UNIVERSITY OF CAMBRIDGE

Numerical Analysis Reports

Waveform Relaxation Method with Toeplitz Operator Splitting

Sigitas Keras

DAMTP 1995/NA4
August 1995

Department of Applied Mathematics and Theoretical Physics
Silver Street
Cambridge CB3 9EW
England

Waveform Relaxation Method with Toeplitz Operator Splitting

Sigitas Keras *

Abstract

In this paper we consider a waveform relaxation method of the form

$$\frac{du^{n+1}}{dt} + Pu^{n+1} = Qu^n + f, \qquad A = P - Q$$
$$u^{n+1}(0) = u_0,$$

where P is Toeplitz or block Toeplitz matrix for a numerical solution of the equation

$$\frac{du}{dt} + Au = f$$
$$u(0) = u_0$$

We show that under suitable conditions this method converges and apply this result to linear parabolic equations with variable coefficients.

1 Introduction

In this paper we consider a waveform relaxation method for the initial value problem

$$\frac{du}{dt} + Au = f,$$
$$u(0) = u_0.$$

where A is a positive definite Hermitian matrix, i.e. $A = \bar{A}^T$ and the Euclidean inner product $\langle Au, u\rangle > 0$ whenever $u \neq 0$. The waveform relaxation method (also known as a dynamic iteration method) is an iterative method of the following form

$$\frac{du^{n+1}}{dt} + F(u^{n+1}, u^n) = f, \tag{1.1}$$
$$u^{n+1}(0) = u_0, \tag{1.2}$$

where the function F satisfies the identity

$$A(u) = F(u, u)$$

*Department of Applied Mathematics and Theoretical Physics, University of Cambridge, England (email S.Keras@damtp.cam.ac.uk)

and u^0 is chosen arbitrarily provided it satisfies the initial condition $u^0(0) = u_0$. The method was known already more than one hundred years ago (see [5], [1]), however, it did not attract attention as a practical method for solving ODEs until Lelarasmee et al. ([2]) published a paper on its applications in the process of simulating large circuits. It is especially efficient when applied to stiff equations, since decoupling makes it possible to solve the equations separately, applying different time scales for different parts of the system and possibly using parallel computers.

Since semidiscretization of PDEs produces large stiff systems of ODEs, the waveform relaxation technique was lately adopted for solving PDEs, most notably for parabolic equations. In this case A originates in a semidiscretization of a linear elliptic operator, and we will restrict ourselves to the linear decoupling of the system:

$$\frac{du^{n+1}}{dt} + Pu^{n+1} = Qu^n + f, \tag{1.3}$$

$$u^{n+1}(0) = u_0, \tag{1.4}$$

where

$$A = P - Q$$

The splittings which are typically used in this case are *Jacobi*, *Gauss-Seidel* and *Successive Over Relaxation*, where matrices P and Q are chosen as in corresponding static iterative methods. In other words, writing $A = D - L - U$ where D, L and U are diagonal, lower triangular and upper triangular matrices respectively, we have $P = D$ for the Jacobi splitting, $P = D - L$ for the GS splitting and $P = \frac{1}{\omega}D - L$ for the SOR splitting. An extensive analysis of the waveform relaxation method with these splittings has been done in [3] and [4], in particular, necessary and sufficient conditions for the convergence has been established. While each iterative step is computationally cheap and easy to implement in parallel, the main drawback of these methods appears to be a slow rate of convergence. As it was proved in [3], the rate of convergence is approximately $1 - 2\pi^2h^2$ for the SOR splitting and $1 - \pi^2h^2$ for the Gauss-Seidel splitting. Different methods were suggested to overcome this difficulty, most notably *Mutigrid Waveform Relaxation* ([6]). In this paper we suggest a different splitting, where matrix P is a Toeplitz or block Toeplitz approximation of A. In this case the rate of convergence does not depend on h, and in case of linear parabolic equations is significantly less than 1. Since matrix P is Toeplitz, it is possible to apply fast Helmholtz equation solvers for the equation (1.3) which guarantees the fast convergence of the methods.

2 Toeplitz Waveform Relaxation

We commence this section by stating the problem. Consider the equation

$$\frac{du}{dt} + Au = f, \tag{2.1}$$

$$u(0) = u_0, \tag{2.2}$$

where A is a hermitian positive definite matrix, i.e. $A = \bar{A}^T$ and $\langle Au, u\rangle > 0$ whenever $u \neq 0$. We solve this equation by a waveform relaxation method

$$\frac{du^{n+1}}{dt} + Pu^{n+1} = Qu^n + f, \tag{2.3}$$

$$u^{n+1}(0) = u_0, \tag{2.4}$$

where $A = P - Q$.

