SF3580
HW 2

Anna Broms & Fredrik Fryklund

2018/11/29

Task 2

(a) & (b)

For all α there is an isolated eigenvalue. All predicted rates of convergence are based on either one circle containing all the eigenvalues or two circles, where one is a point consisting of the outlying eigenvalue. The circles are found in Figure 3–6.
We observe faster convergence and faster predicted convergence for larger α, see Figure (1) and Figure (2). This is expected, as most eigenvalues are clustered around the origin for $\alpha = 1$. Setting $\alpha = 5$ translates the centre of the cluster away from zero, although some eigenvalues are still close to the origin. Note that we may still have convergence within fewer iterations than the rank of A, which is the case for $\alpha = 5$, but not for $\alpha = 1$. For the latter 100 iterations is required, which is the maximum rank for a 100×100 matrix. As expected the estimated rates of convergence are not of any use and diverges. For $\alpha = 5$ we have convergence and the predicted rates of convergence are indeed bounds, but not very useful due to the underestimation. Furthermore, we see that the predicted rate of convergence for one disc is better than the corresponding for two discs. This is not the case for $\alpha = 10$, 100, where we also have faster convergence.

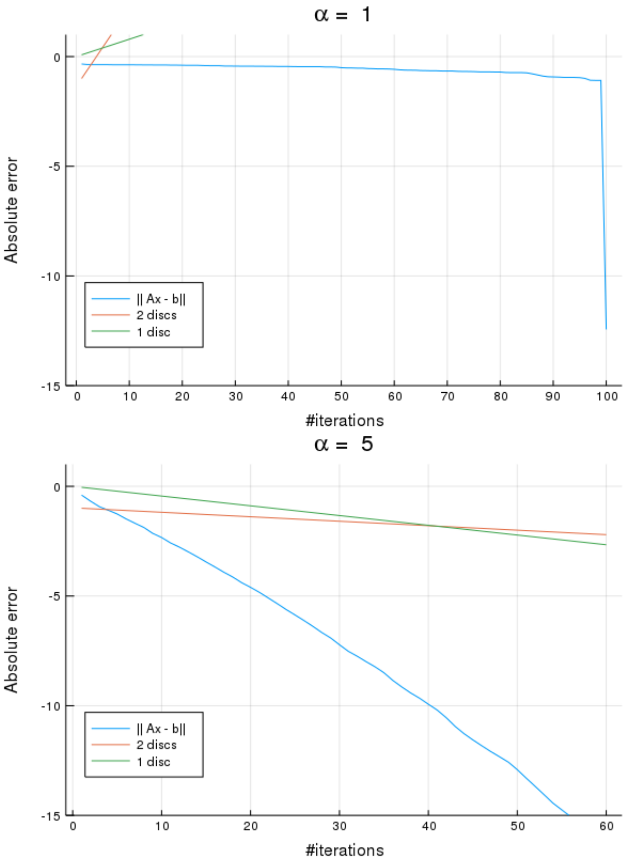


Figure 1: Task 1, (a) & (b): The convergence of the residual for gmres, plotted against the number of iterations. Here for $\alpha = 1$ and $\alpha = 5$

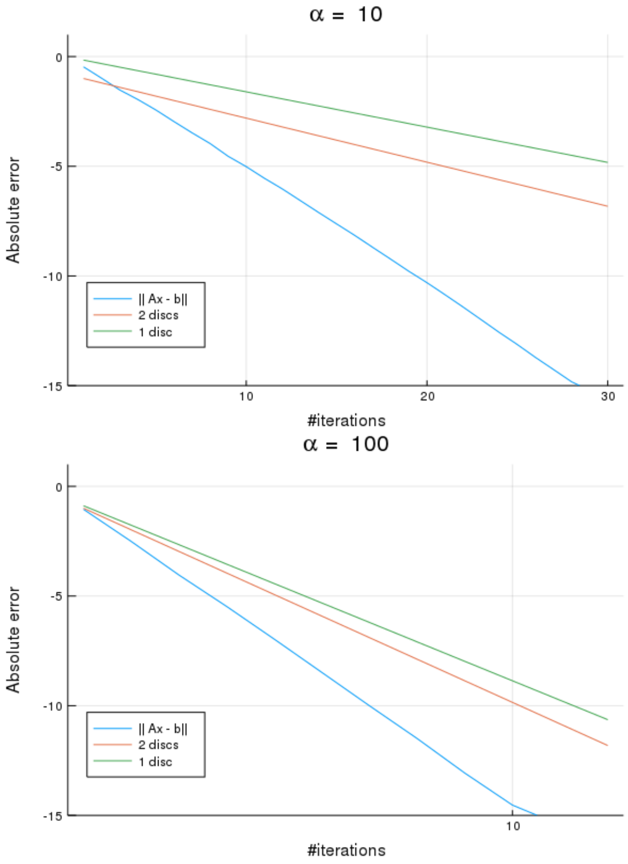


Figure 2: Task 1, (a) & (b): The convergence of the residual for gmres, plotted against the number of iterations. Here for $\alpha = 10$ and $\alpha = 100$

