TMA4205 - PROJECT 1

OTTAR HELLAN

1. Introduction

In this project some iterative techniques are explored for solving sparse linear systems, primarily applied to a partial differential equation in two dimensions. An acceleration of the examined methods is presented, in view of similarities to gradient descent methods, and is shown to improve convergence under certain parameters.

2. Exercise 1

In this exercise we explore and compare three different iterative techniques for solving the linear system

$$A\boldsymbol{u} = \boldsymbol{b}, \tag{2.1}$$

namely the Jacobi, forward Gauss-Seidel and successive over relaxation methods. We apply the methods to a finite-difference discretization of a partial differential equation.

The iterative scheme is defined by the recursive relation

$$A_1\boldsymbol{u}^{k+1} = A_2\boldsymbol{u}^{k+1} + \boldsymbol{b}, \tag{2.2}$$

where $A = A_1 - A_2$ and $\det(A_1) \neq 0$ ensures a unique solution in each iteration. A fixed point of this iteration will be the unique solution of the linear system (2.1), assuming A is non-singular.

Let A_d, A_l, and A_u be the diagonal, the negative strictly lower triangular, and negative strictly upper triangular parts of A, such that $A = A_d - A_l - A_u$. We then define the three methods by

$$\text{Jacobi:} \qquad A_1 = A_d, \tag{2.3}$$

$$\text{Forward Gauss-Seidel:} \qquad A_1 = A_d - A_l, \tag{2.4}$$

$$\text{Successive over relaxation:} \qquad A_1 = A_d - \omega A_l, \tag{2.5}$$

where ω can be chosen to maximize the convergence rate. Typically $1 < \omega < 2$.

Dependent on the method chosen and matrix A, it can be practical and cost-efficient to calculate the inverse of A_1 and state the iteration instead as

$$\boldsymbol{u}^{k+1} = G\boldsymbol{u}^{k+1} + \boldsymbol{f}, \quad G = A_1^{-1}A_2,\ f = A_1^{-1}\boldsymbol{b}, \tag{2.6}$$

simplifying the iteration somewhat.

Our principal test problem is the five-point central-difference discretization of the Poisson equation in two dimensions with homogeneous Dirichlet boundary conditions,

$$\nabla^2 u(x,y) = f(x,y), \quad (x,y) \in \Omega = [0,1] \times [0,1], \tag{2.7}$$

$$u(x,y) = 0, \quad (x,y) \in \partial\Omega \tag{2.8}$$

on a regular $n \times n$-grid of interior points. The system is represented in vector form $\boldsymbol{u}, \boldsymbol{f} \in \mathbb{R}^{n^2}$ and $u_{I(x,y)} = u(x,y)$, $f_{I(x,y)} = f(x,y)$ as shown in figure 1.

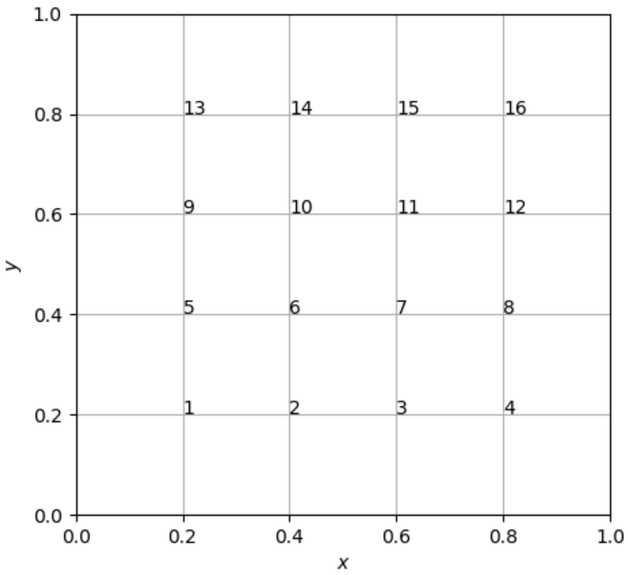

FIGURE 1. The indexing scheme for turning two-dimensional discrete functions into vectors.