Solving (2.3-4) explicitly by integration of constants, u^{n+1} can be formally written as

$$u^{n+1} = \mathcal{K}u^n + \phi,$$

where

$$\mathcal{K}u(t) = \int_0^t e^{(s-t)P} Qu(s)ds,$$

$$\phi(t) = e^{-tP}u_0 + \int_0^t e^{(s-t)P} f(s)ds,$$

and it follows from the Banach fixed point theorem that the method (2.3–4) converges for all f and u^0 if and only if $\rho(\mathcal{K}) < 1$, where ρ denotes the spectral radius of the operator. The following theorem is crucial to our analysis.

Theorem 2.1 Let A be hermitian matrix. Then the method (2.3–4) converges if and only if

$$2|\langle Qu, u\rangle| < \langle (P + P^H)u, u\rangle. \tag{2.5}$$

for any $u \neq 0$

Proof. As it was proved in [3], as long as all the eigenvalues of A and P have positive real parts, the spectral radius of $\mathcal{K}$ can be represented by means of the formula

$$\rho(\mathcal{K}) = \max_{\xi \in R} \rho((i\xi I + P)^{-1}Q). \tag{2.6}$$

Let $\lambda = x + iy$ be an eigenvalue of $(i\xi I + P)^{-1}Q$. Then

$$Qu = \lambda(i\xi I + P)u \tag{2.7}$$

for some u and

$$\langle Qu, u\rangle = \lambda\langle (i\xi I + P)u, u\rangle. \tag{2.8}$$

Without loss of generality we may assume $\langle u, u\rangle = 1$ and write $\langle Pu, u\rangle = r + ip$, $\langle Qu, u\rangle = s + it$ where $r, p, s, t \in \mathbb{R}$. Comparing real and imaginary parts of (2.8) we obtain

$$s = xr - yp - y\xi,$$

$$t = xp + yr + x\xi.$$

Solving this equation with respect to x and y yields

$$x = \frac{t(p+\xi) + sr}{(p+\xi)^2 + r^2},$$
$$y = \frac{tr - s(p+\xi)}{(p+\xi)^2 + r^2},$$

and

$$\lambda^2 = \frac{t^2+s^2}{(p+\xi)^2 + r^2} \leq \frac{t^2+s^2}{r^2}. \tag{2.9}$$

Hence, the method converges if $\frac{t^2+s^2}{r^2} < 1$. However, $|\langle (P+P^H)u, u\rangle|^2 = 4r^2$ and $|\langle Qu, u\rangle|^2 = t^2+s^2$, which proves the "if" part of the theorem. For the second part let us assume that

$$2|\langle Qu, u\rangle| \geq \langle (P+P^H)u, u\rangle \tag{2.10}$$

for some u. Without loss of generality we may assume $\langle u, u\rangle = 1$ and, as above, we write $\langle Qu, u\rangle = s + it$ and $\langle Pu, u\rangle = r + ip$. Let $\xi = -p$. Then

$$\langle (i\xi I + P)u, u\rangle = r = \frac{1}{2}\langle (P+P^H)u, u\rangle. \tag{2.11}$$

Combining (2.11) and assumption (2.10) yields $\rho((i\xi + P)^{-1}Q) \geq 1$ when $\xi = -p$. $\square$

Remark. If $P = P^H$, then the equation (2.5) reduces to the inequality

$$|\langle Qu, u\rangle| < \langle Pu, u\rangle,$$

or, after making the substitution $Q = P - A$,

$$|\langle (P-A)u, u\rangle| < \langle Pu, u\rangle.$$

This is equivalent to Theorem 3.3 in [3] which states that the method (2.3–4) converges if and only if $2P - A$ is positive definite.

Next we apply the above result to linear parabolic equations.

