{\lambda_{\min}^2+1}{2P_{\min}} + \sqrt{\left(\frac{\lambda_{\min}^2+1}{2P_{\min}}\right)^2 - 1}, \tag{3.21}$$

where $\lambda_{\min} = \sqrt{P^2(y_{\min}) + Q^2(y_{\min})}$ and $P_{\min} = P(y_{\min})$. With the optimized parameter α_{opt}^d, the convergence factor of the discrete OWR method (2.3), $\rho_{\text{opt}}^d(\theta,\alpha)$ defined by (3.11), satisfies

$$\rho_{\text{opt}}^d = \widetilde{\mathcal{T}}_{\theta=\frac{1}{2}}\left(\alpha_{\text{opt}}^d\right). \tag{3.22}$$

Proof. Routine calculation yields

$$\text{sign}\left(\partial_\alpha \mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y_{\min})\right) = \text{sign}\left(\alpha^2 - \alpha(\lambda_{\min}^2+1)/P_{\min} + 1\right). \tag{3.23}$$

Then, for $\alpha \in [1,\infty)$, we know that α_1 is the unique local minimizer of $\mathcal{T}_{\theta=\frac{1}{2}}(\alpha, y_{\min})$. It is clear that λ_+^* is the unique local minimizer of $\left(\frac{\alpha-\lambda_+^*}{\alpha\lambda_+^*-1}\right)^2$ and that the minimum is zero. Hence, by using Lemma 2.1, we know that α^* is the unique local minimizer of the function $\widetilde{\mathcal{T}}_{\theta=\frac{1}{2}}(\alpha)$. Since α^{-2} is decreasing for $\alpha \in [1,\infty)$, a simple logical deduction shows that the quantity α_{opt}^d defined by (3.20) is the solution of (3.19b). □

3.2. The case $\theta \in \left(\frac{1}{2}, 1\right]$

We now analyze the min-max problem (3.15) for $\theta > \frac{1}{2}$. In Figure 2, we plotted the profile of the function $\mathcal{T}_\theta$ for $\theta = 1$, $\Delta t = 0.01$, $T = 20$, $d = 10^4$ and $\zeta = 1 + \frac{1}{d}$, where we see that for given $\alpha > 1$ the maximal value of $\mathcal{T}_\theta$ is obtained at either $c = -1$ or $c = c^* \left(= \cos\left(\frac{\Delta t\pi}{T}\right)\right)$, *i.e.*, $\max_{c\in[-1,c^*]} \mathcal{T}_\theta(\alpha, c) = \max\{\mathcal{T}_\theta(\alpha,-1), \mathcal{T}_\theta(\alpha,c^*)\}$. It is difficult to rigorously prove this result owing to the high complicity of $\mathcal{T}_\theta(\alpha,c)$ for $\theta \neq \frac{1}{2}$, but our numerical results for many other values of Δt, T, d and θ indicate that this result always holds. We use this numerical result as a hypothesis and all our analyses in the following for $\theta \in (\frac{1}{2}, 1]$ are based on this hypothesis.

Hypothesis 3.5. *For $\theta \in (\frac{1}{2}, 1]$, $d > 0$ and $\zeta = -a/(2d) \geq 1$, we assume that the function $\mathcal{T}_\theta(\alpha,c)$ defined by (3.14) with $N(\alpha,c)$ and $D(\alpha,c)$ defined by (3.13) satisfies*

$$\max_{c\in[-1,c^*]} \mathcal{T}_\theta(\alpha,c) = \max\{\mathcal{T}_\theta(\alpha,-1), \mathcal{T}_\theta(\alpha,c^*)\},$$

where $c^* = \cos\left(\frac{\Delta t\pi}{T}\right)$.

Under this hypothesis the min-max problem (3.15) can be rewritten as:

$$\min_{\alpha>1}\max\left\{\mathcal{T}_\theta(\alpha,-1),\mathcal{T}_\theta(\alpha,c^*),\ \left(\frac{\alpha-\lambda_+^*}{\alpha\lambda_+^*-1}\right)^2\right\}. \tag{3.24}$$

Similar to the proof of Theorem 3.4, it can be shown that for $\alpha\in[1,+\infty)$ the function $\mathcal{T}_\theta(\alpha,-1)$ (resp. $\mathcal{T}_\theta(\alpha,c^*)$) has a unique local minimizer $\tilde{\alpha}_{-1}$ (resp. $\tilde{\alpha}_1$), where $\tilde{\alpha}_{-1}$ and $\tilde{\alpha}_1$ are defined by:

$$\tilde{\alpha}_{-1}=\frac{\Lambda_R^2(-1)+1}{2\Lambda_R(-1)}+\sqrt{\left(\frac{\Lambda_R^2(-1)+1}{2\Lambda_R(-1)}\right)^2-1},\ \tilde{\alpha}_1=\frac{\lambda_{c^*}^2+1}{2\Lambda_R(c^*)}+\sqrt{\left(\frac{\lambda_{c^*}^2+1}{2\Lambda_R(c^*)}\right)^2-1}, \tag{3.25}$$

where $\lambda_{c^*}:=|\lambda_+|_{c=c^*}=\sqrt{\Lambda_R^2(c^*)+\Lambda_I^2(c^*)}$. Then, applying Lemma 2.1, we get the following result.

