A mathematical analysis of optimized waveform relaxation for a small RC circuit

Mohammad Al-Khaleel

Yarmouk University, Irbid, Jordan

Martin J. Gander

University of Geneva, Switzerland

Albert E. Ruehli

Missouri U. of Science and Techn., Rolla, MO, USA

Abstract

Waveform relaxation techniques are an important tool for the simulation of very large scale circuits. They are based on a partition of the circuit into sub-circuits, and then use an iteration between sub-circuits to converge to the solution of the entire circuit. Their importance has increased with the wide availability of parallel computers with a large number of processors. Unfortunately classical waveform relaxation is hampered by slow convergence, but this can be addressed by better transmission conditions, which led to the new class of optimized waveform relaxation methods. In these methods, both voltage and current information is exchanged in a combination which can be optimized for the performance of the method. We prove in this paper a conjecture for the optimal combination for the particular case of a small RC circuit, and also present and analyze a transmission condition which includes a time derivative.

Keywords: Waveform Relxation, Circuit Simulation, Optimized Transmission Conditions

1. Introduction

Waveform Relaxation (WR) methods are based on partitioning large circuits into sub-circuits which then are solved separately for multiple time steps. Using an iteration and exchanging information between the sub-circuits, one arrives at better and better approximations of the solution of the original large scale circuit. The exchange of waveforms for multiple time steps reduces the importance of processor communication latency and is therefore a very suitable approach for the parallel simulation of large scale circuits. The classical WR algorithm was invented in 1980/81 in the circuit simulation community [21], and only later it was discovered that the algorithm is very much related to the

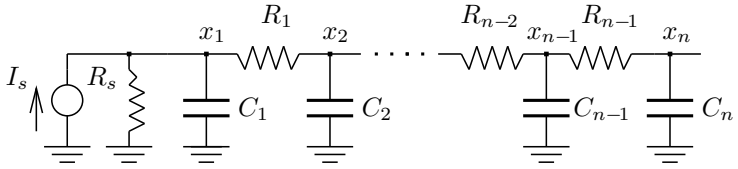

Figure 1: A finite size RC circuit.

classical Picard-Lindelöf iteration, see [22, 23, 24]. WR algorithms have been tested on a multitude of problems in the circuits area, see [25] and references therein, and also for evolution partial differential equations under the name Schwarz waveform relaxation, see [4, 20, 5, 16, 17, 18].

The limiting factor for classical waveform relaxation is the rather slow convergence for an important class of problems. This can however be overcome with the newly developed class of optimized WR or oWR, see [14, 12, 13, 8, 15], which has its roots in optimized Schwarz waveform relaxation methods, see [11, 9, 10, 3] and also the related approach for circuits in [7, 6]. In oWR, combinations of voltage and current values are transmitted at the interfaces between sub-circuits, and one can optimize the combination in order to obtain rapidly converging algorithms, which also led to the name oWR. In [2] a conjecture was stated on the optimal combination for a specific class of RC circuits. We prove in this paper the optimality conjecture for a small RC circuit of this class. We then introduce a transmission condition which also includes a time derivative, and leads to an even better algorithm.

Circuit equations are nowadays derived using Modified Nodal Analysis (MNA), see [19], which leads to equations of the form $\mathbf{C}\dot{\boldsymbol{x}}(t) + \mathbf{G}\boldsymbol{x}(t) = \mathbf{B}\boldsymbol{u}(t)$, where $\mathbf{C}$ contains the reactive elements, $\mathbf{G}$ the other elements, while $\mathbf{B}$ is the input selector matrix, and $\boldsymbol{u}(t)$ are the excitation or forcing functions. In [2], the model RC circuit in Fig. 1 was considered, and the MNA circuit equations were rewritten in tridiagonal form,

$$\dot{\boldsymbol{x}} = \begin{bmatrix} b_1 & c_1 & & & \\ a_1 & b_2 & c_2 & & \\ & a_2 & b_3 & c_3 & \\ & & a_3 & b_4 & \ddots \\ & & & \ddots & \ddots \end{bmatrix} \boldsymbol{x} + \boldsymbol{f}, \tag{1}$$

where

$$a_i = \frac{1}{R_i C_{i+1}}, \quad b_i = \begin{cases} -(\frac{1}{R_s} + \frac{1}{R_1})\frac{1}{C_1}, & i = 1 \\ -(\frac{1}{R_{i-1}} + \frac{1}{R_i})\frac{1}{C_i}, & i = 2, 3, \dots \\ -\frac{1}{R_{i-1}C_i}, & i = n \end{cases} \quad c_i = \frac{1}{R_i C_i},$$

and n is the number of nodes in the circuit. The source term on the right hand side is given by $\boldsymbol{f} = (I_s(t)/C_1, 0, \dots, 0)^T$, for some source function $I_s(t)$, and the initial voltages are given by $\boldsymbol{x}(0) = (v_1^0, v_2^0, \dots, v_n^0)^T$ at the time $t = 0$.

While in [2] classical and optimized WR algorithms were introduced for RC circuits of arbitrary size, we focus here on the particular case of $n = 4$ in order to prove the conjecture from [2]. For this case, the classical WR algorithm with two equal size sub-circuits is given by (see [14])

$$\begin{aligned} \dot{\boldsymbol{u}}^{k+1} &= \begin{bmatrix} b_1 & c_1 \\ a_1 & b_2 \end{bmatrix} \boldsymbol{u}^{k+1} + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} 0 \\ c_2 w_1^k \end{pmatrix}, \\ \dot{\boldsymbol{w}}^{k+1} &= \begin{bmatrix} b_3 & c_3 \\ a_3 & b_4 \end{bmatrix} \boldsymbol{w}^{k+1} + \begin{pmatrix} f_3 \\ f_4 \end{pmatrix} + \begin{pmatrix} a_2 u_2^k \\ 0 \end{pmatrix}, \end{aligned} \tag{2}$$

with corresponding initial conditions $\boldsymbol{u}^{k+1}(0) = (v_1^0, v_2^0)^T$ and $\boldsymbol{w}^{k+1}(0) = (v_3^0, v_4^0)^T$. To start the WR iteration, we need to specify two initial waveforms $u_2^0(t)$ and $w_1^0(t)$ for $t \in [0, T]$.

We see that in the relaxed terms in (2) the so called classical transmission conditions

$$u_3^{k+1} = w_1^k, \quad w_0^{k+1} = u_2^k \tag{3}$$

were used. Using a Laplace transform in time with Laplace parameter s, it was shown in [14] that $\hat{u}_2^{2k} = (\rho_{cla})^k \hat{u}_2^0$, and $\hat{w}_1^{2k} = (\rho_{cla})^k \hat{w}_1^0$, with the convergence factor

$$\rho_{cla}(s, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}) = \tfrac{c_2(s-b_1)}{(s-b_1)(s-b_2)-a_1c_1} \cdot \tfrac{a_2(s-b_4)}{(s-b_3)(s-b_4)-a_3c_3}, \quad s = \eta + i\omega. \tag{4}$$

Furthermore, it was shown in [14] that ρ_{cla} reaches its maximum for $\omega = 0$. Hence, the low angular frequency components ω close to zero will cause difficulties and slow down the convergence of the algorithm. New transmission conditions were therefore proposed in [14], namely

$$\begin{aligned} (u_3^{k+1} - u_2^{k+1}) + \alpha u_3^{k+1} &= (w_1^k - w_0^k) + \alpha w_1^k, \\ (w_1^{k+1} - w_0^{k+1}) + \beta w_0^{k+1} &= (u_3^k - u_2^k) + \beta u_2^k. \end{aligned} \tag{5}$$

Applying these transmission condition (5) in the WR iteration leads to an optimizable WR algorithm

$$\begin{aligned} \dot{\boldsymbol{u}}^{k+1} &= \begin{bmatrix} b_1 & c_1 \\ a_1 & b_2 + \frac{c_2}{\alpha+1} \end{bmatrix} \boldsymbol{u}^{k+1} + \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} + \begin{pmatrix} 0 \\ c_2 w_1^k - \frac{c_2}{\alpha+1} w_0^k \end{pmatrix}, \\ \dot{\boldsymbol{w}}^{k+1} &= \begin{bmatrix} b_3 - \frac{a_2}{\beta-1} & c_3 \\ a_3 & b_4 \end{bmatrix} \boldsymbol{w}^{k+1} + \begin{pmatrix} f_3 \\ f_4 \end{pmatrix} + \begin{pmatrix} a_2 u_2^k + \frac{a_2}{\beta-1} u_3^k \\ 0 \end{pmatrix}, \end{aligned} \tag{6}$$

where the values u_3^k and w_0^k are determined by the transmission conditions (5). It was shown in [14], that $\hat{u}_2^{2k} = (\rho_{opt})^k \hat{u}_2^0$, and $\hat{w}_1^{2k} = (\rho_{opt})^k \hat{w}_1^0$, where the convergence factor ρ_{opt} is given by

$$\begin{aligned} &\rho_{opt}(s, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \alpha, \beta) = \\ &\quad -\tfrac{c_2(s-b_1)(\beta-1)+(s-b_1)(s-b_2)-a_1c_1}{((s-b_3)(s-b_4)-a_3c_3)(\beta-1)+a_2(s-b_4)} \cdot \tfrac{-a_2(s-b_4)(\alpha+1)+(s-b_3)(s-b_4)-a_3c_3}{((s-b_1)(s-b_2)-a_1c_1)(\alpha+1)+c_2(b_1-s)}. \end{aligned} \tag{7}$$

The best choice in the Laplace transformed domain for α and β in the transmission conditions (5) is thus

$$\alpha := \frac{-a_3c_3}{(s-b_4)a_2} + \frac{s-b_3}{a_2} - 1,\ \beta := \frac{a_1c_1}{(s-b_1)c_2} - \frac{s-b_2}{c_2} + 1,\ s \in \mathbb{C}, \quad (8)$$

since then the convergence factor (7) vanishes, and hence we have converges in two iterations, independently of the guess for the initial waveforms [14]. This optimal choice in the Laplace transformed domain depends however on the Laplace parameter s, and thus represents symbols of operators in time in the real domain. Therefore, in [14], an approximation by a constant was proposed, which leads to a very practical algorithm with remarkable improvement over the classical WR algorithm. The optimal choice of this constant was given as a conjecture in [2] for an RC circuit of arbitrary size. We will prove in the next section the optimality of this constant approximation. We then also analyze a first order approximation in detail.

2. Optimization of the constant approximation

We now assume that α and β are constants, and start by studying the convergence factor ρ_{opt} in (7), whose analyticity in the right half of the complex plane is proved in [14] under the conditions that

$$\alpha > \frac{c_2|b_1|}{b_1b_2 - a_1c_1} - 1 =: \underline{\alpha}, \qquad \beta < -\frac{a_2|b_4|}{b_3b_4 - a_3c_3} + 1 =: \overline{\beta}. \quad (9)$$

Under these conditions, the maximum of ρ_{opt} for $s = \eta + i\omega$, $\eta \geq 0$, is attained on the boundary. Since, the limit of ρ_{opt} for $s := re^{i\theta}$, where $-\frac{\pi}{2} < \theta < \frac{\pi}{2}$, as $r \to \infty$ is $\left(\frac{-1}{(\alpha+1)(\beta-1)}\right)$, the same limit in all directions, we see that it suffices to minimize the convergence factor for $\eta = \text{const}$. In order to get an optimized algorithm for the L^2 norm in time, we need hence to solve the min-max problem

$$\min_{\alpha>\underline{\alpha},\beta<\overline{\beta}} \left(\max_{|\omega|<\infty} |\rho_{opt}(i\omega, \boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}, \alpha, \beta)| \right). \quad (10)$$

One can see from (7) that the modulus of ρ_{opt} for $s = i\omega$ depends on ω^2 only, since $|c_2(i\omega - b_1)(\beta - 1) + (i\omega - b_1)(i\omega - b_2) - a_1c_1|$ depends only on ω^2, and similarly for the other terms. Hence, it suffices to optimize for non-negative frequencies, $\omega \geq 0$.