Theorem 2.2 Consider a parabolic equation

$$u_t - \nabla(a(x)\nabla u(x)) = f, \quad (x,t) \in \Omega \times (0,\infty), \tag{2.12}$$
$$u(0,x) = u_0(x), \quad x \in \Omega, \tag{2.13}$$
$$u(t,x) = 0, \quad (x,t) \in \partial\Omega \times [0,\infty). \tag{2.14}$$

where Ω is a rectangular domain in $\mathbb{R}^d$. Let A be a discretization of the elliptic operator $-\nabla(a(x)\nabla)$, $0 < a_- \leq a(x) \leq a_+ < \infty$ and P be a discretization of the operator $-\nabla(b(x)\nabla)$ for some function b such that $0 < b_- \leq b(x) \leq b_+ < \infty$ and let both A and P satisfy the following assumptions

1. A and P are positive definite

2. $c\langle Pu, u\rangle < \langle Au, u\rangle < C\langle Pu, u\rangle$ for any vector $u \neq 0$ as long as $cb(x) < a(x) < Cb(x)$ for all $x \in \Omega$, where $c, C \in \mathbb{R}$

Then the method (2.3–4) converges if $\max\left|\frac{a(x)-b(x)}{b(x)}\right| < 1$.

Proof. Since bot A and P are positive definite, we can again use the result from [3], which says that $\rho(\mathcal{K}) = \max_{\xi\in R}\rho((i\xi I+P)^{-1}Q)$, so that it suffices to estimate the spectral radius of the latter operator. Again, let λ be an eigenvalue of $K_\xi = (i\xi I + P)^{-1}Q$. Then we can write

$$Qu = \lambda(i\xi I + P)u.$$

Taking an inner product with u we obtain

$$\langle Qu, u\rangle = \lambda\langle i\xi I + Pu, u\rangle. \tag{2.15}$$

As in the previous theorem, let $\lambda = x+iy$, $\langle u, u\rangle = 1$, $\langle Pu, u\rangle = p$, $\langle Qu, u\rangle = s$. Comparing real and imaginary parts of (2.15) yields

$$\begin{aligned} s &= -y\xi + xp, \\ 0 &= x\xi + yp. \end{aligned}$$

Multiplying the first equation by x and the second equation by y results in

$$xs = |\lambda|^2 p.$$

However, under our assumptions about P and A

$$\langle Au, u\rangle \le \max\frac{a(x)}{b(x)}\langle Px, x\rangle$$

which yields

$$s = \langle (A-P)u, u\rangle \le \max\left|\frac{a(x)-b(x)}{b(x)}\right|\langle Pu, u\rangle = \max\left|\frac{a(x)-b(x)}{b(x)}\right| p.$$

According to the remark to the previous theorem, this ensures the convergence of the method. Moreover, we can also estimate where exactly the eigenvalues of the operator $(i\xi I + P)^{-1}Q$ are. The following inequalities hold

$$\begin{aligned} \left|\lambda + \max\left|\frac{a(x)-b(x)}{2b(x)}\right|\right| &\le \max\left|\frac{a(x)-b(x)}{2b(x)}\right|, \\ \left|\lambda - \max\left|\frac{a(x)-b(x)}{2b(x)}\right|\right| &\le \max\left|\frac{a(x)-b(x)}{2b(x)}\right|, \end{aligned}$$

i.e. the eigenvalues lie in two symmetric circles, both with the radius $r = \max\left|\frac{a(x)-b(x)}{2b(x)}\right|$, and the centres at $\pm\max\left|\frac{a(x)-b(x)}{2b(x)}\right|$, which completes the proof. Figure 2 shows the distribution of the eigenvalues for the matrix $(i\xi I + P)^{-1}Q$ when $a(x) = 1 + 0.5\sin\pi x$ and $b(x) = 1$. $\square$

Remark. The method also converges for nonrectangular domains, however, in this case the matrix P need not be Toeplitz or block Toeplitz and it is no longer possible to apply fast solvers based on Fast Fourier Transform techniques.

Corollary 2.1 If $b(x) = C$ then

$$\rho(\mathcal{K}) \leq \max \left| \frac{a(x) - C}{C} \right| \tag{2.16}$$

and

$$\min_{C \in \mathbb{R}^+} \rho(\mathcal{K}) \leq \frac{a_+ - a_-}{a_+ + a_-} \tag{2.17}$$

Proof. The first statement is a direct consequence of the theorem. It is easy to check that (2.16) is minimised when $C = \frac{a_{max}+a_{min}}{2}$, where $a_{max} = \max_{x\in\Omega} a(x)$ and $a_{min} = \min_{x\in\Omega} a(x)$. In this case

$$\min_{C \in \mathbb{R}^+} \rho(\mathcal{K}) \leq \frac{a_{max} - a_{min}}{a_{max} + a_{min}} \leq \frac{a_+ - a_-}{a_+ + a_-}.$$

□

Corollary 2.2 A standard $2d + 1$ point finite difference discretization in a rectangular domain satisfies the assumptions of the theorem.