Proposition 3.6. *Let $g_1(\alpha)=\mathcal{T}_\theta(\alpha,-1)$, $g_2(\alpha)=\mathcal{T}_\theta(\alpha,c^*)$ and $g_3(\alpha)=\left(\frac{\alpha-\lambda_+^*}{\alpha\lambda_+^*-1}\right)^2$. Let $x_1^*=\tilde{\alpha}_{-1}$, $x_2^*=\tilde{\alpha}_{-1}$ and $x_3^*=\lambda_+^*$, and X_3^* be the quantity determined by the minimizing-procedure given in Lemma 2.1. Then, a reliable parameter α used in the discrete OWR method (2.3) can be chosen as $\alpha_{\rm opt}^d=X_3^*$. With the choice $\alpha=\alpha_{\rm opt}^d$, the convergence factor of the discrete OWR satisfies*

$$\rho_{\rm opt}^d=\max_{j=1,2,3}g_j(\alpha_{\rm opt}^d). \tag{3.26}$$

4. Asymptotic analysis

In this section, we analyze the asymptotic dependence of the convergence factor $\rho_{\rm opt}^d$ on T (the length of time interval) and Δt (the mesh size). For $\theta\in(\frac{1}{2},1]$, we assume that the function $\mathcal{T}_\theta(\alpha,c)$ defined by (3.14) satisfies Hypothesis 3.5.

4.1. Asymptotic results with respect to Δt

For Δt small, we have $c^*=\cos\left(\frac{\Delta t\pi}{T}\right)=1-\frac{\pi^2}{2T^2}\Delta t^2+O\left(\Delta t^4\right)$. To analyze the asymptotic dependence of $\rho_{\rm opt}^d$ on Δt, we make an ansatz $\alpha_{\rm opt}^d=1+C\Delta t^{-\beta}$ with $\beta>0$. Then, for all $\theta\in\left[\frac{1}{2},1\right]$ a tedious but routine calculation yields the following results

$$\begin{aligned}
&\left(\frac{\alpha_{\rm opt}^d-\lambda_+^*}{\alpha_{\rm opt}^d\lambda_+^*-1}\right)^2=\left(\frac{\frac{1}{d\theta}-C\Delta t^{1-\beta}+O\left(\Delta t\right)}{\frac{C}{d\theta}+\frac{\Delta t^\beta}{d\theta}+O\left(\Delta t\right)}\right)^2\Delta t^{2\beta},\\
&\mathcal{T}_\theta\left(\alpha_{\rm opt}^d,-1\right)=\left(\frac{2-dC(2\theta-1)\Delta t^{1-\beta}+O\left(\Delta t\right)}{2C+2\Delta t^\beta+O\left(\Delta t\right)}\right)^2\Delta t^{2\beta},\\
&\mathcal{T}_\theta\left(\alpha_{\rm opt}^d,c^*\right)=\frac{C^2+2C(1-m_0)\Delta t^\beta+\left[(m_0-1)^2+n_0^2\right]\Delta t^{2\beta}+O\left(\Delta t^{1+\beta}\right)}{C^2(m_0^2+n_0^2)+2C[m_0(m_0-1)+n_0^2]\Delta t^\beta+[(m_0-1)^2+n_0^2]\Delta t^{2\beta}+O\left(\Delta t\right)},\\
&\Lambda_R(c^*)=m_0+O\left(\Delta t\right),\ \Lambda_I(c^*)=n_0+O\left(\Delta t\right),
\end{aligned} \tag{4.1a}$$

where m_0 and n_0 are given by:

$$\begin{aligned}
m_0&=\zeta+\sqrt{\frac{1}{8d^2}\left[\sqrt{\left(4d^2(\zeta^2-1)-\frac{\pi^2}{T^2}\right)^2+16d^2\zeta^2\frac{\pi^2}{T^2}}+4d^2(\zeta^2-1)-\frac{\pi^2}{T^2}\right]},\\
n_0&=\frac{\pi}{2{\rm d}T}+\sqrt{\frac{1}{8d^2}\left[\sqrt{\left(4d^2(\zeta^2-1)-\frac{\pi^2}{T^2}\right)^2+16d^2\zeta^2\frac{\pi^2}{T^2}}-4d^2(\zeta^2-1)+\frac{\pi^2}{T^2}\right]}.
\end{aligned} \tag{4.1b}$$