We use now the same simplifying assumptions that led to the conjecture in [2], namely that all circuit elements are the same,

$$c_i = a_i = a \text{ for } i = 1, 2, 3, \quad b_i = b \text{ for } i = 1, 2, 3, 4. \quad (11)$$

In this case, we see from the optimal choice in (8) that the symmetry assumption $\beta = -\alpha$ also made in [2] is reasonable. This reduces the optimization parameters to α, and leads to the simplified convergence factor

$$\rho_{opt0}(i\omega, a, b, \alpha) = \left(\frac{\alpha + 1 - \lambda}{(\alpha+1)\lambda - 1}\right)^2, \quad (12)$$

where

$$\lambda := \frac{(s-b)^2 - a^2}{a(s-b)} = \frac{-b}{a} + \frac{1}{a}\frac{s(s-b)-a^2}{s-b}, \quad s = i\omega. \tag{13}$$

We next show an important property of λ.

Lemma 2.1. *Let $b < 0$, $a > 0$, and $-b \geq 2a$. Then the modulus of λ in (13) is larger than one in the right half of the complex plane.*

Proof The modulus of λ is given by

$$|\lambda| = \frac{|(s-b)^2 - a^2|}{|a(s-b)|} = \frac{|(s+|b|)^2 - a^2|}{|a(s+|b|)|},$$

and using a time scaling of the equations, i.e. taking $a = 1$, $b = -2c^2$, and $c \geq 1$, we have

$$|\lambda| = \sqrt{\frac{(\eta^2+1+\omega^2+4\eta c^2+2\eta+4c^2+4c^4)(\eta^2+1+\omega^2+4\eta c^2-2\eta-4c^2+4c^4)}{\eta^2+4\eta c^2+4c^4+\omega^2}}.$$

This shows that the modulus $|\lambda|$ is bigger than one, since the first factor in the numerator of the argument under the square root is bigger than the denominator, and the second factor is bigger than one for $c \geq 1$ and $\eta > 0$. Hence, $|\lambda| > 1$ in the right half of the complex plane. ■

Using this lemma, we see that it suffices in this case to require $\alpha > 0$ and the convergence factor ρ_{opt0} is still analytic in the right half of the complex plane. Nevertheless, as we will see later, the optimized value of α will be bigger than $\underline{\alpha}$, but for the change of variables that will follow, it is convenient to work with the simpler assumption $\alpha > 0$.

We now use a change of variables based on the real part of $z := \frac{s(s-b)-a^2}{s-b}$, $s = i\omega$, $\omega \geq 0$, which appears in λ. We write z as

$$z := x + iy = \Re\left(\frac{s(s-b)-a^2}{s-b}\right) + \Im\left(\frac{s(s-b)-a^2}{s-b}\right) i. \tag{14}$$

Hence, we have

$$\lambda = \frac{-b}{a} + \frac{1}{a}(x+iy). \tag{15}$$

The optimal value of α in (8) can be written in terms of λ as $\alpha = \lambda - 1$. Moreover, if p is a free parameter corresponding to a constant approximation of $x + iy$ in (15), then a constant approximation of α is $\alpha = \frac{-b}{a} - 1 + \frac{p}{a}$.

In the new variable x, and the new parameter p for the constant approximation, the convergence factor (12) in modulus becomes

$$|\rho_{opt0}(x,a,b,p)| = \frac{-a^2((2b^2+2bp-a^2)x^2+(-2ba^2+b^3-bp^2)x-b^2a^2)}{(-4b^4-4b^2p^2+8b^3p+3b^2a^2+a^2p^2-4bpa^2)x^2+(2a^2p^2b+a^4b-2b^2pa^2)x+b^4a^2-2b^3pa^2+p^2b^2a^2}.$$

Factorizing a^6 from the denominator and numerator, implies

$$|\rho_{opt0}(x,a,b,p)| = \frac{Q_1(x,a,b,p)}{Q_2(x,a,b,p)},$$

where

$$
\begin{aligned}
Q_1 &:= -\left(\left(2\left(\tfrac{b}{a}\right)^2+2\left(\tfrac{b}{a}\right)\left(\tfrac{p}{a}\right)-1\right)\left(\tfrac{x}{a}\right)^2+\left(\left(\tfrac{b}{a}\right)^3-2\left(\tfrac{b}{a}\right)-\left(\tfrac{b}{a}\right)\left(\tfrac{p}{a}\right)^2\right)\left(\tfrac{x}{a}\right)-\left(\tfrac{b}{a}\right)^2\right),\\
Q_2 &:= \left(-4\left(\tfrac{b}{a}\right)^4-4\left(\tfrac{b}{a}\right)^2\left(\tfrac{p}{a}\right)^2+8\left(\tfrac{b}{a}\right)^3\left(\tfrac{p}{a}\right)+3\left(\tfrac{b}{a}\right)^2+\left(\tfrac{p}{a}\right)^2-4\left(\tfrac{b}{a}\right)\left(\tfrac{p}{a}\right)\right)\left(\tfrac{x}{a}\right)^2\\
&\quad+\left(2\left(\tfrac{p}{a}\right)^2\left(\tfrac{b}{a}\right)+\left(\tfrac{b}{a}\right)-2\left(\tfrac{b}{a}\right)^2\left(\tfrac{p}{a}\right)\right)\left(\tfrac{x}{a}\right)\\
&\quad+\left(\tfrac{b}{a}\right)^4-2\left(\tfrac{b}{a}\right)^3\left(\tfrac{p}{a}\right)+\left(\tfrac{p}{a}\right)^2\left(\tfrac{b}{a}\right)^2.
\end{aligned}
$$

We set $\tilde{p}=\frac{p}{a}$, and in addition, we set $\frac{b}{a}=-2c^2$, $c\geq 1$, since $|b|\geq 2a$, to eliminate one parameter, and assume $\tilde{x}=\frac{x}{a}$, where $\tilde{x}\in[\frac{1}{b/a},0)\equiv[-\frac{1}{2c^2},0)$. Since we have $\alpha>0$, the new parameter $\tilde{p}$ should satisfy $\tilde{p}>1-2c^2$. Furthermore, the modulus of the convergence factor (12) now simplifies to

$$
R_0(\tilde{x},c,\tilde{p})=-\frac{(4c^2\tilde{p}-8c^4+1)\tilde{x}^2+(8c^6-4c^2-2c^2\tilde{p}^2)\tilde{x}+4c^4}{(64c^8-8c^2\tilde{p}-12c^4+64c^6\tilde{p}+16c^4\tilde{p}^2-\tilde{p}^2)\tilde{x}^2+(2c^2+4c^2\tilde{p}^2+8c^4\tilde{p})\tilde{x}-16c^8-16c^6\tilde{p}-4c^4\tilde{p}^2},
\tag{16}
$$

where $\tilde{x}\in[-\frac{1}{2c^2},0)$, and the min-max problem (10) becomes

$$
\min_{\tilde{p}>(1-2c^2)}\left(\max_{\frac{-1}{2c^2}\leq\tilde{x}<0}R_0(\tilde{x},c,\tilde{p})\right),\ c\geq 1.
\tag{17}
$$

To analyze the min-max problem (17), we need the following nine technical lemmas:

Lemma 2.2. *For $\tilde{p}>1-2c^2$, and $c\geq 1$, the polynomial L defined by*

$$
L(c,\tilde{p}):=(16c^4-1)\tilde{p}^2+4c^2\tilde{p}-64c^8-1+28c^4
\tag{18}
$$

has a unique real root at

$$
\tilde{p}_+=\frac{-2c^2+\sqrt{48c^4+1024c^{12}-512c^8-1}}{16c^4-1}.
\tag{19}
$$

Moreover, $\tilde{p}_+>0$, and $L(c,\tilde{p})>0$ for $\tilde{p}>\tilde{p}_+$, and $L(c,\tilde{p})<0$ for $1-2c^2<\tilde{p}<\tilde{p}_+$.

Proof The polynomial L has two real roots $\tilde{p}_\pm$, which are given by

$$
\tilde{p}_\pm=\frac{-2c^2\pm\sqrt{48c^4+1024c^{12}-512c^8-1}}{16c^4-1}.
\tag{20}
$$

The first root $\tilde{p}_+$ satisfies $\tilde{p}_+>0>1-2c^2$, since

$$
\begin{aligned}
\tilde{p}_+>0 &\iff -2c^2+\sqrt{48c^4+1024c^{12}-512c^8-1}>0\\
&\iff 48c^4+1024c^{12}-512c^8-1>4c^4\\
&\iff 44c^4+1024c^{12}-512c^8-1>0,
\end{aligned}
$$

and the last inequality is true for $c \geq 1$, since the coefficient of the term c^{12} is the dominant one.

For the second root $\tilde{p}_-$, we have $\tilde{p}_- < 1 - 2c^2$ since

$$\begin{aligned} \tilde{p}_- < 1 - 2c^2 &\iff -2c^2 - \sqrt{48c^4 + 1024c^{12} - 512c^8 - 1} < (1 - 2c^2)(16c^4 - 1) \\ &\iff 32c^6 - 16c^4 - 4c^2 + 1 < \sqrt{48c^4 + 1024c^{12} - 512c^8 - 1}, \end{aligned}$$

and now observing that both sides are positive, we can square and simplify to obtain

$$\iff -1024c^{10} + 192c^6 - 64c^4 - 8c^2 + 2 + 512c^8 < 0,$$

and this last inequality holds for $c \geq 1$. Hence, $\tilde{p}_-$ can be discarded.

Since $L(c, \tilde{p})$ is positive for large $\tilde{p}$, because of the sign of the coefficient of $\tilde{p}^2$ which is positive, we have $L(c, \tilde{p}) > 0$ for $\tilde{p} > \tilde{p}_+$, and for $1 - 2c^2 < \tilde{p} < \tilde{p}_+$ the polynomial $L(c, \tilde{p}) < 0$. ■

Lemma 2.3. *For $\tilde{p} > 1 - 2c^2$, and $c \geq 1$, the polynomial d given by*

$$\begin{aligned} d(c, \tilde{p}) = (1 - 16c^4)\tilde{p}^4 - 4c^2\tilde{p}^3 + (128c^8 - 32c^4 + 2)\tilde{p}^2 \\ +(80c^6 - 4c^2)\tilde{p} + 240c^8 - 32c^4 - 256c^{12} + 1 \end{aligned} \tag{21}$$

has only two real roots, say $\tilde{p}_1$ and $\tilde{p}_2$ with $\tilde{p}_1 < \tilde{p}_2$, which are both larger than zero, and has no roots in the interval $(1 - 2c^2, 0]$. Furthermore, d satisfies the inequalities

i) $d(c, \tilde{p}) < 0$ for $\tilde{p} \in (1 - 2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$,

ii) $d(c, \tilde{p}) \geq 0$ for $\tilde{p} \in [\tilde{p}_1, \tilde{p}_2]$.