The discrete Laplacian corresponding to the five-point central-difference scheme in our grid can be written in block-tridiagonal form as

$$\Delta x^2 L = \begin{pmatrix} B & I_n & & \\ I_n & \ddots & \ddots & \\ & \ddots & \ddots & I_n \\ & & I_n & B \end{pmatrix} \in \mathbb{R}^{n^2 \times n^2},\ B = \begin{pmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{pmatrix} \in \mathbb{R}^{n \times n}, \tag{2.9}$$

with I_n being the $n \times n$ identity matrix. Solving our discretization of the Poisson equation is now equivalent to solving the linear system

$$L\boldsymbol{u} = \boldsymbol{b},\ \boldsymbol{b} := \Delta x^2 \boldsymbol{f}. \tag{2.10}$$

We test the methods using grid size $n = 10$ and define the residual at each step $\boldsymbol{r}^k := \boldsymbol{b} - A\boldsymbol{u}^{k+1}$. We halt the iteration and accept the current value as soon as $||\boldsymbol{r}^k||_2 \leq \text{TOL}$ or $||\boldsymbol{r}^k||_2/||\boldsymbol{r}^0||_2 \leq \text{RTOL}$ for some specified tolerances. If the number of iterations, k, surpasses some number $I_{\max}$, we halt the iteration without accepting the answer.

In figure 2 the relative residuals of the three methods are plotted against the number of iterations. The iterations are run until a tolerance of $\text{TOL} = \text{RTOL} = 10^{-7}$ is achieved. For successive over relaxation (SOR), the value of $\omega = 1.5$ is chosen.

The figure shows that, in this case, the Jacobi method needs about 350 iterations to converge, with forwards Gauss-Seidel needing approximately half of that, and SOR with $\omega = 1.5$ needing approximately half again of what Gauss-Seidel requires. Table 1 and as figure 3 show the run time of each method as well as the average time spent per iteration in each of the three methods in two typical test runs with tolerances $\text{TOL} = \text{RTOL} = 10^{-7}$ and $\text{TOL} = \text{RTOL} = 10^{-14}$. Although SOR and Gauss-Seidel methods spend more time on each iteration, because of how few iterations are needed, they are much more efficient.

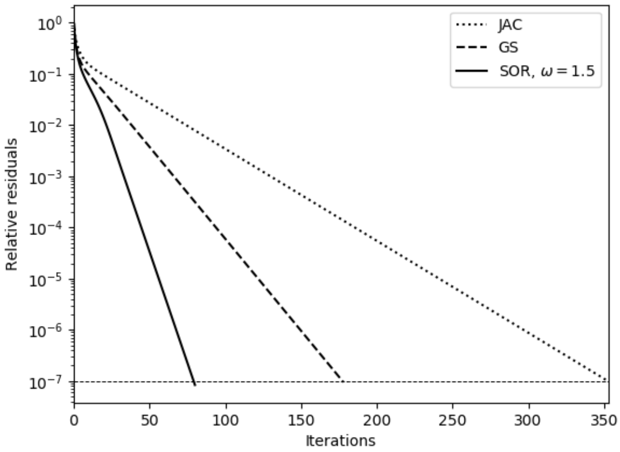

FIGURE 2. The progression of the relative residuals of the Jacobi, forward Gauss-Seidel and successive over relaxation methods on the Poisson test problem. The direct and relative tolerances are equal.

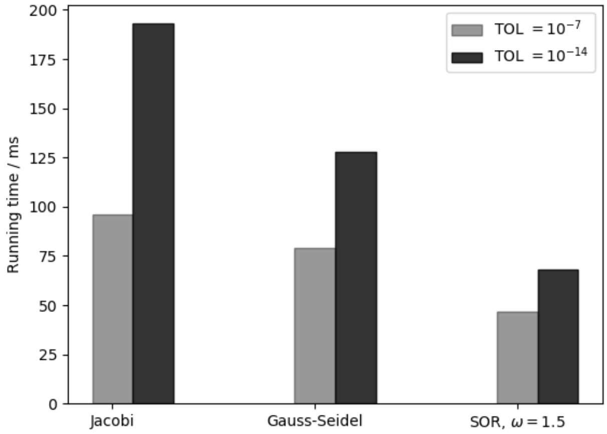

FIGURE 3. The time needed to converge for the Jacobi, forward Gauss-Seidel and successive over relaxation methods on the Poisson test problem under two different tolerances. The direct and relative tolerances are equal.

Figure 4 shows the iterations needed to converge for a selection of ω-values from 1 to 2. As can be seen, the most efficient values of ω are around 1.57. With $\omega = 1.57$, the iteration converges in 53 steps. Also, with choices of ω approaching 2, SOR is less worthwhile than Gauss-Seidel, and eventually also less efficient than the Jacobi method, because of the extra time needed for each step.