From the analysis in Section 3, we have

$$\rho^d_{\text{opt}}(\theta,\alpha)=\max\left\{\mathcal{T}_\theta(\alpha,-1),\mathcal{T}_\theta(\alpha,c^*),\left(\frac{\alpha-\lambda^*_+}{\alpha\lambda^*_+-1}\right)^2\right\}. \tag{4.2}$$

Hence, if $\beta>0$ from (4.1a) we have $\rho^d_{\text{opt}}\left(\theta,\alpha^d_{\text{opt}}\right)=\mathcal{T}_\theta(\alpha^d_{\text{opt}},c^*)$ for Δt small. Since the function $\mathcal{T}_\theta(\alpha,c^*)$ attains its global minimum at $\alpha=\frac{\Lambda^2_R(c^*)+\Lambda^2_I(c^*)+1}{2\Lambda_R(c^*)}+\sqrt{\left(\frac{\Lambda^2_R(c^*)+\Lambda^2_I(c^*)+1}{2\Lambda_R(c^*)}\right)^2-1}$, using (4.1a) again we have for Δt small that $\alpha^d_{\text{opt}}=\frac{m_0^2+n_0^2+1}{2m_0}+\sqrt{\left(\frac{m_0^2+n_0^2+1}{2m_0}\right)^2-1}$. Clearly, this contradicts with the assumption $\alpha^d_{\text{opt}}=1+C\Delta t^{-\beta}$ with $\beta>0$. By a similar analysis, we can also exclude the case $\alpha^d_{\text{opt}}=1+C\Delta t^{\beta}$ with $\beta>0$. Therefore, there is only one possibility left, $\alpha^d_{\text{opt}}=1+C$ with $C>0$. In this case, using (4.1a) again, for Δt small we have

$$\left(\frac{\alpha^d_{\text{opt}}-\lambda^*_+}{\alpha^d_{\text{opt}}\lambda^*_+-1}\right)^2\approx\frac{1}{(1+C)^2},\ \mathcal{T}_\theta\left(\alpha^d_{\text{opt}},-1\right)\approx\frac{1}{(1+C)^2},\ \mathcal{T}_\theta\left(\alpha^d_{\text{opt}},c^*\right)\approx\frac{(C+1-m_0)^2+n_0^2}{\left[(C+1)m_0-1\right]^2+(C+1)^2n_0^2}.$$

Let $\tilde{c}=C+1$. Then, for Δt small, it holds that $\rho^d_{\text{opt}}(\theta,\alpha^d_{\text{opt}})=\max\left\{\frac{1}{\tilde{c}^2},\frac{(\tilde{c}-m_0)^2+n_0^2}{[\tilde{c}m_0-1]^2+\tilde{c}^2n_0^2}\right\}$ and that the best choice of $\tilde{c}$ is $\tilde{c}=m_0+\sqrt{m_0^2-1}$, which gives $\alpha^d_{\text{opt}}=1+C=m_0+\sqrt{m_0^2-1}$.

Proposition 4.1. *Let $T>0$, $\zeta\geq1$, $\theta\in[\frac{1}{2},1]$ and m_0 be the quantity defined by (4.1b). Then, for Δt small, we have the following asymptotic results*

$$\alpha^d_{opt}\approx m_0+\sqrt{m_0^2-1},\ \rho^d_{\text{opt}}\approx\left(m_0+\sqrt{m_0^2-1}\right)^{-2}. \tag{4.3}$$

4.2. Asymptotic results with respect to T

For fixed Δt and T large, we have $c^*=\cos\left(\frac{\pi\Delta t}{T}\right)=1-\frac{(\pi\Delta t)^2}{2}T^{-2}+O\left(T^{-4}\right)$. Let

$$\lambda^*_{-1}=\frac{\frac{2}{2\theta-1}+2\gamma\zeta+\sqrt{\left(\frac{2}{2\theta-1}+2\gamma\zeta\right)^2-4\gamma^2}}{2\gamma}. \tag{4.4}$$

Then, it holds that

$$\mathcal{T}_\theta(\alpha,-1)=\left(\frac{\alpha-\lambda^*_{-1}}{\alpha\lambda^*_{-1}-1}\right)^2. \tag{4.5}$$

Note that, for $\theta=\frac{1}{2}$ we have $\lambda^*_{-1}=+\infty$, which gives $\mathcal{T}_\theta(\alpha,-1)=\frac{1}{\alpha^2}$ and therefore the min-max problem (4.2) is reduced to (3.19b). Our analysis in what follows is divided into two cases, $\zeta=1$ and $\zeta>1$.