Proof The polynomial $d(c, \tilde{p})$ is negative for large $\tilde{p}$, because of the sign of the coefficient of $\tilde{p}^4$ which is negative. Moreover, d takes positive values, e.g. at $\tilde{p} = 2c^2$, $d(c, 2c^2) = (16c^4 - 1)^2 > 0$, hence, it must have by continuity and the Intermediate Value Theorem at least one real root $\tilde{p}_2(c) > 2c^2 > 1 - 2c^2$, $d(c, \tilde{p}_2) = 0$. An example of d with $c = 1$ and $\tilde{p} > 1 - 2c^2$ is given in Figure 2. To show that d has exactly two roots bigger than $1 - 2c^2$, say $\tilde{p}_1$ and $\tilde{p}_2$, and $\tilde{p}_2 > \tilde{p}_1 > 0$, we use the derivative of $d(c, \tilde{p})$ with respect to $\tilde{p}$. The derivative

$$\frac{d}{d\tilde{p}}(d(c, \tilde{p})) = 4(1 - 16c^4)\tilde{p}^3 - 12c^2\tilde{p}^2 + 2(128c^8 - 32c^4 + 2)\tilde{p} - 4c^2 + 80c^6$$

has two real roots, say $r_1, r_2 > 1 - 2c^2$, and a third real root, say r_3, less than $1 - 2c^2$, since

$$\begin{aligned} \tfrac{d}{d\tilde{p}}(d(c, \tilde{p}))\big|_{\tilde{p}=-2c^2} &= 4c^2(32c^4 - 3) > 0, \\ \tfrac{d}{d\tilde{p}}(d(c, \tilde{p}))\big|_{\tilde{p}=1-2c^2} &= -8(4c^4 - 2c^2 - 1)(4c^2 - 1)^2 < 0, \\ \tfrac{d}{d\tilde{p}}(d(c, \tilde{p}))\big|_{\tilde{p}=0} &= 4c^2(20c^4 - 1) > 0, \\ \tfrac{d}{d\tilde{p}}(d(c, \tilde{p}))\big|_{\tilde{p}=2c^2} &= -4c^2(16c^4 - 1) < 0. \end{aligned}$$

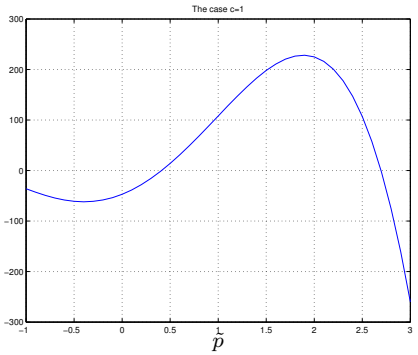

Figure 2: The polynomial d for the case $c = 1$.

Therefore, by the Intermediate Value Theorem, we have $r_3 \in (-2c^2, 1-2c^2)$, which can be discarded, $r_2 \in (1-2c^2, 0)$ which is a minimum, and $r_1 \in (0, 2c^2)$ which is a maximum.

Now, we have $d(c, 1-2c^2) = -4(4c^4 - 2c^2 - 1)(4c^2 - 1)^2 < 0$, and then d decreases to more negative values until d reaches its minimum at r_2, after that d starts increasing to its maximum at $r_1 \in (0, 2c^2)$, which is a positive value since $d(c, 2c^2) = (16c^4 - 1)^2 > 0$, and r_1, where the maximum is attained, is less than $2c^2$, so here $d(c, \tilde{p})$ has a real root which is $\tilde{p}_1 > 1 - 2c^2$, and more than that, we have $\tilde{p}_1 > 0$ since at $\tilde{p} = 0$, $d(c, \tilde{p})$ is negative. After d reaches its maximum at r_1, it starts decreasing again to minus infinity, so here d has its second root $\tilde{p}_2 > \tilde{p}_1 > 0$, and there are no more roots, since d decreases to minus infinity. Therefore, $d(c, \tilde{p})$ has only two roots, $\tilde{p}_2 > \tilde{p}_1 > 0$, for $\tilde{p} > 1 - 2c^2$, and $c \geq 1$, and no roots in the interval $(1 - 2c^2, 0]$. Moreover, d falls under one of the two cases which are stated in the Lemma, and the proof is complete. ■

Lemma 2.4. *For $c \geq 1$, the root $\tilde{p}_+$ given in (19) lies in the interval $[\tilde{p}_1, \tilde{p}_2]$, where $\tilde{p}_1$ and $\tilde{p}_2$ are the two real roots of d which are characterized by Lemma 2.3.*

Proof By Lemma 2.2, we have $\tilde{p}_+ > 0$ for $c \geq 1$. Now, since $d(c, \tilde{p}_+) > 0$, $\tilde{p}_+$ must lie in the interval $[\tilde{p}_1, \tilde{p}_2]$, by Lemma 2.3. ■

Lemma 2.5. *For $\tilde{p} \in [\tilde{p}_1, \tilde{p}_2]$, and $c \geq 1$, the polynomial P_2 defined by*

$$P_2(c, \tilde{p}) = (1 - 16c^4)\tilde{p}^2 - 4c^2(1 + 4c^4)\tilde{p} + 32c^8 - 28c^4 + 1$$

is always negative.

Proof Using the Intermediate Value Theorem, one can show that the two roots of P_2 are $r_- \in (-2c^2, 1 - 2c^2)$ and $r_+ \in (0, 2c^2)$. By finding the precise r_+, and substituting it into d in (21), we have $d(c, r_+) < 0$, and since $d(c, \tilde{p}) \geq 0$ for $\tilde{p} \in [\tilde{p}_1, \tilde{p}_2]$, Lemma 2.3, and $r_+ < 2c^2$, we have $r_- < r_+ < \tilde{p}_1$, and the two zeros

$r_\pm$ are not in $[\tilde{p}_1, \tilde{p}_2]$. In addition, the coefficient of $\tilde{p}^2$ is negative, which implies that the sign of P_2 is positive only for $\tilde{p} \in (r_-, r_+)$, and is negative everywhere else. Hence, the polynomial P_2 is always negative for $\tilde{p} \in [\tilde{p}_1, \tilde{p}_2]$. ■

Lemma 2.6. *For $\tilde{p} > 1 - 2c^2$, and $c \geq 1$, the polynomial P_4 defined by*

$$P_4(c, \tilde{p}) = L(c, \tilde{p}) \left((-16c^8 + 1 - 16c^4)\tilde{p}^2 - (4c^2 + 32c^6)\tilde{p} + 64c^{12} - 16c^8 + 1 - 28c^4 \right),$$

where L is given in (18), has only two real roots, $\tilde{p}_+$ given in (19) and another real root, say $\hat{\tilde{p}}$, and $1 - 2c^2 < \tilde{p}_1 < \hat{\tilde{p}} < \tilde{p}_+$, where $\tilde{p}_1$ is the real root of d in Lemma 2.3. Moreover, P_4 is negative for $\tilde{p} \in (1 - 2c^2, \hat{\tilde{p}}) \bigcup (\tilde{p}_+, \infty)$, and positive for $\tilde{p} \in (\hat{\tilde{p}}, \tilde{p}_+)$.

Proof For $\tilde{p} = \tilde{p}_+$, $\tilde{p}_-$ given in (20), the roots of L, we have $P_4(\tilde{p}) = 0$, which means $\tilde{p}_+ > 0$, $\tilde{p}_- < 1 - 2c^2$ are roots for $P_4(\tilde{p})$. One can also find the other two roots from $(1 - 16c^8 - 16c^4)\tilde{p}^2 - (4c^2 + 32c^6)\tilde{p} + 64c^{12} - 16c^8 + 1 - 28c^4 = 0$, which implies two roots, one is less than $1 - 2c^2$, and hence, it can be discarded, and another root $\hat{\tilde{p}} > 1 - 2c^2$, where $\hat{\tilde{p}} < \tilde{p}_+$, since $L(c, \hat{\tilde{p}}) < 0$, and $L(c, \tilde{p})$ is negative for all $\tilde{p} \in (1 - 2c^2, \tilde{p}_+)$, by Lemma 2.2. Therefore, $P_4(\tilde{p})$ has exactly two roots bigger than $1 - 2c^2$, which are $\tilde{p}_+$ and $\hat{\tilde{p}}$. Furthermore, $d(c, \hat{\tilde{p}}) > 0$, which means $\tilde{p}_1 < \hat{\tilde{p}}$, since $1 - 2c^2 < \hat{\tilde{p}}$, and $d(c, \tilde{p})$ is positive in $(\tilde{p}_1, \tilde{p}_2)$, by Lemma 2.3. Therefore, from the sign of P_4, the polynomial P_4 is negative for $\tilde{p} \in (1 - 2c^2, \hat{\tilde{p}}) \bigcup (\tilde{p}_+, \infty)$, and positive for $\tilde{p} \in (\hat{\tilde{p}}, \tilde{p}_+)$. ■

Lemma 2.7. *For $\tilde{p} > 1 - 2c^2$, and $c \geq 1$, let $x_1(c, \tilde{p})$ be given by*

$$x_1(c, \tilde{p}) = \frac{(16c^4 + 8c^2\tilde{p} + 2\sqrt{d})c^2}{L},$$

where L and d are given in (18) and (21) respectively. Then x_1 is not defined for $\tilde{p} = \tilde{p}_+$, and is complex for $\tilde{p} \in (1 - 2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$. Furthermore, $x_1 < \frac{-1}{2c^2}$ for $[\tilde{p}_1, \tilde{p}_+)$, and $x_1 > 0$ for $(\tilde{p}_+, \tilde{p}_2]$.

Proof By Lemma 2.2, the denominator of x_1 is zero at $\tilde{p} = \tilde{p}_+$, and by Lemma 2.3, $d(c, \tilde{p}) < 0$ for $\tilde{p} \in (1 - 2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$. Hence, x_1 is not defined for $\tilde{p} = \tilde{p}_+$, and is complex for $\tilde{p} \in (1 - 2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$. For $\tilde{p} \in (\tilde{p}_+, \tilde{p}_2]$, we have $x_1 > 0$, since it is a fraction of two positive quantities, Lemmas 2.2 and 2.3. Consider now the interval $[\tilde{p}_1, \tilde{p}_+)$, then

$$\begin{aligned}
x_1 < \frac{-1}{2c^2} \quad &\iff \quad \frac{(8\tilde{p}c^2 + 16c^4 + 2\sqrt{d})c^2}{L} < \frac{-1}{2c^2} \\
&\iff \quad (8\tilde{p}c^2 + 16c^4 + 2\sqrt{d})c^2 > \frac{-L}{2c^2} \\
&\qquad \text{(note that } L < 0 \text{ in the interval considered)} \\
&\iff \quad 4c^4\sqrt{d} > -L - 16\tilde{p}c^6 - 32c^8 \\
&\iff \quad 4c^4\sqrt{d} > (1 - 16c^4)\tilde{p}^2 - 4c^2(1 + 4c^4)\tilde{p} + 32c^8 - 28c^4 + 1.
\end{aligned}$$

Now, since in the interval considered, the left hand side is positive, Lemma 2.3, and the right hand side is negative, Lemma 2.5, the last inequality is true and we have $x_1 < \frac{-1}{2c^2}$. ■

Lemma 2.8. *For $\tilde{p} > 1 - 2c^2$, and $c \geq 1$, let $x_2(c, \tilde{p})$ be given by*

$$x_2(c, \tilde{p}) = \frac{(16c^4 + 8c^2\tilde{p} - 2\sqrt{d})c^2}{L},$$

where L and d are given in (18) and (21) respectively. Then x_2 is not defined for $\tilde{p} = \tilde{p}_+$, and is complex for $\tilde{p} \in (1 - 2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$. Furthermore, $x_2 \geq \frac{-1}{2c^2}$ for $\tilde{p} \in [\hat{\tilde{p}}, \tilde{p}_+) \bigcup (\tilde{p}_+, \tilde{p}_2]$, and $x_2 < \frac{-1}{2c^2}$ for $[\tilde{p}_1, \hat{\tilde{p}})$. In addition, $x_2 < 0$ for $\tilde{p} \in [\tilde{p}_1, \tilde{p}_+) \bigcup (\tilde{p}_+, \tilde{\tilde{p}})$, and $x_2 \geq 0$ for $[\tilde{\tilde{p}}, \tilde{p}_2]$, where $\tilde{\tilde{p}} = \sqrt{4c^4 - 1}$, and $\tilde{p}_+ < \tilde{\tilde{p}} < \tilde{p}_2$.