By theorem 4.1 in Saad[1], a necessary and sufficient condition for iterative schemes of the form (2.2) to converge is that the spectral radius

$$\rho(G) = \rho(A_1^{-1}A_2) < 1, \tag{2.11}$$

Tolerance	Method	Iterations	Run time / ms	Step time / ms
10^{-7}	Jacobi	353	95.8	0.271
	Gauss-Seidel	178	79.3	0.446
	SOR, $\omega = 1.5$	80	46.7	0.584
10^{-14}	Jacobi	353	193	0.260
	Gauss-Seidel	178	128	0.342
	SOR, $\omega = 1.5$	80	67.9	0.424

Table 1. The iterations needed to converge, time needed to converge and average time spent per step for the Jacobi, forward Gauss-Seidel and successive over relaxation methods on the Poisson test problem under two different tolerances. The direct and relative tolerances are equal.

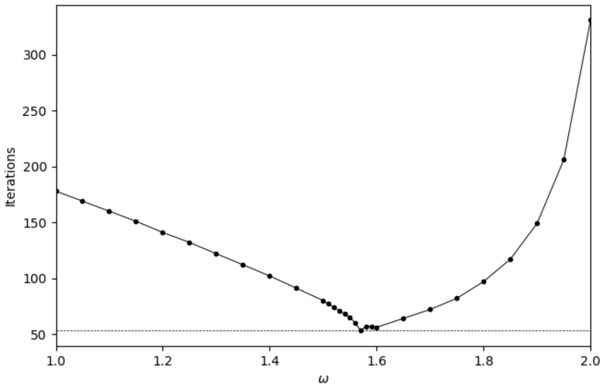

Figure 4. The number of iterations needed for the successive over relaxation method to converge for different values of ω between 1 and 2 on the Poisson test problem with direct and relative tolerance TOL $= 10^{-7}$.

independent of constant $\boldsymbol{b}$ and starting vector $\boldsymbol{u}^0$. For the three methods, the calculated spectral radii are shown in table 2. All three methods have spectral radius of $\rho(A_1^{-1}A_2) < 1$, matching our findings of convergence.

Method	$\rho(A_1^{-1}A_2)$
Jacobi	0.959
Gauss-Seidel	0.921
SOR, $\omega = 1.57$	0.724

Table 2. The spectral radii of $A_1^{-1}A_2$ for the Jacobi, forward Gauss-Seidel and successive over relaxation methods on the Poisson test problem.

If we let $\boldsymbol{u}$ be the unique fixed point of iteration scheme (2.6), such that

$$\boldsymbol{u} = G\boldsymbol{u} + \boldsymbol{f} \iff \boldsymbol{f} = \boldsymbol{u} - G\boldsymbol{u},$$

and let the error in each iteration be $\boldsymbol{d}^k = \boldsymbol{u}^{k+1} - \boldsymbol{u}$, we get that

$$\begin{aligned} \boldsymbol{u}^{k+1} &= G\boldsymbol{u}^{k+1} + \boldsymbol{f} \\ \boldsymbol{u}^{k+1} - \boldsymbol{u} &= G\boldsymbol{u}^{k+1} + \boldsymbol{u} - G\boldsymbol{u} - \boldsymbol{u} \\ \boldsymbol{u}^{k+1} - \boldsymbol{u} &= G\left(\boldsymbol{u}^{k+1} - \boldsymbol{u}\right) \\ \boldsymbol{d}^{k+1} &= G\boldsymbol{d}^k. \end{aligned} \tag{2.12}$$

From this we get that

$$||\boldsymbol{d}^{k+1}|| \leq ||G|| \cdot ||\boldsymbol{d}^k|| \leq \rho(G)||\boldsymbol{d}^k|| \tag{2.13}$$

$$\iff ||\boldsymbol{d}^k|| \leq \rho(G)^k||\boldsymbol{d}^0||. \tag{2.14}$$

Since $||\boldsymbol{r}^k|| = ||A\boldsymbol{u}^k - \boldsymbol{b}|| = ||A\boldsymbol{u}^k - A\boldsymbol{u}|| = ||A\boldsymbol{d}^k||$, this means that

$$||\boldsymbol{r}^k|| \leq C\rho(G)^k \qquad \frac{||\boldsymbol{r}^k||}{||\boldsymbol{r}^0||} \leq C_r\rho(G)^k. \tag{2.15}$$

Assuming this is not an overly pessimistic bound, we will get that

$$\ln ||\boldsymbol{r}^k|| \approx \ln C + k \ln \rho(G) = \alpha + \beta k.$$

Performing simple linear regression on the measured residuals with the three methods, we find that the performance very closely matches the above estimate, as shown in table 3.