For $\zeta=1$, by assuming

$$\alpha^d_{\text{opt}}=1+C_\dagger T^{-\beta}\text{ with }\beta\in\left(0,\frac{1}{2}\right), \tag{4.6}$$

we have

$$\mathcal{T}_\theta\left(\alpha^d_{\text{opt}},-1\right)\approx1-2C_\dagger\left(\frac{\lambda^*_{-1}+1}{\lambda^*_{-1}-1}\right)T^{-\beta},\left(\frac{\alpha^d_{\text{opt}}-\lambda^*_+}{\alpha^d_{\text{opt}}\lambda^*_+-1}\right)^2\approx1-2C_\dagger\left(\frac{\lambda^*_++1}{\lambda^*_+-1}\right)T^{-\beta},$$

$$\mathcal{T}_\theta\left(\alpha^d_{\text{opt}},c^*\right)\approx1-\frac{4\sqrt{\frac{\pi\Delta t}{2d}}}{C_\dagger}T^{-\left(\frac{1}{2}-\beta\right)}. \tag{4.7}$$

Define

$$\lambda^*_{\min} = \min\left\{\frac{\lambda^*_{-1}+1}{\lambda^*_{-1}-1}, \frac{\lambda^*_{+}+1}{\lambda^*_{+}-1}\right\}. \tag{4.8}$$

Then, it holds that

$$\max\left\{\mathcal{T}_\theta\left(\alpha^d_{\text{opt}}, -1\right), \left(\frac{\alpha^d_{\text{opt}}-\lambda^*_+}{\alpha^d_{\text{opt}}\lambda^*_+ - 1}\right)^2\right\} \approx 1 - 2C_\dagger\lambda^*_{\min}T^{-\beta}. \tag{4.9}$$

Balancing (4.9) and the third term in (4.7) gives

$$\beta = \frac{1}{2} - \beta \Rightarrow \beta = \frac{1}{4},\ 2C_\dagger\lambda^*_{\min} = \frac{4\sqrt{\frac{\pi\Delta t}{2d}}}{C_\dagger} \Rightarrow C_\dagger = \left(\frac{2\pi\Delta t}{d\lambda^2_{\min}}\right)^{\frac{1}{4}}. \tag{4.10}$$

Proposition 4.2. *Let $\zeta := \frac{-a}{2d} = 1$, Δt be a fixed number and $\theta \in [\frac{1}{2}, 1]$. Then, for T large we have*

$$\alpha^d_{\text{opt}} \approx 1 + \left(\frac{2\pi\Delta t}{d\lambda^2_{\min}}\right)^{\frac{1}{4}} T^{-\frac{1}{4}},\ \rho^d_{\text{opt}} \approx 1 - \left(\frac{32\lambda^2_{\min}\pi\Delta t}{d}\right)^{\frac{1}{4}} T^{-\frac{1}{4}}, \tag{4.11}$$

where $\lambda^*_{\min}$ is defined by (4.8).

For $\zeta > 1$, we can expect constant convergence factor when T is large. In this case, by noticing $c^* = \lim_{T\to+\infty}\cos\left(\frac{\pi\Delta t}{T}\right) = 1$, from equation (3.7) we have $\hbar_R(1) = \hbar_I(1) = 0$, which implies $\Lambda_R(1) = \tilde{\lambda}^* := \zeta + \sqrt{\zeta^2-1}$ and $\Lambda_I(1) = 0$. Hence, $\mathcal{T}_\theta(\alpha, c^*) \to \left(\frac{\alpha-\tilde{\lambda}^*}{\alpha\tilde{\lambda}^*-1}\right)^2$ as $T \to +\infty$. Let $g_1(\alpha) := \mathcal{T}_\theta(\alpha, -1) = \left(\frac{\alpha-\lambda^*_{-1}}{\alpha\lambda^*_{-1}-1}\right)^2$, $g_2(\alpha) := \mathcal{T}_\theta(\alpha, 1) = \left(\frac{\alpha-\tilde{\lambda}^*}{\alpha\tilde{\lambda}^*-1}\right)^2$ and $g_3(\alpha) := \left(\frac{\alpha-\lambda^*_+}{\alpha\lambda^*_+-1}\right)^2$, where λ^*_+ and λ^*_{-1} are defined by (3.4) and (4.4). Let $x^*_1 = \lambda^*_{-1}$, $x^*_2 = \tilde{\lambda}^*$, $x^*_3 = \lambda^*_+$ and X^*_3 be the quantity determined by the minimizing-procedure given in Lemma 2.1. Then, for $T \to +\infty$ from (4.2) and (4.5) we know that Proposition 3.6 also holds and $\rho^d_{\text{opt}} = \max_{j=1,2,3} g_j\left(\alpha^d_{\text{opt}}\right) < 1$.