Proof The proof is similar to the proof of Lemma 2.7. By Lemma 2.2, the denominator of x_2 is zero at $\tilde{p} = \tilde{p}_+$, and by Lemma 2.3, $d(c, \tilde{p}) < 0$ for $\tilde{p} \in (1 - 2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$. Hence, x_2 is not defined for $\tilde{p} = \tilde{p}_+$, and is complex for $\tilde{p} \in (1 - 2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$. For $\tilde{p} \in [\tilde{p}_1, \tilde{p}_+)$, we have

$$\begin{aligned}
x_2 > \frac{-1}{2c^2} &\iff \frac{(8\tilde{p}c^2 + 16c^4 - 2\sqrt{d})c^2}{L} > \frac{-1}{2c^2} \\
&\iff (8\tilde{p}c^2 + 16c^4 - 2\sqrt{d})c^2 < \frac{-L}{2c^2} \\
&\iff 4c^4\sqrt{d} > L + 16\tilde{p}c^6 + 32c^8 \\
&\iff 4c^4\sqrt{d} > -((1 - 16c^4)\tilde{p}^2 - 4c^2(1 + 4c^4)\tilde{p} + 32c^8 - 28c^4 + 1),
\end{aligned}$$

and now the right hand side equals minus the polynomial P_2 studied in Lemma 2.5, hence both sides are positive, and we can square and simplify to obtain

$$\iff L(c, \tilde{p}) \left((-16c^8 + 1 - 16c^4)\tilde{p}^2 - (4c^2 + 32c^6)\tilde{p} + 64c^{12} - 16c^8 + 1 - 28c^4\right) > 0.$$

The left hand side is the polynomial P_4 studied in Lemma 2.6. Therefore, by Lemma 2.6, we have $x_2 < \frac{-1}{2c^2}$ for $\tilde{p} \in [\tilde{p}_1, \hat{\tilde{p}})$, and for $\tilde{p} \in [\hat{\tilde{p}}, \tilde{p}_+)$, we have $x_2 \geq \frac{-1}{2c^2}$.

Consider now the interval $(\tilde{p}_+, \tilde{p}_2]$, in which we have

$$\begin{aligned}
x_2 > \frac{-1}{2c^2} &\iff \frac{(8\tilde{p}c^2 + 16c^4 - 2\sqrt{d})c^2}{L} > \frac{-1}{2c^2} \\
&\iff (8\tilde{p}c^2 + 16c^4 - 2\sqrt{d})c^2 > \frac{-L}{2c^2} \\
&\iff 4c^4\sqrt{d} < L + 16\tilde{p}c^6 + 32c^8 \\
&\iff 4c^4\sqrt{d} < -((1 - 16c^4)\tilde{p}^2 - 4c^2(1 + 4c^4)\tilde{p} + 32c^8 - 28c^4 + 1) \\
&\iff L(c, \tilde{p}) \left((-16c^8 + 1 - 16c^4)\tilde{p}^2 - (4c^2 + 32c^6)\tilde{p}\right. \\
&\qquad \left. + 64c^{12} - 16c^8 + 1 - 28c^4\right) < 0.
\end{aligned}$$

The left hand side is again the polynomial P_4 studied in Lemma 2.6, and thus, $x_2 > \frac{-1}{2c^2}$ for $\tilde{p} \in (\tilde{p}_+, \tilde{p}_2]$. Therefore, $x_2 \geq \frac{-1}{2c^2}$ for $\tilde{p} \in [\hat{\tilde{p}}, \tilde{p}_+) \bigcup (\tilde{p}_+, \tilde{p}_2]$, and $x_2 < \frac{-1}{2c^2}$ for $[\tilde{p}_1, \hat{\tilde{p}})$.

For $\tilde{p} \in [\tilde{p}_1, \tilde{p}_+)$, we also have

$$\begin{aligned} x_2 < 0 &\iff \frac{(8\tilde{p}c^2 + 16c^4 - 2\sqrt{d})c^2}{L} > \frac{-1}{2c^2} \\ &\iff 4\tilde{p}c^2 + 8c^4 - \sqrt{d} > 0 \\ &\iff (\tilde{p} - \tilde{\tilde{p}})(\tilde{p} + \tilde{\tilde{p}})(\tilde{p} - \tilde{p}_+)(\tilde{p} - \tilde{p}_-) > 0, \end{aligned}$$

where $\tilde{\tilde{p}} = \sqrt{4c^4 - 1}$, and $\tilde{p}_+$, $\tilde{p}_-$ are given in (20). Therefore, the only two roots of the left hand side in the last inequality above, which are bigger than $1 - 2c^2$, are $\tilde{p}_+$ and $\tilde{\tilde{p}}$. Moreover, since $L(c, \tilde{p})$ is positive for $\tilde{p} = \tilde{\tilde{p}} > 1 - 2c^2$, we have $\tilde{p}_+ < \tilde{\tilde{p}}$, by Lemma 2.2. Also, since $d(c, \tilde{p})$ at $\tilde{p} = \tilde{\tilde{p}}$ is positive, and $\tilde{\tilde{p}} > \tilde{p}_+$, we have $\tilde{\tilde{p}} \in (\tilde{p}_1, \tilde{p}_2)$, by Lemma 2.3. By studying the sign of the left hand side expression in the last inequality, $[(\tilde{p} - \tilde{\tilde{p}})(\tilde{p} + \tilde{\tilde{p}})(\tilde{p} - \tilde{p}_+)(\tilde{p} - \tilde{p}_-)]$, we see that $x_2 < 0$ for $\tilde{p} \in [\tilde{p}_1, \tilde{p}_+)$.

Consider now the interval $(\tilde{p}_+, \tilde{p}_2]$, in which we have

$$\begin{aligned} x_2 < 0 &\iff \frac{(8\tilde{p}c^2 + 16c^4 - 2\sqrt{d})c^2}{L} > \frac{-1}{2c^2} \\ &\iff 4\tilde{p}c^2 + 8c^4 - \sqrt{d} < 0 \\ &\iff (\tilde{p} - \tilde{\tilde{p}})(\tilde{p} + \tilde{\tilde{p}})(\tilde{p} - \tilde{p}_+)(\tilde{p} - \tilde{p}_-) < 0, \end{aligned}$$

which implies that, $x_2 < 0$ for $\tilde{p} \in (\tilde{p}_+, \tilde{\tilde{p}})$, and $x_2 \geq 0$ for $\tilde{p} \in [\tilde{\tilde{p}}, \tilde{p}_2]$. Therefore, $x_2 < 0$ for $\tilde{p} \in [\tilde{p}_1, \tilde{p}_+) \bigcup (\tilde{p}_+, \tilde{\tilde{p}})$, and $x_2 \geq 0$ for $[\tilde{\tilde{p}}, \tilde{p}_2]$. ■

Lemma 2.9. *For $c \geq 1$, the function $\tilde{x} \mapsto R_0(\tilde{x}, c, \tilde{p})$ defined in (16) has a unique local minimum in $[\frac{-1}{2c^2}, 0)$, located at*

$$\underline{\tilde{x}}(c, \tilde{p}) = \frac{(16c^4 + 8c^2\tilde{p} - 2\sqrt{d(c, \tilde{p})})c^2}{(16c^4 - 1)\tilde{p}^2 + 4c^2\tilde{p} - 64c^8 - 1 + 28c^4}, \tag{22}$$

where $d(c, \tilde{p})$ is given in (21), if $\tilde{p} \in [\hat{\tilde{p}}, \tilde{p}_+) \bigcup (\tilde{p}_+, \tilde{\tilde{p}})$, where $\tilde{p}_+$, $\hat{\tilde{p}}$ and $\tilde{\tilde{p}}$ are determined by Lemmas 2.2, 2.6, and 2.8 respectively. For any other value of $\tilde{p} > 1 - 2c^2$, R_0 has no extrema in $[\frac{-1}{2c^2}, 0)$.

Proof A partial derivative of $R_0(\tilde{x}, c, \tilde{p})$ with respect to $\tilde{x}$ shows that the roots of the polynomial

$$Q(\tilde{x}) = (\tilde{p} - 1 + 2c^2)(\tilde{p} + 1 + 2c^2)P(\tilde{x}),$$

where $P(\tilde{x})$ is given by

$$\begin{aligned} P(\tilde{x}) = &-2c^2(16c^4\tilde{p}^2 - \tilde{p}^2 + 4c^2\tilde{p} + 28c^4 - 64c^8 - 1)\tilde{x}^2 \\ &+2c^2(16c^4\tilde{p} + 32c^6)\tilde{x} - 2c^2(4c^4\tilde{p}^2 + 4c^4 - 16c^8), \end{aligned} \tag{23}$$

determine the extrema of R_0. Since $(\tilde{p}-1+2c^2)(\tilde{p}+1+2c^2) > 0$ for $c \geq 1$ and $\tilde{p} > 1-2c^2$, we have $Q(\tilde{x}) = 0 \Longleftrightarrow P(\tilde{x}) = 0$, with the same coefficient signs. The polynomial $P(\tilde{x})$ has two roots $\overline{\tilde{x}}$ and $\underline{\tilde{x}}$ given by

$$\overline{\tilde{x}}(c,\tilde{p}) = \frac{(16c^4+8c^2\tilde{p}+2\sqrt{d(c,\tilde{p})})c^2}{(16c^4-1)\tilde{p}^2+4c^2\tilde{p}-64c^8-1+28c^4},$$
$$\underline{\tilde{x}}(c,\tilde{p}) = \frac{(16c^4+8c^2\tilde{p}-2\sqrt{d(c,\tilde{p})})c^2}{(16c^4-1)\tilde{p}^2+4c^2\tilde{p}-64c^8-1+28c^4},$$

and $d(c,\tilde{p})$ is given in (21). Note that, $\overline{\tilde{x}}$ and $\underline{\tilde{x}}$ are the same x_1 and x_2 which are given in Lemmas 2.7 and 2.8, respectively. By Lemmas 2.7 and 2.8, $\overline{\tilde{x}}$ and $\underline{\tilde{x}}$ are not defined for $\tilde{p} = \tilde{p}_+$, and are complex for $\tilde{p} \in (1-2c^2, \tilde{p}_1) \bigcup (\tilde{p}_2, \infty)$. Therefore, we analyze for the intervals $[\tilde{p}_1, \tilde{p}_+)$ and $(\tilde{p}_+, \tilde{p}_2]$. Now, by Lemmas 2.7 and 2.8, R_0 has only one extremum in $[\frac{-1}{2c^2}, 0)$ at $\tilde{x} = \underline{\tilde{x}}$ if $\tilde{p} \in [\hat{\tilde{p}}, \tilde{p}_+) \bigcup (\tilde{p}_+, \tilde{\tilde{p}})$. By studying the sign of $Q(\tilde{x})$, it is a minimum.