Proof. The positive definiteness of the matrices A and P follows from the fact that both A and P are symmetric and diagonally dominant with positive entries along their diagonals. If $a(x) > cb(x)$ then $A - cP$ is symmetric and diagonally dominant with positive entries along the diagonal, hence positive definite. Similarly, if $a(x) < Cb(x)$ then $CP - A$ is symmetric and diagonally dominant with positive entries along the diagonal, hence positive definite. □

3 The algorithm

The last corollary in the previous section allows us immediately to construct a method for solving the equation (2.12–14). We approximate the coefficient $a(x)$ by a constant $C = \frac{a_{max}+a_{min}}{2}$ and use its discretization as an operator P in the equation 2.3. Since P in this case is a Toeplitz or a block Toeplitz matrix, we will call this method a *Toeplitz* waveform relaxation. The rate of convergence in this case is not greater than $\frac{a_{max}-a_{min}}{a_{max}+a_{min}}$, which, in the case of mildly varying coefficient $a(x)$, is close to 0 and does not depend on h! As it is proved in [4], any A stable method produces a convergent discrete waveform relaxation method, so that for the discretization in time one can chose, for instance, a backward Euler method or a trapezoidal rule.

At each time step a new iterate can be calculated using the value of the new iterate at the previous time step (or, in case of multistep methods, in several previous steps) and the value of the old iterate at the present time step. We describe this in Figure 1 where the arrows show the values needed for the calculation of the next iterate. It is easy to

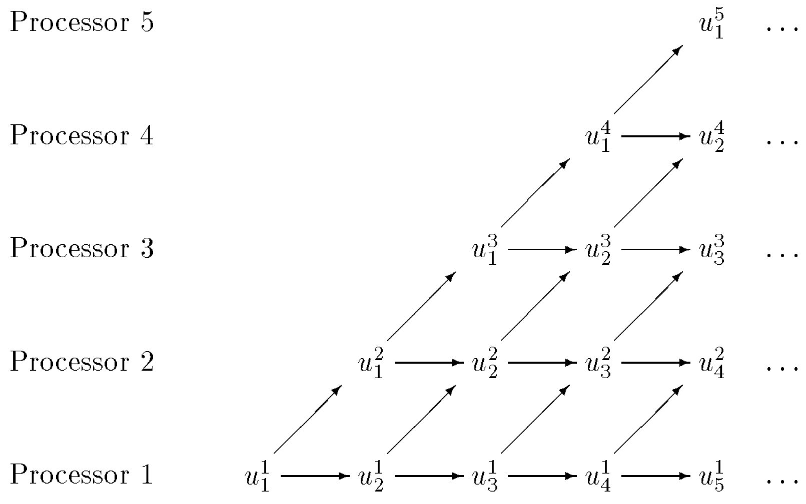

Figure 1: A graphical description of the iterative procedure for parabolic equations when using waveform relaxation method.

see that, for instance, u_2^1 and u_2^1 (where a superscript denotes the number of iteration and a subscript denotes the number of the time step) can be calculated simultaneously and independently of each other. The same is true for u_3^1, u_2^2, u_1^3. It is trivial to show by induction that this is true for any u_m^n along the vertical line $m + n = const$. This means that the algorithm can be efficiently implemented in parallel, each new iterate being calculated on an independent processor immediately using the values obtained by another processor for the previous iterate.

4 Numerical Experiments

In this section we present the results of numerical experiments with Toeplitz operator splitting. As a test equation we chose the following parabolic problem:

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left((1+\alpha \sin 4\pi x \sin 4\pi y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left((1+\alpha \sin 4\pi x \sin 4\pi y)\frac{\partial u}{\partial y}\right) \quad (4.1)$$

$$(x, y, t) \in \Omega \times (0,1), \quad \Omega = (0,1) \times (0,1)$$

$$u(x, y, t) = 0, \quad (x, y, t) \in \partial\Omega \times (0,1) \quad (4.2)$$

$$u(x, y, 0) = \sin \pi x \sin \pi y \quad (4.3)$$

Here) $< \alpha < 1$, which ensures that the problem is parabolic. This equation was discretized using five-point approximation for the space variables and backward Euler approximation for the time variable and was solved using SOR and Toeplitz methods with different grid sizes and different vales of α. The value of ω for the SOR method was chosen so that to minimise the total number of iterations. Since for the SOR method the calculation time is virtually independent of α, we only present the results with $\alpha = 0.5$. For both methods the stopping criterion was $||u^{n+1} - u^n|| < TOL$, where u^n, u^{n+1} denote two successive iterates and $TOL = 10^{-6}$. For the Toeplitz method we have chosen $b(x, y) = 1$, in which case the spectral radius of the operator $\mathcal{K}$ is $\rho_{\mathcal{K}} \approx \alpha$. All calculations were carried out on a Sun SPARC-10 workstation.