Remark 4.3 (Results from continuous analysis)**.** For the continuous OWR method (2.3), the best parameter α^c_{opt} (the superscript 'c' denote 'continuous'), is determined by (see [1], Chap. 3 for details):

$$\mathcal{R}(\tau_{\min}, \alpha) = \mathcal{R}(\tau_{\max}, \alpha), \tag{4.12a}$$

where $\mathcal{R}(\tau, \alpha) = \frac{2\alpha^2\zeta - \alpha^2\tau - 4\alpha\tau\zeta + 2\alpha\tau^2 + \tau}{-4\alpha\tau\zeta + 2\alpha\tau^2 + 2\zeta - \tau + \alpha^2\tau}$, $\tau_{\min,\max} = \zeta + \frac{\sqrt{2}}{4}\sqrt{\sqrt{\xi_{\min,\max}} - \tilde{\omega}^2_{\min,\max} + 4(\zeta^2-1)}$ and

$$\xi_{\min,\max} = \tilde{\omega}^4_{\min,\max} + 8\zeta^2\tilde{\omega}^2_{\min,\max} + 8\tilde{\omega}^2_{\min,\max} + 16(\zeta^2-1)^2 \text{ with } \tilde{\omega}_{\min} = \frac{\pi}{\mathrm{d}T} \text{ and } \tilde{\omega}_{\max} = \frac{\pi}{\mathrm{d}\Delta t}. \tag{4.12b}$$

For a given temporal discretization, we will compare in the next section the convergence rates of the OWR methods using the two choices of the parameter α, *i.e.*, $\alpha = \alpha^c_{\text{opt}}$ and $\alpha = \alpha^d_{\text{opt}}$.

5. Numerical results

In this section, we compare the performance of the OWR method using the parameters from the continuous and discrete analyses. We consider a model RC circuit with 100 nodes with parameters $R = \frac{1}{3}$ Ohm and $C = \frac{3}{200}$ pF. The resistor value is chosen to be $\underline{R} = \frac{40}{3}$ Ohm. This setting gives state equation (2.1a) with $d = 200$ and $a = -(2d+5)$. The source term $\mathbf{f}(t)$ is determined as this: we use $I_s(t) = 1 + N$ for $t \in [N, N+1]$

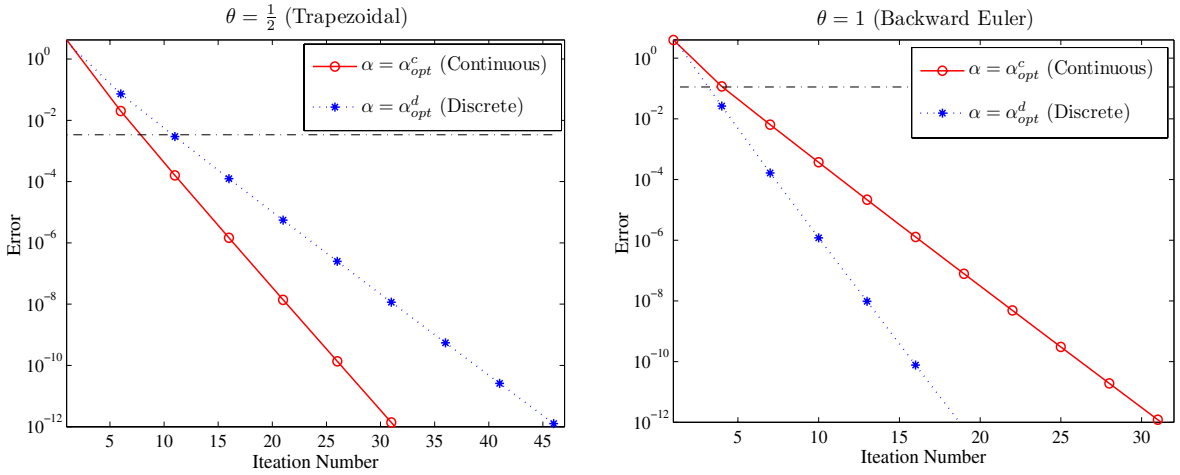

FIGURE 3. Comparisons of the convergence rates of the OWR methods for two time-integrators: $\theta = \frac{1}{2}$ (*left*) and $\theta = 1$ (*right*). The discretization/problem parameters are $\Delta t = 0.02$, $T = 50$, $a = -(2d+5)$ and $d = 200$. In each subfigure, the horizontal line indicates the truncation error of the time-integrator, which shows how many iterations one should really use in practice.