For any other value of $\tilde{p} > 1-2c^2$, R_0 has no extrema in $\tilde{x}$, because either the extrema are not defined or are not in $[\frac{-1}{2c^2}, 0)$, see Lemmas 2.7 and 2.8. ■

Note that if $\tilde{p} = \tilde{p}_+$, which is the zero of the denominator of $\underline{\tilde{x}}$ that makes it not defined, then the polynomial that determines the extrema of R_0 is reduced to a polynomial of degree one, and is given by

$$P_r(\tilde{x}) = 2c^2(16c^4\tilde{p}_+ + 32c^6)\tilde{x} - 2c^2(4c^4\tilde{p}_+^2 + 4c^4 - 16c^8), \tag{24}$$

and has only one zero, given by

$$\tilde{x}_r := \frac{\tilde{p}_+^2 + 1 - 4c^4}{4\tilde{p}_+ + 8c^2} = \frac{-c^2}{16c^4-1},$$

and $\tilde{x}_r \in [\frac{-1}{2c^2}, 0)$. Since the reduced polynomial $P_r(\tilde{x})$ in (24) is just a line, and the coefficient of $\tilde{x}$ is positive, $\tilde{x}_r$ is a minimum. This case is not however of our interest, since if we take $\tilde{p} = \tilde{p}_+$, then we already have the value of the parameter $\tilde{p}$ which we want to optimize, and no more optimization process.

Lemma 2.10. *For fixed $\tilde{x} \in [-\frac{1}{2c^2}, 0)$, and $\tilde{p} > 1-2c^2$, we have $\frac{\partial R_0(\tilde{x},c,\tilde{p})}{\partial \tilde{p}}(\tilde{p} - \underline{\tilde{p}}(\tilde{x},c)) \geq 0$, where $\underline{\tilde{p}}(\tilde{x},c)$ is given by*

$$\underline{\tilde{p}}(\tilde{x},c) = \frac{6c^2\tilde{x} - 4c^4 - 16\tilde{x}c^6 - \tilde{x}^2 + 8\tilde{x}^2c^4 - \sqrt{d(\tilde{x},c)}}{2(4c^4\tilde{x} + 2c^2\tilde{x}^2)}, \tag{25}$$

and $d(\tilde{x},c)$ is given by

$$d(\tilde{x},c) = \left((16c^4-4c^2-1)\tilde{x}^2 + (6c^2-8c^4)\tilde{x} - 4c^4\right) \left((16c^4+4c^2-1)\tilde{x}^2 + (8c^4+6c^2)\tilde{x} - 4c^4\right). \tag{26}$$

Proof A partial derivative of $R_0(\tilde{x},c,\tilde{p})$ with respect to $\tilde{p}$ shows that the roots of the polynomial

$$Q(\tilde{p}) = -\left((16c^4-1)\tilde{x}^2 + 2c^2\tilde{x} - 4c^4\right) P(\tilde{p}),$$

where $P(\tilde{p})$ is given by

$$\begin{aligned} P(\tilde{p}) = &-(4c^2\tilde{x}^2 + 8c^4\tilde{x})\tilde{p}^2 - (32\tilde{x}c^6 - 16\tilde{x}^2c^4 - 12c^2\tilde{x} + 8c^4 + 2\tilde{x}^2)\tilde{p} \\ &-32\tilde{x}c^8 + 48\tilde{x}^2c^6 - 16c^6 + 16c^4\tilde{x} - 8c^2\tilde{x}^2, \end{aligned} \tag{27}$$

determine the extrema of R_0. For $\tilde{x} \in [-\frac{1}{2c^2}, 0)$ with $c \geq 1$, we have $-((16c^4 - 1)\tilde{x}^2 + 2c^2\tilde{x} - 4c^4) > 0$. Therefore, $Q(\tilde{p}) = 0 \iff P(\tilde{p}) = 0$, and they have the same coefficient signs. The polynomial $P(\tilde{p})$ has two roots $\overline{\tilde{p}}$ and $\underline{\tilde{p}}$ given by

$$\begin{aligned} \overline{\tilde{p}}(\tilde{x}, c) &= \frac{6c^2\tilde{x} - 4c^4 - 16\tilde{x}c^6 - \tilde{x}^2 + 8\tilde{x}^2c^4 + \sqrt{d(\tilde{x}, c)}}{2(4c^4\tilde{x} + 2c^2\tilde{x}^2)}, \\ \underline{\tilde{p}}(\tilde{x}, c) &= \frac{6c^2\tilde{x} - 4c^4 - 16\tilde{x}c^6 - \tilde{x}^2 + 8\tilde{x}^2c^4 - \sqrt{d(\tilde{x}, c)}}{2(4c^4\tilde{x} + 2c^2\tilde{x}^2)}, \end{aligned}$$

and $d(\tilde{x}, c)$ is given by (26). One can show, using the first derivative with respect to $\tilde{x}$, where $\tilde{x} \in [\frac{-1}{2c^2}, 0),\ c \geq 1$, and finding the minimum that the two factors of d are negative for $\tilde{x} \in [\frac{-1}{2c^2}, 0),\ c \geq 1$, i.e.

$$\begin{aligned} (16c^4 - 4c^2 - 1)\tilde{x}^2 + (6c^2 - 8c^4)\tilde{x} - 4c^4 < 0, \\ (16c^4 + 4c^2 - 1)\tilde{x}^2 + (6c^2 + 8c^4)\tilde{x} - 4c^4 < 0, \end{aligned} \tag{28}$$

and hence, $d > 0$ for $\tilde{x} \in [\frac{-1}{2c^2}, 0)$. Now, we want to show that $\overline{\tilde{p}} < 1 - 2c^2$, and hence $\overline{\tilde{p}}$ can be discarded, and $\underline{\tilde{p}} > 1 - 2c^2$. For $\overline{\tilde{p}}$, we have

$$\begin{aligned} \overline{\tilde{p}} < 1 - 2c^2 \quad &\iff \quad (8c^4 - 1)\tilde{x}^2 + (6c^2 - 16c^6)\tilde{x} - 4c^4 + \sqrt{d} \\ &\qquad > 4c^2\tilde{x}(\tilde{x} + 2c^2)(1 - 2c^2) \text{ (since } \tilde{x}(\tilde{x} + 2c^2) < 0) \\ &\iff \quad \sqrt{d} > -((16c^4 - 4c^2 - 1)\tilde{x}^2 + (6c^2 - 8c^4)\tilde{x} - 4c^4), \end{aligned}$$

and now both sides are positive by (28), and squaring and simplifying leads to

$$\iff 8c^2\tilde{x}(\tilde{x} + 2c^2)\left((16c^4 - 4c^2 - 1)\tilde{x}^2 + (6c^2 - 8c^4)\tilde{x} - 4c^4\right) > 0.$$

This last inequality holds since $(16c^4 - 4c^2 - 1)\tilde{x}^2 + (6c^2 - 8c^4)\tilde{x} - 4c^4 < 0$, and $8c^2\tilde{x}(\tilde{x} + 2c^2) < 0$ for $\tilde{x} \in [\frac{-1}{2c^2}, 0)$. Hence, $\overline{\tilde{p}} < 1 - 2c^2$.

For $\underline{\tilde{p}}$, we have, simplifying as before

$$\underline{\tilde{p}} > 1 - 2c^2 \quad \iff \quad \sqrt{d} > (16c^4 - 4c^2 - 1)\tilde{x}^2 + (6c^2 - 8c^4)\tilde{x} - 4c^4.$$

The last inequality holds since the right hand side is negative, and the left hand side is positive for $\tilde{x} \in [\frac{-1}{2c^2}, 0)$. Hence, $\underline{\tilde{p}} > 1 - 2c^2$. The coefficient of $\tilde{p}^2$ in the polynomial $P(\tilde{p})$ is positive for $\tilde{x} \in [-\frac{1}{2c^2}, 0)$ with $c \geq 1$, and hence, the larger of the two roots $\underline{\tilde{p}}$ and $\overline{\tilde{p}}$ is a minimum. Therefore, for $1 - 2c^2 < \tilde{p} < \underline{\tilde{p}}$, increasing $\tilde{p}$ decreases R_0, whereas for $\tilde{p} > \underline{\tilde{p}}$, the opposite holds, i.e. increasing $\tilde{p}$ increases R_0. ■

Theorem 2.11 (Optimized constant transmission conditions). *The best performance of the optimized waveform relaxation algorithm (6) with constant transmission conditions and the simplifying assumption (11) is obtained for* $\alpha = \alpha^*$*, where*

$$\alpha^* = 2c^2 - 1 + \tilde{p}^*, \tag{29}$$

and $\tilde{p}^*$*, the solution of the min-max problem (17), is given by*

$$\tilde{p}^* = \frac{-1 + \sqrt{1 + 16c^8 - 12c^4}}{2c^2}, \tag{30}$$

and $c = \sqrt{\frac{-b}{2a}} \geq 1$*. Furthermore,* $\alpha^* > \frac{a|b|}{b^2 - a^2} - 1 = \underline{\alpha}$*.*

Proof By Lemma 2.10, the optimal $\tilde{p}^*$ must lie in the interval $[\frac{-1}{2c^2}, \infty)$, since with $\tilde{p}$ outside this interval, R_0 can be uniformly decreased for all $\frac{-1}{2c^2} \leq \tilde{x} < 0$ by moving $\tilde{p}$ towards this interval. The left endpoint of this interval is $\underline{\tilde{p}}(\tilde{x} = \frac{-1}{2c^2})$, and the right endpoint is $\underline{\tilde{p}}(\tilde{x} = 0^-) := \lim_{\tilde{x} \to 0^-} (\underline{\tilde{p}}(\tilde{x}))$. Now, by Lemma 2.9, the maximum of the min-max problem can only be attained on the boundaries, at $\tilde{x} = \frac{-1}{2c^2}$ and at $\tilde{x} = 0^-$, since R_0 has no interior maxima. By the notation $\tilde{x} = 0^-$ we mean that $\tilde{x}$ approaches 0 from the left, since we have $\tilde{x} \in [\frac{-1}{2c^2}, 0)$, open from the right. Now, for $\tilde{p} = \underline{\tilde{p}}(\tilde{x} = \frac{-1}{2c^2}) = \frac{-1}{2c^2}$, we have $R_0(\frac{-1}{2c^2}, c, \frac{-1}{2c^2}) = 0$, and so increasing $\tilde{p}$ increases $R_0(\frac{-1}{2c^2}, c, \tilde{p})$ monotonically, by Lemma 2.10. On the other hand, for $\tilde{p} = \frac{-1}{2c^2}$, we have $R_0(0^-, c, \frac{-1}{2c^2}) = \frac{4c^2}{1 + 16c^8 - 8c^4} > 0$, $c \geq 1$, and increasing $\tilde{p}$ decreases $R_0(0^-, c, \tilde{p}) = \frac{1}{(2c^2 + \tilde{p})^2}$ to $\lim_{\tilde{p} \to \infty} (\frac{1}{(2c^2 + \tilde{p})^2}) = 0$. Therefore, by increasing $\tilde{p}$ we reach $R_0(\frac{-1}{2c^2}, c, \tilde{p}) = R_0(0^-, c, \tilde{p})$. Solving the equation for $\tilde{p}$ gives the solution in (30), and three other solutions, $\tilde{p} = 1 - 2c^2$, $-1 - 2c^2$, $\frac{-1 - \sqrt{1 + 16c^8 - 12c^4}}{2c^2}$. Those three solutions can be discarded, since $\tilde{p} > 1 - 2c^2$. Therefore, we have $\alpha^* = 2c^2 - 1 + \frac{p^*}{a} = 2c^2 - 1 + \tilde{p}^*$, $c = \sqrt{\frac{-b}{2a}} \geq 1$, where $\tilde{p}^*$ is given in (30), and $\alpha^* > \frac{a|b|}{b^2 - a^2} - 1 := \underline{\alpha}$. ■