Method	$\rho(G)$	Est. $\exp\beta$
Jacobi	0.959	0.959
Gauss-Seidel	0.921	0.920
SOR, $\omega = 1.57$	0.724	0.753

Table 3. The spectral radius of G and estimated convergence constant the observed convergence would correspond to for the Jacobi, forward Gauss-Seidel and successive over relaxation methods on the Poisson test problem.

3. Exercise 2

It can be shown that iterations of the form (2.2) are equivalent to gradient descent methods under the A_1-inner product $\langle \boldsymbol{u}, \boldsymbol{v}\rangle_{A_1} = \langle \boldsymbol{u}, A_1\boldsymbol{v}\rangle$ for minimizing the related quadratic function

$$E : \mathbb{R}^n \to \ R, \qquad E(\boldsymbol{u}) = \frac{1}{2}\boldsymbol{u}^T A\boldsymbol{u} - \boldsymbol{b}^T\boldsymbol{u}, \tag{3.1}$$

when both A and A_1 are symmetric, positive-definite.

Gradient descent methods can be accelerated by momentum techniques such as the following two-step Polyak heavy ball method

$$\boldsymbol{p}^{k+1} = \boldsymbol{p}^k - hA_1^{-1}\left(A\boldsymbol{u}^k - \boldsymbol{b}\right) - h\lambda\boldsymbol{p}^k \tag{3.2}$$

$$\boldsymbol{u}^{k+1} = \boldsymbol{u}^k + h\boldsymbol{p}^{k+1}, \tag{3.3}$$

where h and λ are chosen as to maximize convergence, and $\boldsymbol{p}^0 = \boldsymbol{0}$. Here, h is the step size constant common in optimization methods and λ is a weighting for how much the previous step should be used.

We test this heavy ball method on the Jacobi iteration (2.3) and a new method using the maximum eigenvalue of $A-A_d$. Let σ be the maximum eigenvalue of $A-A_d$ and $\boldsymbol{v}$ its associated eigenvector. We then define the method by

$$A_1 = A_d + \sigma \boldsymbol{v}\boldsymbol{v}^T \tag{3.4}$$

and refer to it as the maximum eigenvalue method. Assuming $A-A_d$ is non-singular, this can be seen as a simplification of the spectral decomposition of $A-A_d = V\Sigma V^T$, where the columns of V are the eigenvectors of $A-A_d$ and Σ is the diagonal matrix with corresponding eigenvectors, in which the elements of Σ not corresponding to the maximum eigenvalue are set to zero.

Because of the Sherman-Morrison formula, as stated in appendix A.27 of Nocedal and Wright[2],

$$\left(X + \boldsymbol{v}\boldsymbol{w}^T\right)^{-1} = X^{-1} - \frac{X^{-1}\boldsymbol{w}\boldsymbol{v}^T X^{-1}}{1+\boldsymbol{w}^T X^{-1}\boldsymbol{v}}, \tag{3.5}$$

assuming X and $\left(X + \boldsymbol{w}\boldsymbol{v}^T\right)$ are non-singular, the inverse of A_1 can easily be directly computed and

$$A_1^{-1} = A_d^{-1} - \sigma\frac{\boldsymbol{v}A_d^{-2}\boldsymbol{v}}{1+\sigma\boldsymbol{v}^T A_d^{-1}\boldsymbol{v}}, \tag{3.6}$$

since the diagonal matrix A_d has an easily found inverse.

In the case of the Laplace test problem, the eigenvalue and eigenvector are given by

$$\sigma = 4\cos\left(\frac{\pi}{n+1}\right), \qquad v_{l,m} = \sin\left(l\frac{\pi}{n+1}\right)\sin\left(m\frac{\pi}{n+1}\right), \tag{3.7}$$

in the $n \times n$-system. Since $L_d = \operatorname{diag}(-4, ..., -4)$, the explicit form for L_1^{-1} in the maximum eigenvalue method (3.4) is given by

$$L_1^{-1} = -\frac{1}{4}\left(I_{n^2} + \frac{\sigma}{4-\sigma}\boldsymbol{v}\boldsymbol{v}^T\right), \tag{3.8}$$

with σ and $\boldsymbol{v}$ given by (3.7).