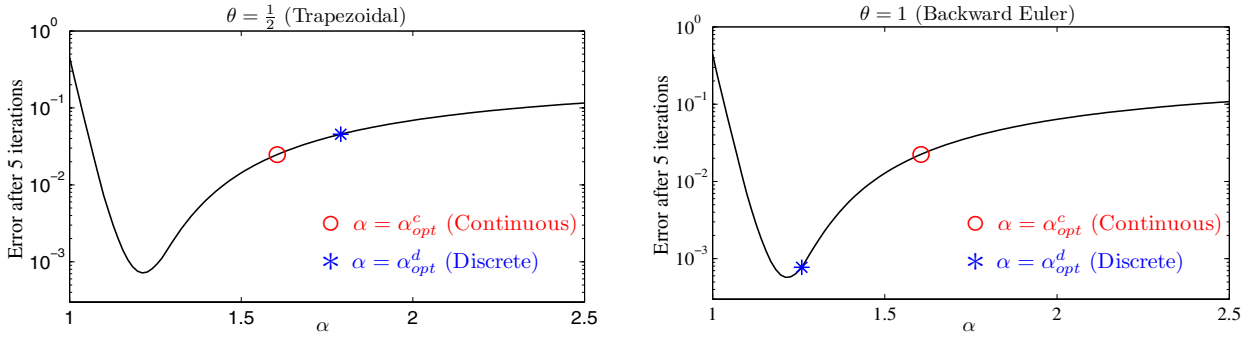

FIGURE 4. Measured error of the OWR method after 5 iterations for various values of the parameter α. The parameters $\alpha = \alpha^c_{\rm opt}$ from the continuous analysis and $\alpha = \alpha^d_{\rm opt}$ from the discrete analysis are denoted by '$\circ$' and '$*$'. *Left*: $\theta = \frac{1}{2}$; *Right*: $\theta = 1$.

as the input function with $N \geq 0$ being an integer, and then $\mathbf{f}(t) = (I_s(t)/C, 0, \ldots, 0)^\top \in \mathbb{R}^{200}$. The initial iterate for the OWR method is chosen randomly and the iteration stops when the global error satisfies

$$\max_n \left\| \tilde{\mathbf{x}}^k(n) - \tilde{\mathbf{x}}(n) \right\|_\infty \leq 10^{-12}, \tag{5.1}$$

where $\{\tilde{\mathbf{x}}(n)\}$ is the reference solution obtained by using the same time-integrator as that for $\{\tilde{\mathbf{x}}^k(n)\}$.

In Figure 3, we compare the convergence rates of the OWR methods using different parameters α. The left and right subfigures correspond to $\theta = \frac{1}{2}$ (*i.e.*, the Trapezoidal rule) and $\theta = 1$ (*i.e.*, the Backward–Euler method), respectively. We see clearly that, for $\theta = \frac{1}{2}$ it is better to use $\alpha = \alpha^c_{\rm opt}$ instead of $\alpha = \alpha^d_{\rm opt}$, while in the case of $\theta = 1$, $\alpha = \alpha^d_{\rm opt}$ is a better choice than $\alpha = \alpha^c_{\rm opt}$.

With the same discretization/problem parameters, this conclusion is further confirmed by the results shown in Figure 4, where we show the measured error of the OWR method after 5 iterations using various values for the parameter α; the choice $\alpha = \alpha^c_{\rm opt}$ is denoted by a circle and $\alpha = \alpha^d_{\rm opt}$ is denoted by a star. Besides the message

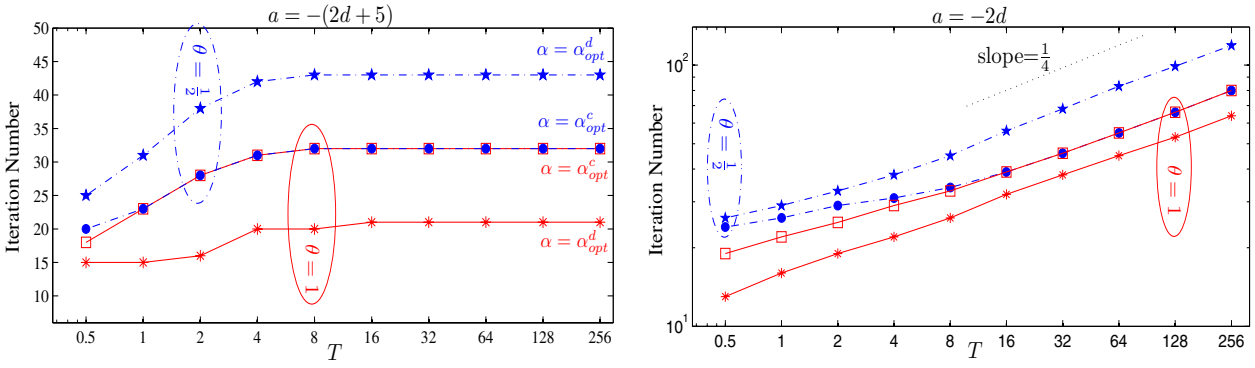

FIGURE 5. For $\Delta t = 0.02$, dependence of the iteration number of the OWR method on T. *Left*: $a = -(2d+5)$; *Right*: $a = -2d$.