Figures 3 to 5 present the plots of the execution times and of the number of iterations for different values of α and different grid sizes. As expected, the SOR algorithm shows no sensitivity to the values of α, while Toeplitz algorithm performs considerably better when α is close to 0 (the case of mildly varying coefficients) and its performance deteriorates when α approaches 1 (the case of highly varying coefficients). Since the convergence of the waveform relaxation methods is linear, in both cases the number of iterations L can be estimated by means of the inequality

$$\rho^L ||e^0|| < TOL$$

where $||e^0|| = ||u^0 - u||$ is the norm of the initial error and ρ is the spectral radius of the operator $\mathcal{K}$. This yields

$$L \approx \frac{\ln(TOL/||e^0||)}{\ln \rho}, \tag{4.4}$$

In our example $\rho \approx \alpha$ and, as $\alpha \to 1$, $\ln \alpha \approx \alpha - 1$, so that the number of iterations grows inverse linearly as α approaches 1

Next we compare asymptotically the number of the operations required for the two methods in order to achieve the given tolerance TOL when space discretization step h tends to 0. In both cases it is equal to the number of iterations L times the number of the operations in each iteration M. L can be estimated as in (4.4). It is known (see [3]) that for the SOR method $\rho \approx 1 - 2\pi^2 h^2$ so that

$$L_{SOR} \approx \frac{\ln(TOL/||e^0||)}{\ln(1 - 2\pi^2 h^2)} \approx -\frac{\ln(TOL/||e^0||)}{2\pi^2 h^2} = O(n^2),$$

where $n = 1/h$ is a number of grid points at the unit interval. On the other hand at each step the resulting linear equations are lower triangular with $O(n^2)$ nonzero entries, so that the solutions of such equations requires $O(n^2)$ operations. Thus, the total cost of the SOR method can be estimated as

$$L_{SOR} M_{SOR} = O(n^4).$$

For the Toeplitz method we know that ρ is independent of h, which means that

$$L_T \approx \frac{\ln(TOL/||e^0||)}{\ln \rho_T} = O(1)$$

as n grows to infinity. The number of the operations required to solve a block tridiagonal matrix (which is the case if the domain is rectangular) is $M_T = O(n^2 \ln n^2) = O(n^2 \ln n)$, provided that n can be factorised into a product of small primes. This allows us to write

$$\frac{L_{SOR} M_{SOR}}{L_T M_T} = O(n^2 / \ln n)$$

Figures 6 depicts the results obtained in numerical experiments. In this case α was fixed and we have chosen different values of n. The first graph shows the ratio of the times

