Norwegian University of Science
and Technology
Department of Mathematical
Sciences

TMA4205 Numerical Linear Algebra Fall 2017

Exercise set 2

1 We return to solving the one dimensional Poisson problem that was discussed in the first exercise set.

$$\begin{aligned} -\frac{\partial^2 u}{\partial x^2} &= 4\pi^2 \sin(2\pi x), \quad && x \in [0,1], \\ u &= 0, && x \in \{0,1\}. \end{aligned}$$

Let us again use the finite difference method on a uniform grid with step-size $h = 1/n$ and grid points $x_j = jh$. The discretized equations can be expressed as $Au = b$ where A represents the discrete Laplacian. We now want to solve this system of linear equations by three different iterative methods: the Jacobi iteration, steepest descent (SD), and minimal residual (MR) iteration.

a) Suppose that we want to reduce the initial error by 5 orders of magnitude. Estimate the number of iterations required in the Jacobi method and with SD.

b) Suppose that we want to reduce the initial residual by 5 orders of magnitude. Estimate the number of iterations required in MR.

c) Discuss the computational cost (complexity) of the three iterative methods.

d) What are the relative advantages (if they exist) of the various methods, both in terms of solving the Poisson problem, and in the more general context of solving linear systems?

Possible solution

In general, if we have an error/residual behaviour given by

$$\|e_k\| \sim \rho^k \|e_0\|,$$

then

$$\log \frac{\|e_k\|}{\|e_0\|} \sim k \log \rho.$$

For this exercise we want to obtain an error/residual reduction by 10^{-5}, such that $\log(\|e_k\|/\|e_0\|) \sim -5$, i.e.,

$$-5 = k \log \rho. \tag{1}$$

We will need the following useful approximations based on Taylor expansions:

$$\begin{aligned} \cos(x) &\approx 1 - \frac{1}{2}x^2, \\ \log(1+x) &\approx x, \\ (1+x)^{1/2} &\approx 1 + \frac{1}{2}x. \end{aligned}$$

Jacobi. For this case we have that ρ is the spectral radius of the iteration matrix, so $\rho = \cos(\pi/n) \approx 1 - \pi^2/2n^2$. From (1) and the approximation of the logarithm, we find that $k \approx 10n^2/\pi^2$.

Steepest descent. From the lectures we know that

$$\|e_k\|_A \leq \rho^k \|e_0\|_A,$$

where

$$\rho = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\kappa - 1}{\kappa + 1} \approx 1 - \frac{2}{\kappa}$$

for large values of $\kappa = \lambda_{\max}/\lambda_{\min}$, the condition number based on the Euclidean norm of A. From (1) and the approximation of the logarithm, we find that $k \approx 5\kappa/2$. We also have that

$$\kappa = \frac{\lambda_{\max}}{\lambda_{\min}} = \frac{\frac{2}{h^2}\Big(1 - \cos\big(\frac{(n-1)\pi}{n}\big)\Big)}{\frac{2}{h^2}\Big(1 - \cos\big(\frac{\pi}{n}\big)\Big)} \approx \frac{2}{\pi^2/2n^2} = \frac{4n^2}{\pi^2},$$

and thus

$$k \approx \frac{10n^2}{\pi^2}. \tag{2}$$

b) We now consider MR iteration for the same problem, and wish to reduce the initial residual with 5 orders of magnitude. In this case,

$$\|r_k\|_2 \leq \rho^k \|r_0\|_2$$

with

$$\rho = \Big(1 - \frac{\mu^2}{\sigma^2}\Big)^{1/2},$$

where

$$\mu = \lambda_{\min}\Big(\frac{A + A^T}{2}\Big) \quad \text{and} \quad \sigma = \|A\|_2.$$

In our case, A is symmetric and positive-definite (SPD), andwe thus get the simplified expressions

$$\mu = \lambda_{\min}(A) \quad \text{and} \quad \sigma = \lambda_{\max}(A).$$

Hence,

$$\rho = \Big(1 - \frac{\lambda_{\min}^2}{\lambda_{\max}^2}\Big)^{1/2} = \Big(1 - \frac{1}{\kappa^2}\Big)^{1/2} \approx 1 - \frac{1}{2\kappa^2}.$$

From (1) and the approximation of the logarithm, we find that $k \approx 10\kappa^2$. We insert the expression for the condition number from (2), and find

$$k \approx \frac{160n^4}{\pi^4}.$$

We may now compare the value for k for Jacobi, SD and MR, and observe that

$$\frac{k_{\mathrm{SD/J}}}{k_{\mathrm{MR}}} \approx \frac{1}{4\kappa} = \frac{\pi^2}{16n^2}.$$

Here, $k_{\mathrm{SD/J}}$ is the number of iterations for Jacobi and steepest descent to reduce the error by 5 orders of magnitude, while k_{MR} is the number of iterations for MR to reduce the initial

residual by the same amount. Since we expect κ to be large for large n, this difference is significant!

c) We now discuss the computational cost for the three iterative methods. For one iteration, each method requires one matrix-vector multiplication. The Jacobi method in addition requires the addition of two vectors, while both SD and MR method require in total six additions of two vectors, multiplications with a scalar, or inner products. Since the matrix A is tri-diagonal, each matrix-vector multiplication can be performed in $5n$ multiplications and additions; in the case of the Jacobi method, the iteration matrix only has two non-zero diagonals, and thus this reduces to $3n$ multiplications. Thus each step of the Jacobi method is roughly half as expensive than each step of the other two methods.