Figure 5 shows a run of the Jacobi and maximum eigenvalue methods along with their non-heavy ball counterparts. The parameters h, λ are here not chosen to be optimal in any way, but are some values that give an improvement on the standard non-accelerated methods. As can be seen in the progression of the accelerated Jacobi iteration, the residual does not monotonously decrease. This is because the step direction $u^{k+1} - u^k$ is no longer guaranteed to be a descent direction.

Determining the optimal values of h and λ in the Polyak heavy ball method (3.2), (3.3) is non-trivial, but an approximate answer can be obtained none the less. We assume the optimal values for each factor are between zero and two and then check how fast each value of (h, λ) converges in a regular grid of size $N_1 \times N_1$ covering $[0, 2] \times [0, 2]$. Centered on the (h, λ) giving fastest convergence, we make a smaller grid of size $N_2 \times N_2$ of width 0.8. The (h, λ) giving the fastest convergence is taken as our approximation to the optimal choice of (h, λ). Such arguments are not very useful for understanding the techniques, but show some of the potential of the heavy ball methods.

Figure 6 shows the second, finer grid plotted along with the number of iterations to reach convergence. The maximum number of iterations was set to 400, which is why the graph tops out evenly. The fineness of the first grid was set to $N_1 = 5$ and the

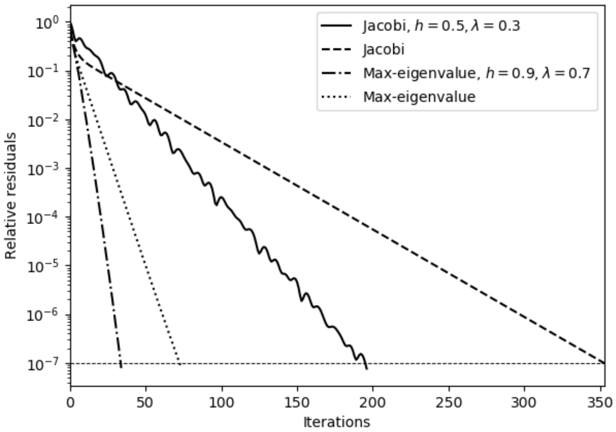

FIGURE 5. The progression of the relative residuals of the accelerated and un-accelerated versions of the Jacobi and maximum eigenvalue methods on the Poisson test problem.

fineness of the second grid was set to $N_2 = 11$. This gave optimal values of $h = 0.34$ and $\lambda = 1.24$, converging in 58 iterations.

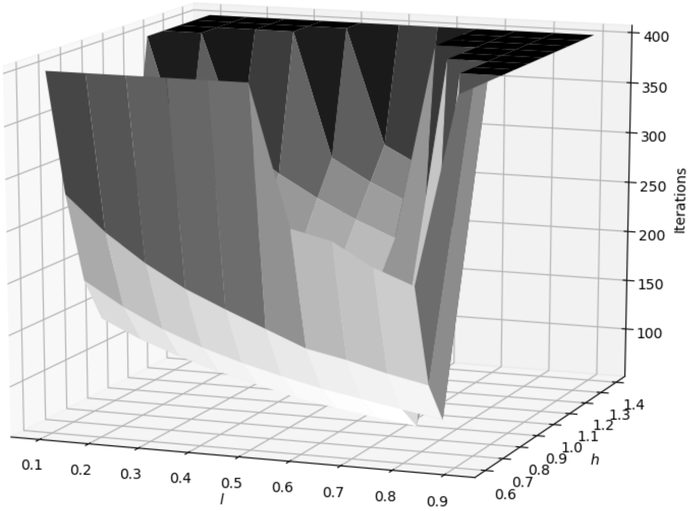

FIGURE 6. 3D plot showing the number of iterations needed to halt for (h, λ)-values in the second grid when finding approximate optimal (h, λ)-values.

To test our methods on a range of different problems, we generate some random symmetric, positive-definite matrices. To this end, we create sparse, block-diagonal matrices,

$$A = \begin{pmatrix} B & & \\ & \ddots & \\ & & B \end{pmatrix} \in \mathbb{R}^{n^2},\ B \in \mathbb{R}^n, \tag{3.9}$$

with the diagonal elements B being symmetric, positive-definite matrices. The matrices B were constructed to be symmetric, positive-definite by letting

$$B = Q\Lambda Q^T, \tag{3.10}$$

where Q is a randomly generated orthogonal matrix and Λ is a diagonal matrix with elements drawn from a uniform distribution between λ_{min} and λ_{max}. The orthogonal matrix was created by performing Householder orthogonalization on a random $n \times n$-matrix with elements uniformly drawn between 0 and 1.