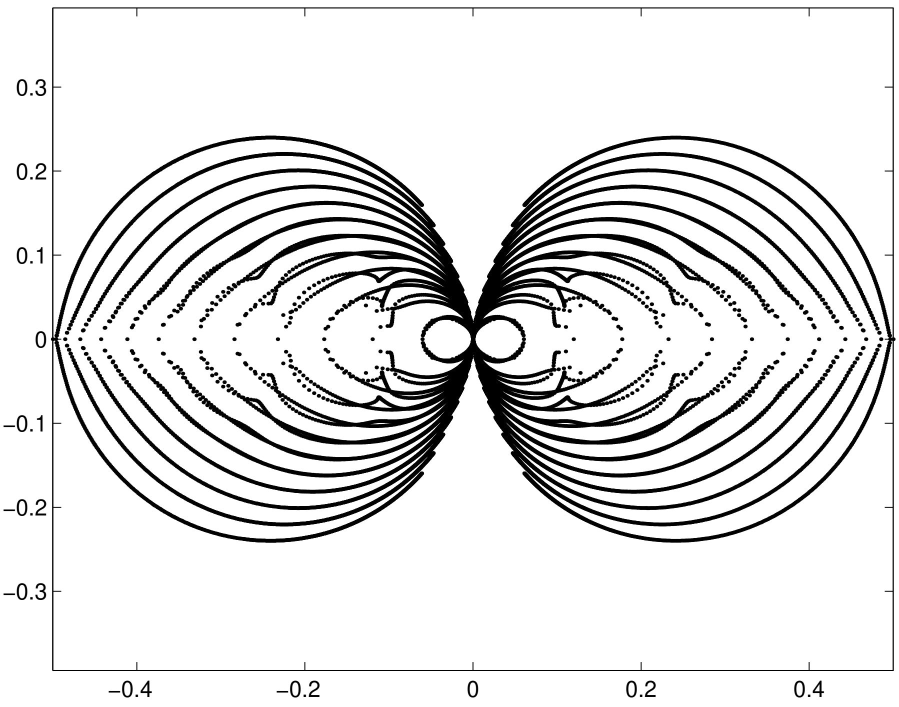

Figure 2: The eigenvalues of $K(\xi)$, $A = \frac{d}{dx}((1+0.5\sin\pi x)\frac{d}{dx})$, $P = \frac{d^2}{dx^2}$ discretized with a mesh size $h = 0.02$, $a_+ = 1.5$, $a_- = 0.5$.

required for two methods for different values of n and the second graph shows the same ratio multiplied by $\ln n/n^2$. One can see that the results of the experiment closely agree with the theoretical predictions.

The results of the experiments allow us to conclude that even for moderate grid sizes ($n = m = 16$) the Toeplitz method outperforms the SOR method if coefficients are mildly varying. The advantage of the Toeplitz method becomes even greater when the number of the grid points increases.

5 Acknowledgments

The author would like to than Dr Arieh Iserles for many useful comments on this work. The author was supported by Leslie Wilson Scholarship from Magdalene College, Cambridge and ORS award.

References

[1] E Lindelöf. Sur l'application des méthodes d'approximations successives á l'etude des intégrales réeles des équations différentielles ordinaires. *Journal de Mathématiques*

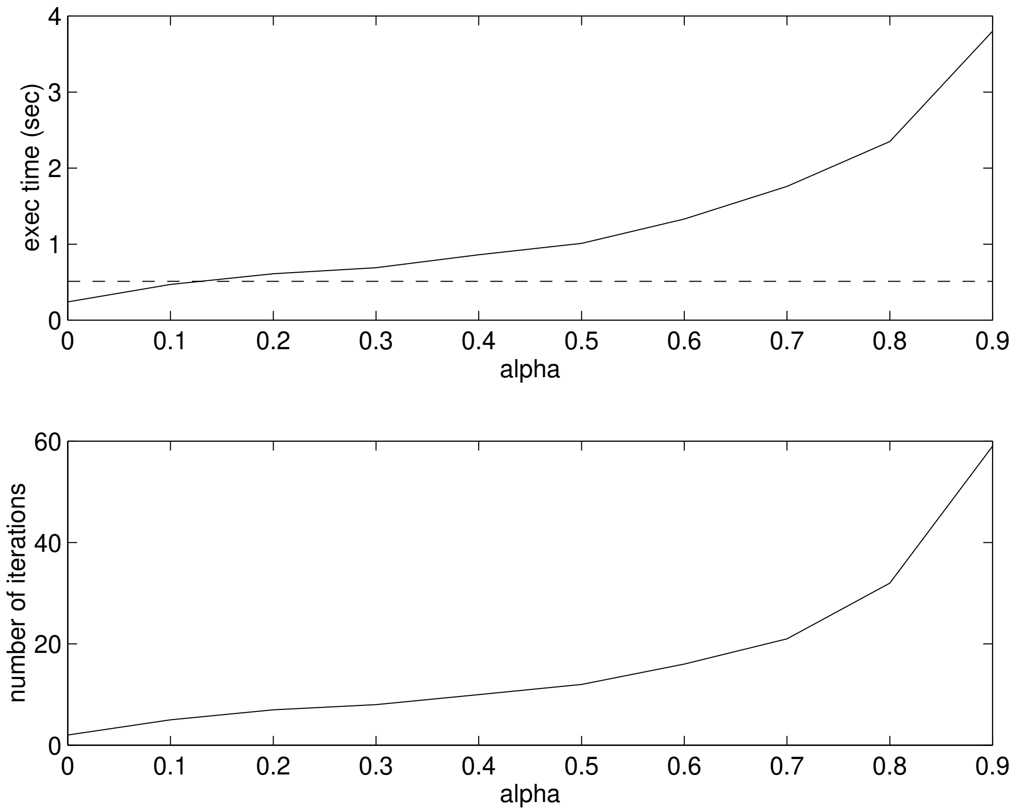

Figure 3: The execution time and the number of iterations for the Toeplitz method with the number of grid points $m = n = 16$. The dashed line represents the execution time for the SOR method