In Figure 3, we show the modulus of the convergence factor for the optimized WR algorithm with the optimized constant approximation, $\rho_{opt0}(\omega, \alpha^*)$. We also show the modulus of the convergence factor using a Taylor approximation, $\rho_{opt0}(\omega, \alpha_T)$, obtained by a zeroth order expansion of the optimal choice (8), which gives using the simplifying assumption $\alpha_T = \frac{a}{b} - \frac{b}{a} - 1$. We finally show the modulus of the classical convergence factor ρ_{cla}. All curves are shown for $c = 1$ from the numerical experiment in Section 4. One can see the remarkable improvement in magnitude and uniformity for the convergence factor with the optimized constant approximation over the classical one and the one with the Taylor approximation.

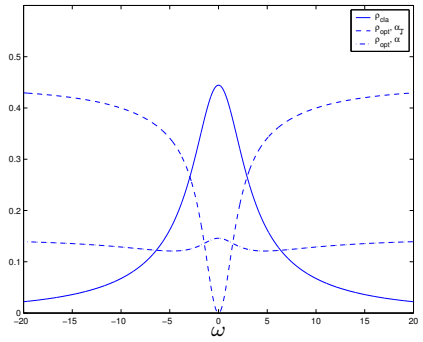

Figure 3: Classical convergence factor $|\rho_{cla}(\omega)|$, versus $|\rho_{opt0}(\omega, \alpha^*)|$ and $|\rho_{opt0}(\omega, \alpha_T)|$.

3. First order transmission conditions

We now approximate the symbols α and β from (8) corresponding to the optimal transmission conditions by a first order polynomial in s,

$$\alpha = \alpha_0 + \alpha_1 s, \quad \beta = \beta_0 + \beta_1 s, \tag{31}$$

where we have four free parameters α_0, α_1, β_0 and β_1 that we can choose to obtain a new optimized waveform relaxation algorithm. This will lead to a time derivative in the time domain in the transmission condition, because of the multiplication by s, and we assume that $\alpha_1, \beta_1 \neq 0$, since otherwise we will get again the constant approximation case.

The simplest first order approximation of the optimal α and β is the low frequency approximation by using a Taylor expansion about $s = 0$ of the optimal choice (8), which is given by

$$\alpha_{0T} = \frac{a_3 c_3}{a_2 b_4} - \frac{b_3}{a_2} - 1, \quad \alpha_{1T} = \frac{1}{a_2} + \frac{a_3 c_3}{a_2 b_4^2},$$
$$\beta_{0T} = -\frac{a_1 c_1}{b_1 c_2} + \frac{b_2}{c_2} + 1, \quad \beta_{1T} = -\frac{1}{c_2} - \frac{a_1 c_1}{b_1^2 c_2}.$$

In Figure 4 on the left, we compare the classical convergence factor with the optimized convergence factor with constant approximation and the first order Taylor approximation for the circuit from the numerical experiment in Section 4. We observe that the first order transmission conditions substantially improve the convergence further, even using only a Taylor approximation.

In order to implement first order transmission conditions, we reformulate now slightly the oWR algorithm: the Laplace transform of the transmission conditions in (5), using the first order expansion (31) for α and β, implies for the transmission conditions

$$\begin{aligned}(\hat{u}_3^{k+1} - \hat{u}_2^{k+1}) + \alpha_0 \hat{u}_3^{k+1} + \alpha_1 s \hat{u}_3^{k+1} &= (\hat{w}_1^k - \hat{w}_0^k) + \alpha_0 \hat{w}_1^k + \alpha_1 s \hat{w}_1^k, \\ (\hat{w}_1^{k+1} - \hat{w}_0^{k+1}) + \beta_0 \hat{w}_0^{k+1} + \beta_1 s \hat{w}_0^{k+1} &= (\hat{u}_3^k - \hat{u}_2^k) + \beta_0 \hat{u}_2^k + \beta_1 s \hat{u}_2^k.\end{aligned} \tag{32}$$

Since a multiplication by s in the frequency domain corresponds to a time derivative, by substituting

$$\dot{w}_1^k = b_3 w_1^k + c_3 w_2^k + a_2 w_0^k + f_3, \quad \dot{u}_2^k = a_1 u_1^k + b_2 u_2^k + c_2 u_3^k + f_2,$$

from the circuit equations into (32), assuming $\alpha_1, \beta_1 \neq 0$, we obtain

$$\begin{aligned} \dot{u}_3^{k+1} &= \tfrac{1}{\alpha_1} u_2^{k+1} - \tfrac{(1+\alpha_0)}{\alpha_1} u_3^{k+1} + \tfrac{(1+\alpha_0+\alpha_1 b_3)}{\alpha_1} w_1^k + \tfrac{(\alpha_1 a_2 - 1)}{\alpha_1} w_0^k + c_3 w_2^k + f_3, \\ \dot{w}_0^{k+1} &= -\tfrac{1}{\beta_1} w_1^{k+1} + \tfrac{(1-\beta_0)}{\beta_1} w_0^{k+1} - \tfrac{(1-\beta_0-\beta_1 b_2)}{\beta_1} u_2^k + \tfrac{(\beta_1 c_2 + 1)}{\beta_1} u_3^k + a_1 u_1^k + f_2. \end{aligned} \tag{33}$$

These ordinary differential equations found from the transmission conditions imply the oWR algorithm

$$\begin{aligned} \begin{pmatrix} \dot{u}_1^{k+1} \\ \dot{u}_2^{k+1} \\ \dot{u}_3^{k+1} \end{pmatrix} &= \begin{bmatrix} b_1 & c_1 & \\ a_1 & b_2 & c_2 \\ & \frac{1}{\alpha_1} & \frac{-(\alpha_0+1)}{\alpha_1} \end{bmatrix} \begin{pmatrix} u_1^{k+1} \\ u_2^{k+1} \\ u_3^{k+1} \end{pmatrix} + \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \\ &\quad + \begin{pmatrix} 0 \\ 0 \\ \frac{(\alpha_1 a_2 - 1)}{\alpha_1} w_0^k + \frac{(1+\alpha_0+\alpha_1 b_3)}{\alpha_1} w_1^k + c_3 w_2^k \end{pmatrix}, \end{aligned} \tag{34}$$

and

$$\begin{aligned} \begin{pmatrix} \dot{w}_0^{k+1} \\ \dot{w}_1^{k+1} \\ \dot{w}_2^{k+1} \end{pmatrix} &= \begin{bmatrix} \frac{1-\beta_0}{\beta_1} & \frac{-1}{\beta_1} & \\ a_2 & b_3 & c_3 \\ & a_3 & b_4 \end{bmatrix} \begin{pmatrix} w_0^{k+1} \\ w_1^{k+1} \\ w_2^{k+1} \end{pmatrix} + \begin{pmatrix} f_2 \\ f_3 \\ f_4 \end{pmatrix} \\ &\quad + \begin{pmatrix} a_1 u_1^k - \frac{(1-\beta_0-\beta_1 b_2)}{\beta_1} u_2^k + \frac{(\beta_1 c_2 + 1)}{\beta_1} u_3^k \\ 0 \\ 0 \end{pmatrix}, \end{aligned} \tag{35}$$

with the initial conditions $\boldsymbol{u}^{k+1}(0) = (v_1^0, v_2^0, v_3^0)^T$ and $\boldsymbol{w}^{k+1}(0) = (v_2^0, v_3^0, v_4^0)^T$, where now the transmission conditions are already implemented in the algorithm. To start the WR iteration, some initial waveforms $\boldsymbol{u}^0(t)$ and $\boldsymbol{w}^0(t)$ are used.

Using the simplifying assumptions (11) from Section 2, and the associated choice $\beta = -\alpha$, we obtain for this variant of the oWR the simplified convergence factor

$$\rho_{opt1}(s, a, b, \alpha_0, \alpha_1) = \left(\frac{(\alpha_0 + \alpha_1 s) + 1 - \lambda}{((\alpha_0 + \alpha_1 s) + 1)\lambda - 1} \right)^2, \tag{36}$$

with λ from (13). Similar to the optimized WR algorithm with constant transmission conditions in Section 2, we use an optimization process to get the best performance of the new WR algorithm. We again want $|\rho_{opt1}| \ll 1$.

Lemma 3.1. *If the circuit parameters satisfy the inequalities*

$$a > 0, \quad b < 0, \quad |b| \geq 2a, \quad \alpha_0 \geq 0, \quad \alpha_1 > 0, \tag{37}$$

then the convergence factor ρ_{opt1} in (36) is an analytic function in the right half of the complex plane, $s = \eta + i\omega$, $\eta > 0$.