Figure 7 shows the Jacobi and maximum eigenvalue methods run on four test matrices generated with (3.9) and (3.10) along with their accelerated versions. The matrices were created with parameters $\lambda_{\text{min}} = 0$ and $\lambda_{\text{max}} = 0.99$, to ensure positive-definiteness and some level of convergence being possible. The parameters are kept the same with each test matrix, for the maximum eigenvalue method we use $h = 0.9$ and $\lambda = 0.7$ and for the Jacobi method we use the $h = 0.34$ and $\lambda = 1.24$ we previously found to approximate the optimal values.

The un-accelerated methods quickly have equal, or close to equal, residuals, while the heavy ball methods differ greatly in their convergence. The maximum eigenvalue heavy ball method converges monotonously, while the Jacobi method oscillates throughout. The generated systems do not converge for all starting seeds however, for unknown reasons.

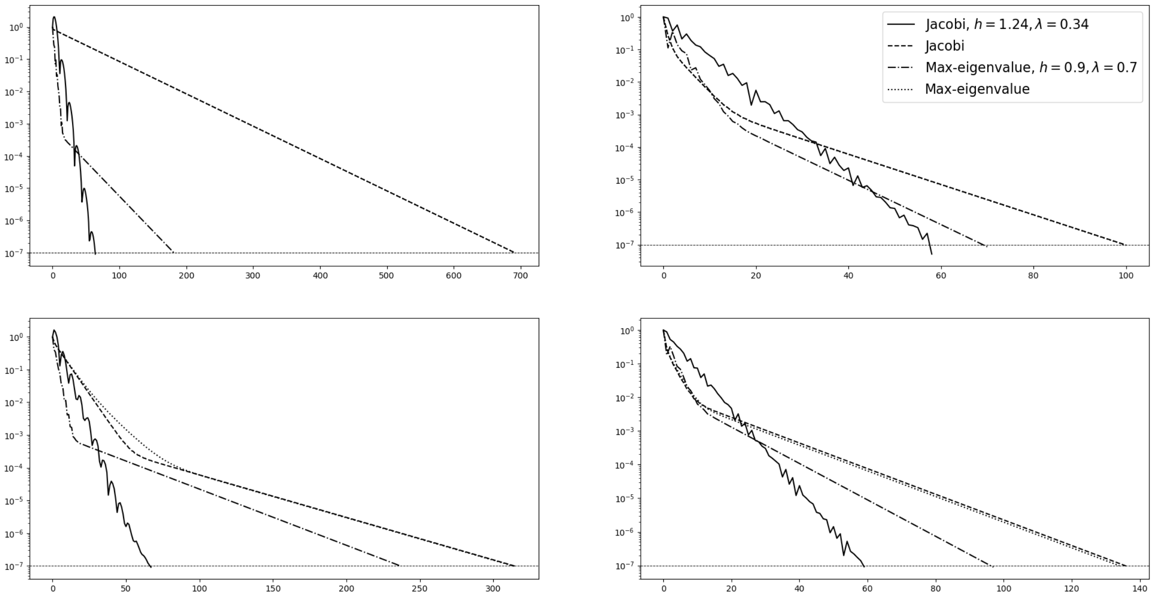

FIGURE 7. The progression of the relative residuals of the accelerated and un-accelerated versions of the Jacobi and maximum eigenvalue methods for four different test problems, with equal h and λ parameters.

4. CONCLUSION

Classical iterative schemes such as the Jacobi, forward Gauss-Seidel, and successive over relaxation methods are used for solving sparse linear systems, in particular one rising from a discretization of the Poisson equation in two dimensions. A method using a greatest eigenvalue is described in (3.4) and it performs competitively with the Jacobi method. Iterative schemes of the form (2.6) are viewed as gradient descent methods under different inner products and accelerated using a Polyak heavy ball momentum technique, giving good results accelerating convergence greatly with good parameter choices. The accelerated and un-accelerated Jacobi and maximum eigenvalue methods are tested with random test systems to mixed results.

References

[1] Yousef Saad. *Iterative Methods for Sparse Linear Systems, second edition.* Society for Industrial and Applied Mathematics, Philadelphia, 2003.

[2] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization, second edition.* Springer, New York City, 2006.