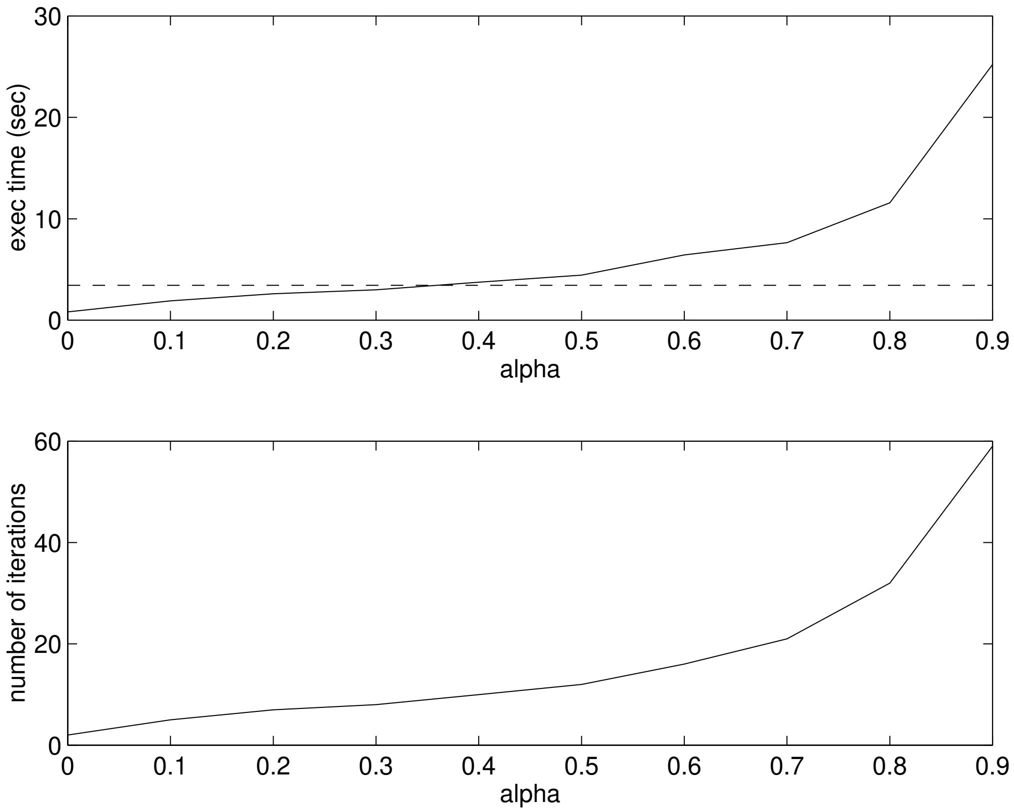

Figure 4: The execution time and the number of iterations for the Toeplitz method with the number of grid points $m = n = 32$. The dashed line represents the execution time for the SOR method