Proof We need to show that the denominator does not have zeros in the right half of the complex plane. We show this by contradiction: Assume there is a zero, $\lambda(1+\alpha_0+\alpha_1 s)-1=0$, then we have $\lambda=\frac{1}{1+\alpha_0+\alpha_1 s}$, which implies $|\lambda|=\frac{1}{(1+\alpha_0+\alpha_1\eta)^2+\omega^2\alpha_1^2}<1$, with the condition (37) on α_0 and α_1. On the other hand, we have by Lemma 2.1 that the modulus $|\lambda|$ is bigger than one in the right half of the complex plane, and thus we have a contradiction. Hence, poles are excluded and the denominator has no zeros in the right half of the complex plane. ■

Since ρ_{opt1} is analytic, we can apply the maximum principle and therefore, the maximum of $|\rho_{opt1}|$ is attained on the boundary. Now, since for $s=re^{i\theta}$, $-\frac{\pi}{2}<\theta<\frac{\pi}{2}$, the limit of ρ_{opt1} is zero as r goes to infinity, i.e. the same limit in all directions, the maximum of $|\rho_{opt1}|$ is attained on the boundary $\eta=\text{const}$. Looking for an L^2 estimate in time as in Section 2, the modulus of ρ_{opt1} for $s=i\omega$ depends on ω^2 only, and it suffices to optimize for non-negative frequencies, $\omega\geq 0$. Therefore, we need to solve the min-max problem

$$\min_{\alpha_0\geq 0,\alpha_1>0}\left(\max_{\omega\geq 0}|\rho_{opt1}(i\omega,a,b,\alpha_0,\alpha_1)|\right). \tag{38}$$

The optimal value of α is given by $\alpha=\lambda-1$, and a first order approximation is

$$\alpha=\alpha_0+\alpha_1 s=\frac{-b}{a}-1+\frac{p}{a}+\frac{q}{a}s,$$

where $p,\ q$ are new parameters. Considering this first order approximation, and using in addition the change of variables (14) for the convergence factor, the modulus of the convergence factor ρ_{opt1} in (36), after factorizing a^7 from the denominator and numerator to eliminate one parameter, is given by

$$|\rho_{opt1}(x,a,b,p,q)|:=\frac{Q_1(x,a,b,p,q)}{Q_2(x,a,b,p,q)},$$

where

$$\begin{aligned}
Q_1 :=& \left(2\tfrac{b}{a}\tfrac{p}{a}-2q\left(\tfrac{b}{a}\right)^2+2\left(\tfrac{b}{a}\right)^2-1\right)\left(\tfrac{x}{a}\right)^3\\
&+\left(-\tfrac{b}{a}\left(\tfrac{p}{a}\right)^2-2q\left(\tfrac{b}{a}\right)^3+2q\tfrac{b}{a}+\left(\tfrac{b}{a}\right)^3+\left(\tfrac{b}{a}\right)^3q^2-2\tfrac{b}{a}\right)\left(\tfrac{x}{a}\right)^2\\
&+\left(2q\left(\tfrac{b}{a}\right)^2-\left(\tfrac{b}{a}\right)^2q^2-\left(\tfrac{b}{a}\right)^2\right)\tfrac{x}{a},\\
Q_2 :=& \left(\left(\tfrac{b}{a}\right)^2q^2+4\tfrac{b}{a}\tfrac{p}{a}+4\left(\tfrac{b}{a}\right)^2\left(\tfrac{p}{a}\right)^2+2q\left(\tfrac{b}{a}\right)^2-4\left(\tfrac{b}{a}\right)^4q^2-8\left(\tfrac{b}{a}\right)^3\tfrac{p}{a}\right.\\
&\left.-\left(\tfrac{p}{a}\right)^2-3\left(\tfrac{b}{a}\right)^2+4\left(\tfrac{b}{a}\right)^4\right)\left(\tfrac{x}{a}\right)^3\\
&+\left(-\tfrac{b}{a}q^2-2q\tfrac{b}{a}+2\left(\tfrac{b}{a}\right)^2\tfrac{p}{a}-\tfrac{b}{a}-2\left(\tfrac{p}{a}\right)^2\tfrac{b}{a}+2q\left(\tfrac{b}{a}\right)^3+6\left(\tfrac{b}{a}\right)^3q^2\right)\left(\tfrac{x}{a}\right)^2\\
&+\left(-\left(\tfrac{b}{a}\right)^4-2q\left(\tfrac{b}{a}\right)^2+\left(\tfrac{b}{a}\right)^4q^2-2\left(\tfrac{b}{a}\right)^2q^2-\left(\tfrac{p}{a}\right)^2\left(\tfrac{b}{a}\right)^2+2\left(\tfrac{b}{a}\right)^3\tfrac{p}{a}\right)\tfrac{x}{a}\\
&-\left(\tfrac{b}{a}\right)^3q^2.
\end{aligned}$$

Letting $\tilde{p} = \frac{p}{a}$, $\frac{b}{a} = -2c^2$, where $c \geq 1$, and $\tilde{x} = \frac{x}{a}$, as for the constant approximation in Section 2, the modulus of the convergence factor ρ_{opt1} in (36) becomes

$$R_1(\tilde{x}, c, \tilde{p}, q) := \frac{P_1(\tilde{x}, c, \tilde{p}, q)}{P_2(\tilde{x}, c, \tilde{p}, q)}, \tag{39}$$

where

$$\begin{aligned}
P_1 := -(((&8qc^4 - 8c^4 + 1 + 4c^2\tilde{p})\tilde{x}^2 \\
&+(-16qc^6 - 2c^2\tilde{p}^2 - 4c^2 + 4qc^2 + 8c^6 + 8c^6q^2)\tilde{x} \\
&+4c^4 - 8qc^4 + 4c^4q^2)\tilde{x}), \\
P_2 := (&4c^4q^2 - 8c^2\tilde{p} + 16c^4\tilde{p}^2 + 8qc^4 - 64c^8q^2 + 64c^6\tilde{p} - \tilde{p}^2 - 12c^4 + 64c^8)\tilde{x}^3 \\
&+(2c^2q^2 + 4qc^2 + 8c^4\tilde{p} + 2c^2 + 4c^2\tilde{p}^2 - 16qc^6 - 48c^6q^2)\tilde{x}^2 \\
&+(-16c^8 - 8qc^4 + 16c^8q^2 - 8c^4q^2 - 4c^4\tilde{p}^2 - 16c^6\tilde{p})\tilde{x} + 8c^6q^2.
\end{aligned}$$

The optimized parameters are given by $\alpha_0 = \frac{-b}{a} - 1 + \frac{p}{a} = 2c^2 - 1 + \tilde{p}$, and $\alpha_1 = \frac{q}{a}$, and since for analyticity in the right half of the complex plane we need $\alpha_0 \geq 0$, and $\alpha_1 > 0$, we require $\tilde{p} \geq 1 - 2c^2$, and $q > 0$. The min-max problem (38) then becomes

$$\min_{\tilde{p} \geq 1-2c^2, q>0} \left(\max_{\frac{-1}{2c^2} \leq \tilde{x} < 0} R_1(\tilde{x}, c, p, q) \right) = \max_{\frac{-1}{2c^2} \leq \tilde{x} < 0} R_1(\tilde{x}, c, p^*, q^*), \ c \geq 1. \tag{40}$$

Since it is hard to solve the optimization problem (40) we use asymptotics, and since for RC type circuits $|b| = 2a$, which corresponds to c going to 1, often holds, we take $c = \sqrt{1+\epsilon}$, and for ϵ small we have the following result.

Theorem 3.2 (Optimized first order transmission conditions). *If in the optimized WR algorithm with first order transmission conditions (34), (35) the free parameters are chosen to be* $\alpha_0 = \alpha_0^* = 2c^2 - 1 + \tilde{p}^*$, *and* $\alpha_1 = \alpha_1^* = \frac{q^*}{a}$, *where* $c = \sqrt{\frac{-b}{2a}} = \sqrt{1+\epsilon} \geq 1$ *and* a, b *are the entries of the matrices in (34), (35), and* $\tilde{p}^*$ *and* q^* *are defined by the system of equations*

$$R_1(\tilde{x}_0, c, \tilde{p}^*, q^*) = R_1(\overline{\tilde{x}}, c, \tilde{p}^*, q^*), \qquad \frac{\partial}{\partial q} R_1(\overline{\tilde{x}}, c, \tilde{p}^*, q^*) = 0, \tag{41}$$

where $\tilde{x}_0 = \frac{-1}{2c^2}$, $R_1(\tilde{x}, c, \tilde{p}, q)$ *is given in (39), and* $\overline{\tilde{x}}$ *is given by the root of the polynomial* $P(\tilde{x})$ *giving the maximum of* R_1, *then for* ϵ *small,* $R_1(\tilde{x}, c, \tilde{p}^*, q^*) \leq R_1(\tilde{x}_0, c, \tilde{p}^*, q^*) := \bar{R}_{O1}$ *for all* $\tilde{x} \in [\tilde{x}_0, 0)$. *Moreover, we have the asymptotic result*

$$\tilde{p}^* \sim -0.4655, \ q^* \sim 1.1378, \ \bar{R}_{O1} \sim 0.0007.$$

Proof A partial derivative of R_1 with respect to $\tilde{x}$ shows that the roots of the polynomial $P(\tilde{x})$ determine the extrema of R_1. First, to see that there is indeed a solution as stated in (41) for ϵ small, we substitute $c = \sqrt{1+\epsilon}$, and we use the ansatz $\tilde{p} = C_p\epsilon^{\gamma_1}$, $q = C_q\epsilon^{\gamma_2}$, and $\overline{\tilde{x}} = C_1\epsilon^{\delta}$, and determine the leading asymptotic terms as ϵ goes to zero of the root of the polynomial $P(\tilde{x})$, and the

equations (41). After extensive calculations, which can be found in [1], we get by equating the exponents in these three equations $\gamma_1 = \gamma_2 = \delta = 0$, which implies the same equations as for the case $c = 1$. Since the constants need to match as well, we obtain $C_p = -0.4655$, $C_q = 1.1378$, and $C_1 = -0.2617$, by solving the resulting equations. Since $c = \sqrt{\frac{-b}{2a}} = \sqrt{1+\epsilon}$, we have $\epsilon = \frac{-b}{2a} - 1$, and using these results we get

$$\begin{aligned} \tilde{p}^* &:= C_p(\tfrac{-b}{2q} - 1)^{\gamma_1} = C_p = -0.4655, \\ q^* &:= C_q(\tfrac{-b}{2a} - 1)^{\gamma_2} = C_q = 1.1378. \end{aligned} \tag{42}$$

Now, to see that there is indeed only one interior maximum, which we denote by $\overline{\tilde{x}}$, we take $c = \sqrt{1+\epsilon}$, and we substitute $\tilde{p}^*$, q^* from (42) into $P(\tilde{x})$. The leading terms of the polynomial $P(\tilde{x})$ as ϵ goes to zero are

$$\begin{aligned} P(\tilde{x}) = 7.1756762\tilde{x}^4 - 2.0577882\tilde{x}^3 - 10.88760728\tilde{x}^2 \\ -5.58470978\tilde{x} - 0.78638698 + O(\epsilon), \end{aligned}$$

which is a polynomial of degree 4 in $\tilde{x}$ plus higher order terms. As ϵ goes to 0, finding the roots of this 4^{th} degree polynomial implies the four roots -0.5737, -0.4610, -0.2617, and 1.5832. Only two roots lie in the interval $[-\frac{1}{2}, 0)$, which are one maximum given by $\overline{\tilde{x}} = -0.2617$ and one minimum. Therefore, as ϵ goes to zero, $R_1(\tilde{x}, \tilde{p}^*, q^*)$ has only one interior maximum at $\overline{\tilde{x}}$, where $\tilde{p}^*$ and q^* are given in (42). Since R_1 has only one interior maximum for $\tilde{x} \in [\frac{-1}{2c^2}, 0)$, $c = \sqrt{1+\epsilon}$, and ϵ small, and no other interior maximum as we have shown above, and since in addition, $R_1 \longrightarrow 0$ as $\tilde{x} \longrightarrow 0$, the maximum of R_1 can be attained either on the boundary at $\tilde{x} = \tilde{x}_0$ or at the maximum $\overline{\tilde{x}}$. Balancing the value of R_1 at the two locations as stated in (41) guarantees then that R_1 is uniformly bounded by R_1 at $\tilde{x}_0$ for ϵ small. Now, expanding $\bar{R}_{O1}$ for ϵ small, we get

$$\bar{R}_{O1} \sim \frac{1 + 4C_p + 4C_p^2}{24C_p + 9C_p^2 + 16},$$

and substituting from (42) we obtain the asymptotic result

$$\bar{R}_{O1} \sim 0.0007.$$

Finally, α_0^* and α_1^* are given by

$$\alpha_0^* = 2c^2 - 1 + \frac{p^*}{a} = 2c^2 - 1 + \tilde{p}^*, \ \alpha_1^* = \frac{q^*}{a}, \ c = \sqrt{\frac{-b}{2a}} \geq 1,$$

and $\tilde{p}^*$, q^* are given in (42). ■

In Fig. 4 on the left, we observe the better convergence we get by using the first order approximation over the classical convergence and the convergence using the optimized constant approximation. We show in Fig. 4 on the right the result of the optimization with respect to α_0 and α_1 using the circuit elements from the numerical experiment in Section 4. The solution of the min-max problem