Concerning memory requirements, all methods need to store enough information about A to be able to performmatrix-vector products. The sparsity of A should here be exploited, and we thus only need to store the non-zero entries, which in this case is $O(n)$ (in fact, the non-zero entries in each row are the same). Both SD and MR need to store three vectors x, p and r. For Jacobi we need to store x as well as a working array. Thus, the memory requirement is $O(n)$ for all methods.

We now consider the cost for k iterations. The memory requirement remains the same, while the number of operations is given by

$$\mathcal{N}_{\text{tot}} = k\mathcal{N}_1,$$

where $\mathcal{N}_1$ is the number of operations in one iteration and $\mathcal{N}_{\text{tot}}$ the total number of iterations. Using the results from a) and b), we find that the Jacobi and SD methods require $O(n^3)$ iterations, while the MR method requires n^5 iterations.

d) Of the three iterative methods considered here, MR is the most general one, since the requirement is only that A is positive-definite (and it still may converge without that assumption). However, we see that this method is much slower than the other two for the one-dimensional Poisson problem. Steepest descent is guaranteed to converge as long as A is SPD, while Jacobi iteration has an even stronger requirement, namely that the spectral radius of the iteration matrix must be less than 1 (which is not the case for all SPD matrices). For our particular Poisson problem, Jacobi is a little bit faster than the steepest descent, while Jacobi and steepest descent are both much faster than MR iteration. In conclusion, a method for more general problems is typically slower than a specialized method, and for our Poisson problem, MR iteration is definitely not a good idea. We also note here that we have compared error reduction by 5 orders of magnitude for the first two methods, and residual reduction for the MR method. This is obviously not the same, but we assume that they are comparable and show the same behaviour.

2 **Saad, Exercise 5.3** In Section 5.3.3, it is shown that using a one-dimensional projection method with $\mathcal{K} = \text{span}\{A^T r\}$ and $\mathcal{L} = \text{span}\{AA^T r\}$ is equivalent to using the steepest descent method on the normal equations $A^T Ax = A^T b$ for a matrix $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$.

Show that an orthogonal projection method with solution space $\mathcal{K} = \mathcal{L}$ for solving the equation $A^T Ax = A^T b$ is equivalent to applying a projection method onto $\mathcal{K}$ orthogonally to $\mathcal{L} = A\mathcal{K}$ for the problem $Ax = b$.

Possible solution

We first consider the system $Ax = b$ with $\mathcal{L} = A\mathcal{K}$. Let

$$\mathcal{K} = \text{span}\{v_1, \ldots, v_m\},$$
$$V_m = [v_1 | \ldots | v_m] \in \mathbb{R}^{n \times m}.$$

In the first case, the approximate solution $\tilde{x}$ to $Ax = b$ must satisfy

$$\tilde{x} - x_0 \in \mathcal{K},$$
$$(AV_m)^T(b - A\tilde{x}) = 0.$$

Next, we consider the system $A^TAx = A^Tb$ with $\mathcal{L} = \mathcal{K}$. The approximate solution must now satisfy

$$\tilde{x} - x_0 \in \mathcal{K},$$
$$V_m^T(A^Tb - A^TA\tilde{x}) = 0,$$

but the second equation is the same as

$$(AV_m)^T(b - A\tilde{x}) = 0.$$

Hence we get the same conditions for the two methods, and thus they are equivalent.

3 Assume that a real matrix A is anti-symmetric, that is, $A^T = -A$. Describe the structure of the Hessenberg matrix H_m resulting from Arnoldi process in this case. Explain how this structure can be utilized for an efficient implementation of the Arnoldi algorithm.

Possible solution

fter applying m steps of the Arnoldi process to a matrix A we obtain the matrix V_m containing the orthonormal basis for $\mathcal{K}_m(A, v_1)$, and an upper Hessenberg matrix H_m satisfying the equality $H_m = V_m^TAV_m$. Assuming that $A^T = -A$, the matrix on the right hand side of the equality sign is anti-symmetric. Therefore H_m is also antisymmetric and thus has only two non-zero diagonals:

$$H_m = \begin{pmatrix} 0 & -h_{21} & \ldots & 0 \\ h_{21} & 0 & \ddots & 0 \\ 0 & \ddots & \ddots & -h_{m,m-1} \\ 0 & \ldots & h_{m,m-1} & 0 \end{pmatrix}.$$

As a result, Arnoldi process simplifies to (note: only the sub-diagonal of H is computed):

- Initialise $v_1 := v/k\|v\|_2$, $v_0 := 0$, $h_{1,0} := 0$.
- For $j = 1, \ldots, m$ do:
  - $w \leftarrow Av_j + h_{j,j-1}v_{j-1}$

- $h_{j+1,j} \leftarrow \|w\|_2$
- $v_{j+1} \leftarrow w/h_{j+1,j}$

4 Show that both CG and GMRES are scaling independent. That is, when applied to a system $(\lambda A)x = \lambda b$ for some $\lambda \in \mathbb{R} \setminus \{0\}$ ($\lambda > 0$ in the case of CG in order to preserve positive definiteness) they produce exactly the same sequence of iterates in exact arithmetics. (Of course, since the residual vectors will be scaled by λ, this may affect stopping criteria for the methods.)

Possible solution

One can insert scaled variables into all parts of the algorithm and observe that the output at each step is the same. However, the short story is: all residuals need to be scaled by λ, but the search space and the constraint space remain the same. Indeed, $\text{span}(r_0, Ar_0, \ldots, A^{m-1}r_0) = \text{span}(\lambda r_0, \lambda^2 Ar_0, \ldots, \lambda^m A^{m-1}r_0)$. Furthermore, $r_m \perp \mathcal{L}$ if and only if $\lambda r_m \perp \mathcal{L}$, as $\mathcal{L}$ is a linear space.