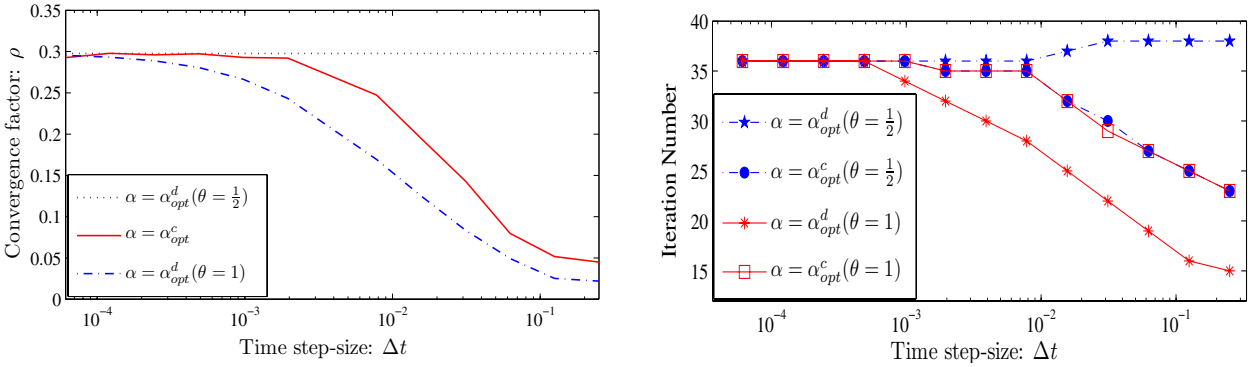

FIGURE 6. For $T = 4$, $a = -(2d+5)$ and $d = 200$, dependence of the convergence factor ρ (*left*) and the measured iteration number (*right*) of the OWR method on Δt.

similar to that implied by Figure 3, another message from Figure 4 is that for $\theta = 1$ the parameter $\alpha = \alpha^d_{\text{opt}}$ from the discrete analysis is very close to the best choice that we can get through numerical implementation, while for $\theta = \frac{1}{2}$ both $\alpha = \alpha^d_{\text{opt}}$ and $\alpha = \alpha^c_{\text{opt}}$ are far away from the best one. This implies that further work is needed to let the OWR method using the Trapezoidal rule as the temporal discretization converge rapidly.

We next verify the asymptotic dependence of the convergence rates of the OWR method on T, the length of time interval, and the mesh size Δt. To this end, in Figure 5 we show the measured iteration number needed to satisfy the stopping criterion (5.1) for several values of T. As we have analyzed in Section 4.2, the convergence factor of the OWR method presents different asymptotic behavior with respect to T, depending on $\zeta := \frac{-a}{2d} > 1$ or $\zeta = 1$, and therefore we consider two cases in Figure 5: in the left subfigure we consider the case $\zeta > 1$ and in the right subfigure we consider $\zeta = 1$. For the first case, we see that the OWR method under different time-integrators and different choices of α behaves robustly as T increases, while for the case $\zeta = 1$ the iteration number increases with a rate of order $O\left(T^{\frac{1}{4}}\right)$, just as Proposition 4.2 predicts.

Then, in Figure 6 we show the asymptotic dependence of the OWR method on Δt. For problem parameters $T = 4$, $a = -(2d+5)$ and $d = 200$, we first show the convergence factor as a function of Δt on the left subfigure. We see that both ρ^c and ρ^d approach to 0.3 as Δt goes to 0 and that $\rho^d_{\theta=1} < \rho^c < \rho^d_{\theta=\frac{1}{2}}$. The first conclusion confirms Proposition 4.1 very well, because from (4.1b) we see that the argument m_0 is independent of θ and therefore for all $\theta \in [\frac{1}{2}, 1]$ the convergence factor ρ^d approaches to the same quantity. These theoretical

predictions are further confirmed by the results shown in the right subfigure, where for different Δt we show the measured number of iterations required to satisfy the stopping criterion (5.1).

6. Conclusions

The efficiency of the classical WR methods for circuit simulations depends very much on finding a good partitioning: the engineer needs to find subcircuits such that the coupling between them is weak; then with the corresponding partitioning the method should converge rapidly. However, finding such a partitioning is not always easy: "*In practice one is interested in knowing what subdivisions yield fast convergence for the iterations... The splitting into subsystems is assumed to be given. How to split in such a way that the coupling remain "weak" is an important question*" [16]. The optimized WR approach is a mathematical technique to reach the two goals concurrently, *i.e.*, maintaining the partitioning procedure as simple as possible and maintaining the convergence as fast as possible. The second goal is realized by optimizing the parameter, namely α, involved in the transmission conditions, which exchange a combination of voltages and currents rather than just voltages or just currents from one subcircuit to its neighboring subcircuits.

In the last decade, optimizing the parameter α is studied by many authors for several different circuits. The optimization is done at the continuous level and the obtained parameter, namely α^c_{opt}, is used for practical circuit simulations. Then, it is natural to ask: *can the convergence rate of the OWR method be further improved by directly optimizing α at the discrete level?* By using the diffusive circuit as the model, this paper provides a positive answer for this question, if we use the Backward–Euler method as the numerical method, which, as we found in the literature, is the most frequently used time-integrator in this field. However, this conclusion is not applicable to the Trapezoidal rule, the simplest 2nd-order, A-stable and one-step numerical method. For this method, it is better to use α^c_{opt} from the continuous analysis instead of α^d_{opt} from the discrete analysis.