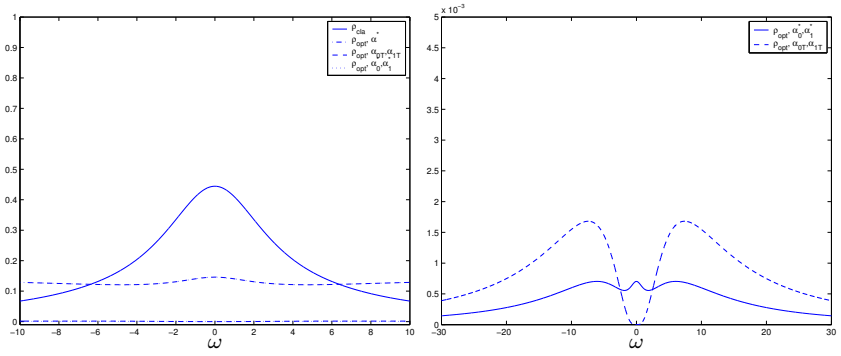

Figure 4: On the left: classical convergence factor $|\rho_{cla}(\omega)|$ versus $|\rho_{opt0}(\omega, \alpha^*)|$, $|\rho_{opt1}(\omega, \alpha_{0T}, \alpha_{1T})|$, and $|\rho_{opt1}(\omega, \alpha_0^*, \alpha_1^*)|$. On the right: convergence factor with the optimized first order approximation $|\rho_{opt1}(\omega, \alpha_0^*, \alpha_1^*)|$ versus $|\rho_{opt1}(\omega, \alpha_{0T}, \alpha_{1T})|$.

occurs when the convergence factor at $\omega = 0$ and at $\omega = \overline{\omega}$ are balanced, where $\overline{\omega} > 0$ is the interior maximum of the modulus of the convergence factor. We also show in Fig. 4 on the right the better convergence we obtain using the optimized values α_0^* and α_1^* over the one using the low frequency first order approximation α_{0T} and α_{1T}.

4. Numerical experiments

The results so far in the reduction of the convergence factor ρ in this work clearly indicate that the oWR does converge faster. To confirm this we use a numerical example to show the improvements in the convergence of the optimized WR algorithms in comparison to a classical WR algorithm. We use the typical values of the RC circuit parameters

$$R_s = R_1 = R_2 = R_3 = \frac{1}{2} \text{ kOhms}, \; C_1 = C_2 = C_3 = C_4 = \frac{63}{100} \text{ pF}.$$

We use a simple backward Euler time integration method to compute the transient solution for $t \in [0, 10]$ ns and a time step of $\Delta t = 1/10$. We start with random initial waveforms and use an input step function with an amplitude of $I_s = 1$ mA and a rise time of 1 time unit. In Figure 5 we show the error as a function of the iterations. One can see the remarkable improvement of the optimized WR algorithm over the classical one. Furthermore, the optimized WR algorithm with first order transmission conditions converges faster than the one with constant transmission conditions.

On the left hand side of Figure 5, we show the example where the simplifying assumptions (11) we used to compute the optimized constant and first order approximations with $\beta = -\alpha$ are satisfied. This leads to the optimized value $\alpha^* = 1.618$ as well as the Taylor approximation $\alpha_T = 0.5$ in the optimized WR algorithm with constant transmission conditions, and for the first order

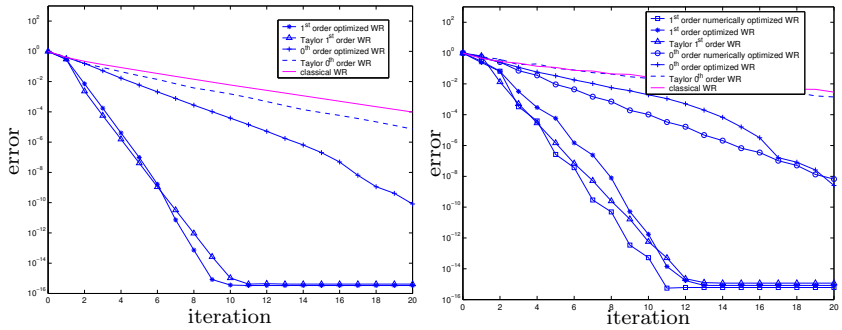

Figure 5: Convergence behavior of classical versus optimized WR algorithms for a model RC circuit.

transmission conditions we obtained the optimized values $\alpha_0^* = 0.5345$, $\alpha_1^* = 0.3585$, and the Taylor approximations $\alpha_{0T} = 0.5$, $\alpha_{1T} = 0.3937$.

On the right hand side of Figure 5, we show an example where the simplifying assumptions (11) are not satisfied, $b_4 = \frac{b_1}{2}$, and we find the Taylor approximation $\alpha_T = 0$ and $\beta_T = -0.5$, and the numerically optimized constant approximation $\alpha^* = 2.3002$ and $\beta^* = -0.6953$, which we use in the constant optimized WR algorithm. The optimized constant approximation $\alpha^* = 1.618$ with $\beta^* = -\alpha^*$ computed using the simplifying assumptions is also shown in the constant optimized WR algorithm for comparison purposes. For the first order Taylor approximation, we obtained $\alpha_{0T} = 0$, $\alpha_{1T} = 0.63$, $\beta_{0T} = -0.5$, $\beta_{1T} = -0.3937$, and the numerically optimized first order approximation $\alpha_0^* = 0.3530$, $\alpha_1^* = 0.4092$, $\beta_0^* = -0.5027$, $\beta_1^* = -0.3875$. The optimized first order approximation used here for this case is again $\alpha_0^* = 0.5345$, $\alpha_1^* = 0.3585$ with $\beta^* = -\alpha^*$.

5. Conclusion

We have proved in this paper that the optimal choice of the constant transmission condition conjectured in [2] is indeed true for a small RC circuit. We have also proposed a first order transmission condition which includes time derivatives, and leads to even better convergence behavior of the oWR algorithm. We should note that other circuit topologies like RLC transmission lines have successfully been solved with the oWR approach, see for example [8].

References

[1] M. Al-Khaleel, Optimized waveform relaxation methods for circuit simulations, Ph.D. thesis, McGill University, Montreal, QC, Canada (2007).

[2] M. Al-Khaleel, M. J. Gander, A. Ruehli, Optimization of transmission conditions in waveform relaxation techniques for RC circuits, SIAM J. Num. Anal.Submitted.

[3] D. Bennequin, M. J. Gander, L. Halpern, A homographic best approximation problem with application to optimized Schwarz waveform relaxation, Math. of Comp. 78 (265) (2009) 185–232.

[4] M. Bjørhus, A note on the convergence of discretized dynamic iteration, BIT 35 (1995) 291–296.

[5] M. Bjorhus, A. M. Stuart, Waveform relaxation as a dynamical system, Mathematics of Computation 66 (219) (1997) 1101–1117.

[6] V. B. Dmitriev-Zdorov, Generalized coupling as a way to improve the convergence in relaxation-based solvers, in: Proc. of European Design Autom. Conf., EURO-DAC'96, 1996, pp. 15–20.

[7] V. B. Dmitriev-Zdorov, B. Klaassen, An improved relaxation approach for mixed system analysis with several simulation tools, in: EURO-DAC'95, 1995, pp. 274–279.

[8] M. J. Gander, M. Al-Khaleel, A. E. Ruehli, Waveform relaxation technique for longitudinal partitioning of transmission lines, in: EPEP, Scottsdale, AZ, 2006, pp. 207–210.

[9] M. J. Gander, L. Halpern, Methodes de relaxation d'ondes pour l'equation de la chaleur en dimension 1, C. R. Acad. Sci Paris, Série I 336 (6) (2003) 519–524.

[10] M. J. Gander, L. Halpern, Optimized Schwarz waveform relaxation methods for advection reaction diffusion problems, SIAM J. Numer. Anal. 45 (2) (2007) 666–697.

[11] M. J. Gander, L. Halpern, F. Nataf, Optimal convergence for overlapping and non-overlapping Schwarz waveform relaxation, in: C.-H. Lai, P. Bjørstad, M. Cross, O. Widlund (eds.), Eleventh international Conference of Domain Decomposition Methods, ddm.org, 1999.

[12] M. J. Gander, A. Ruehli, Solution of large transmission line type circuits using a new optimized waveform relaxation partitioning, in: IEEE Transactions on Electromagnetic Compatibility, vol. 2, 2003, pp. 636–641.

[13] M. J. Gander, A. Ruehli, Optimized waveform relaxation methods for RC type circuits, IEEE Transactions on Circuits and Systems, I 51 (4) (2004) 755–767.

[14] M. J. Gander, A. E. Ruehli, Optimized waveform relaxation methods for RC type circuits, IBM Research Report RC 22207, IBM T. J. Watson Research Center, Yorktown Heights (Oct. 2001).

[15] M. J. Gander, A. E. Ruehli, Optimized waveform relaxation solution of electromagnetic and circuit problems, in: Digest of Electr. Perf. Electronic Packaging, vol. 19, Austin, TX, 2010, pp. 65–68.

[16] M. J. Gander, A. M. Stuart, Space-time continuous analysis of waveform relaxation for the heat equation, SIAM Journal on Scientific Computing 19 (6) (1998) 2014–2031.

[17] M. J. Gander, H. Zhao, Overlapping Schwarz waveform relaxation for the heat equation in n-dimensions, BIT 42 (4) (2002) 779–795.

[18] E. Giladi, H. B. Keller, Space time domain decomposition for parabolic problems, Numerische Mathematik 93 (2) (2002) 279–313.

[19] C. Ho, A. Ruehli, P. Brennan, The modified nodal approach to network analysis, IEEE Transactions on Circuits and Systems 22 (6) (1975) 504–509.

[20] R. Jeltsch, B. Pohl, Waveform relaxation with overlapping splittings, SIAM J. Sci. Comput. 16 (1) (1995) 40–49.

[21] E. Lelarasmee, A. E. Ruehli, A. L. Sangiovanni-Vincentelli, The waveform relaxation method for time-domain analysis of large-scale integrated circuits., IEEE Trans. on CAD of Integrated Circuits and Systems CAD-1 (3) (1982) 131–145.

[22] U. Miekkala, O. Nevanlinna, Convergence of dynamic iteration methods for initial value problems, SIAM J. Sci. Stat. Comput. 8 (1987) 459–482.

[23] O. Nevanlinna, Remarks on Picard-Lindelöf iterations part I, BIT 29 (1989) 328–346.

[24] O. Nevanlinna, Remarks on Picard-Lindelöf iterations part II, BIT 29 (1989) 535–562.

[25] A. E. Ruehli, T. A. Johnson, Circuit analysis computing by waveform relaxation, Encyclopedia of Electrical and Electronics Engineering, Wiley New York 3.