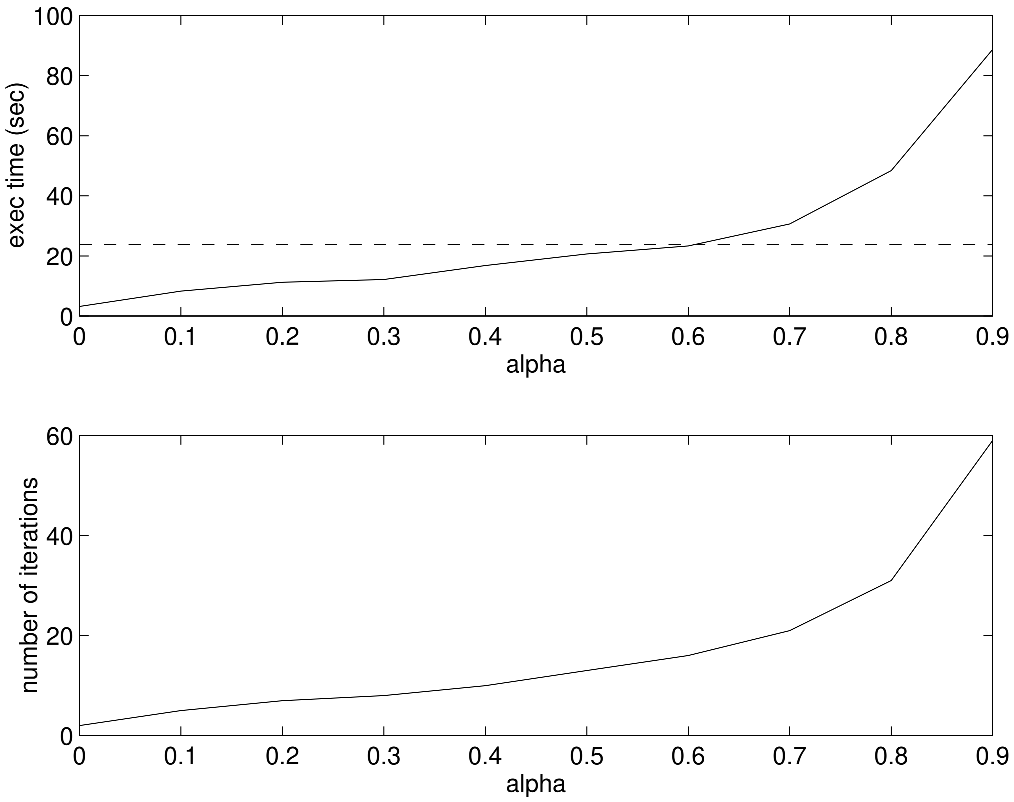

Figure 5: The execution time and the number of iterations for the Toeplitz method with the number of grid points $m = n = 64$. The dashed line represents the execution time for the SOR method

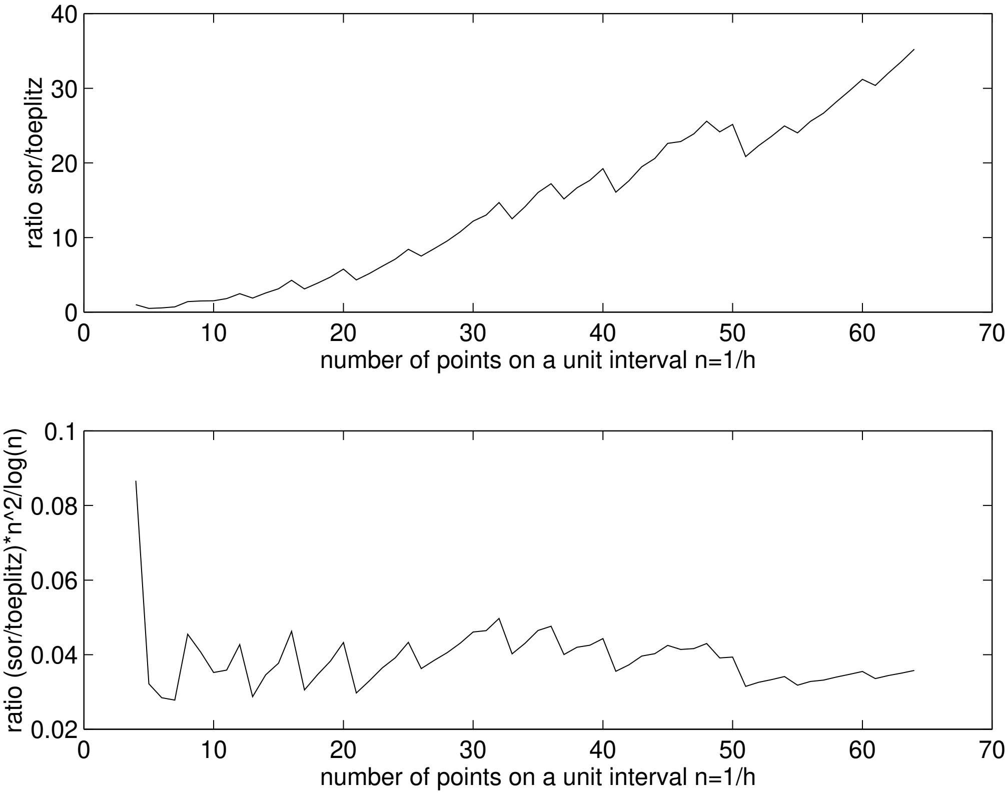

Figure 6: The ratio of the execution times required for the SOR and Toeplitz methods

Pures et Appliquées, **10**:117–128, 1894.

[2] E Lelarasmee, A E Ruehli, and A L Sangiovanni-Vincentelli. The waveform relaxation method for time-domain analysis of large scale integrated circuits. *IEEE Trans. on CAD of IC and Syst.*, **1**:131–145, 1982.

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

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

[5] É Picard. Sur l'application des méthodes d'approximations successives á l'etude de certaines équations différentielles ordinaires. *Journal de Mathématiques Pures et Appliquées*, **9**:217–271, 1893.

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