Our ongoing study is devoted to optimizing the transmission condition parameter α for finite-size RC circuits. For the Trapezoidal rule, from Figure 4 on the left we see that the parameter α^d_{opt} is far from *optimal* in practical computation and one possible reason is that this parameter is analyzed for circuits of infinite-size, while the numerical experiments are carried out for finite-size circuits. From the most recent work by Al-Khaleel *et al.* [3,4], where the authors analyzed the optimized WR method for finite-size RC circuits at the continuous level, we believe that the discrete analysis for finite-size RC circuits shall result in better parameter, since, obviously, in practice the size of a circuit is always finite in a concrete circuit simulation.

Acknowledgements. The authors are very grateful to the anonymous referees for the careful reading of a preliminary version of the manuscript and their valuable suggestions and comments, which greatly improved the quality of this paper. The authors are also very grateful to Professor Annalisa Buffa as the editor of this paper for handling the submission and review of this paper. This work is supported by the NSF of China (11301362, 11371157, 11671074, 61573010), the NSF of Technology & Education of Sichuan Province (2014JQ0035, 15ZA0220), the China Postdoctoral Foundation (2015M580777, 2016T90841) and the project of Sichuan University of Science and Engineering (2015LX01).

References

[1] M. Al-Khaleel, *Optimized waveform relaxation methods for circuit simulations*. Ph.D. dissertation, McGill University (2007).

[2] M.D. Al-Khaleel, M.J. Gander and A.E. Ruehli, Optimized waveform relaxation solution of RLCG transmission line type circuits, In *IEEE 9th International Conference on Innovations in Information Technology (IIT)*. IEEE Publisher (2013) 136–140.

[3] M. Al-Khaleel, M.J. Gander and A.E. Ruehli, A mathematical analysis of optimized waveform relaxation for a small RC circuit. *Appl. Numer. Math.* **75** (2014) 61–76.

[4] M. Al-Khaleel, M.J. Gander and A. Ruehli, Optimization of transmission conditions in waveform relaxation techniques for RC circuits. *SIAM J. Numer. Anal.* **52** (2014) 1076–1101.

[5] M.J. Gander and A. Ruehli, Optimized waveform relaxation methods for RC type circuits. *IEEE Trans. Circuits Syst. I, Reg. Papers* **51** (2004) 755–768.

[6] M.J. Gander and A.M. Stuart, Space-Time continuous analysis of waveform relaxation for the heat equation. *SIAM J. Sci. Comput.* **19** (1998) 2014–2031.

[7] M.J. Gander and L. Halpern, Méthodes de relaxation d'ondes pour l'équation de la chaleur en dimension 1. *C.R. Acad. Sci. Paris, Série I* **336** (2003) 519–524.

[8] M.J. Gander and L. Halpern, Optimized Schwarz waveform relaxation for advection reaction diffusion problems. *SIAM J. Numer. Anal.* **45** (2007) 666–697.

[9] M.J. Gander, M. Al-Khaleel and A.E. Ruehli, Waveform relaxation technique for longitudinal partitioning of transmission lines, in *Digest of Electr. Perf. Electronic Packaging* (2006), pp. 207–210.

[10] M.J. Gander, L. Halpern and F. Nataf, Optimal Schwarz waveform relaxation for the one dimensional wave equation. *SIAM J. Numer. Anal.* **41** (2003) 1643–1681.

[11] M.J. Gander and A.E. Ruehli, Optimized waveform relaxation solution of electromagnetic and circuit problems. *IEEE 19th Conference on Electrical Performance of Electronic Packaging and Systems (EPEPS)* (2010) 65–68.

[12] M.J. Gander, M. Al-Khaleel and A.E. Ruehli, Optimized waveform relaxation methods for longitudinal partitioning of transmission lines. *IEEE Trans. Circuits Syst. I Reg. Papers* **56** (2009) 1732–1743.

[13] E. Giladi and H.B. Keller, Space-time domain decomposition for parabolic problems. *Numer. Math.* **93** (2002) 279–313.

[14] Y.L. Jiang, On time-domain simulation of lossless transmission lines with nonlinear terminations. *SIAM J. Numer. Anal.* **42** (2004) 1018–1031.

[15] E. Lelarasmee, A.E. Ruehli and A.L. Sangiovanni-Vincentelli, The waveform relaxation methods for time-domain analysis of large scale integrated circuits. *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.* **CAD-1** (1982) 131–145.

[16] O. Nevanlinna, Remarks on Picard-Lindelöf iterations, Part I. *BIT Numer. Math.* **29** (1989) 328–346.

[17] A.E. Ruehli and T.A. Johnson, Circuit analysis computing by waveform relaxation, in *Wiley Encyclopedia of Electrical Electronics Engineering*. Wiley, New York (1999).

[18] J.C. Strikwerda, *Finite difference schemes and partial differential equations*. Chapman and Hall (1989).