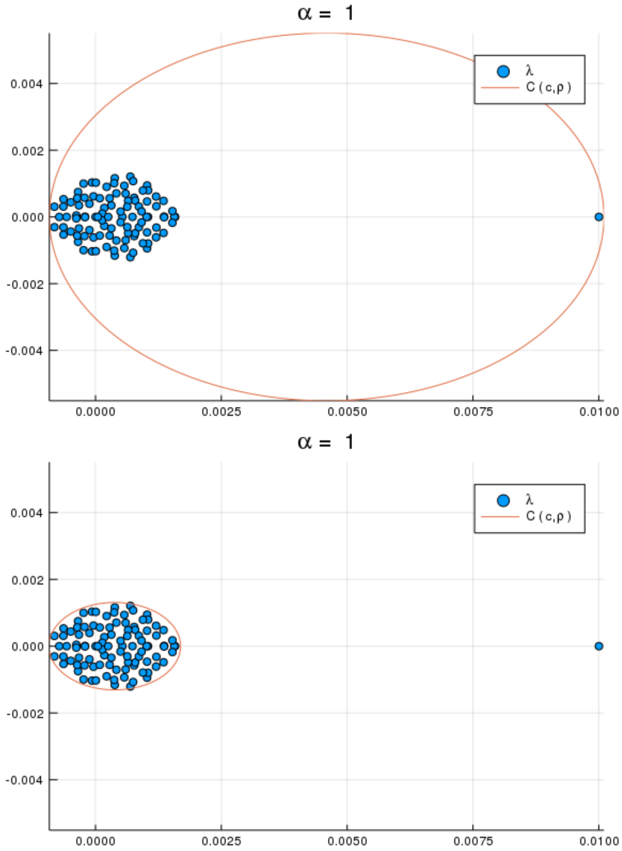


Figure 3: Task 1, (a) & (b): For $\alpha = 1$. The two configurations of discs containing all eigenvalues.

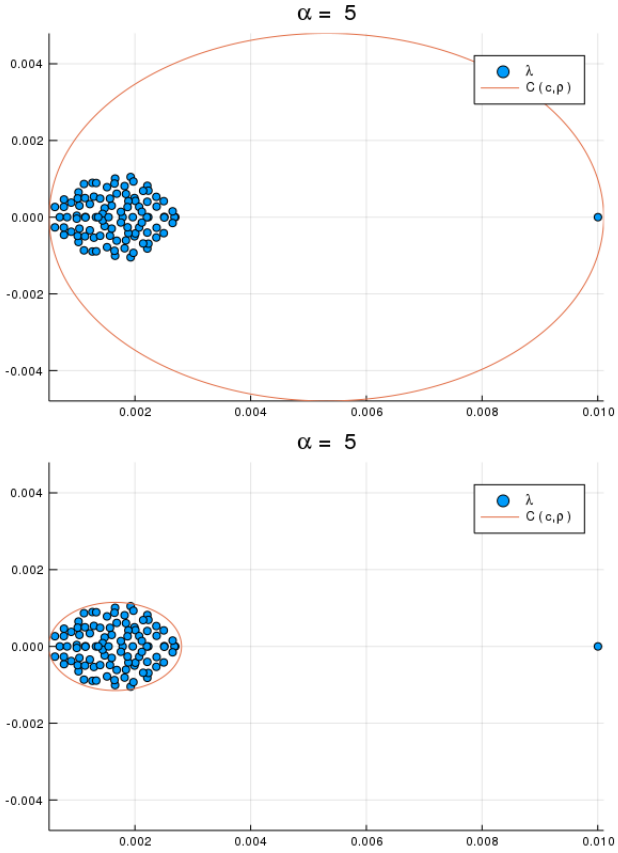


Figure 4: Task 1, (a) & (b): For $\alpha = 5$. The two configurations of discs containing all eigenvalues.

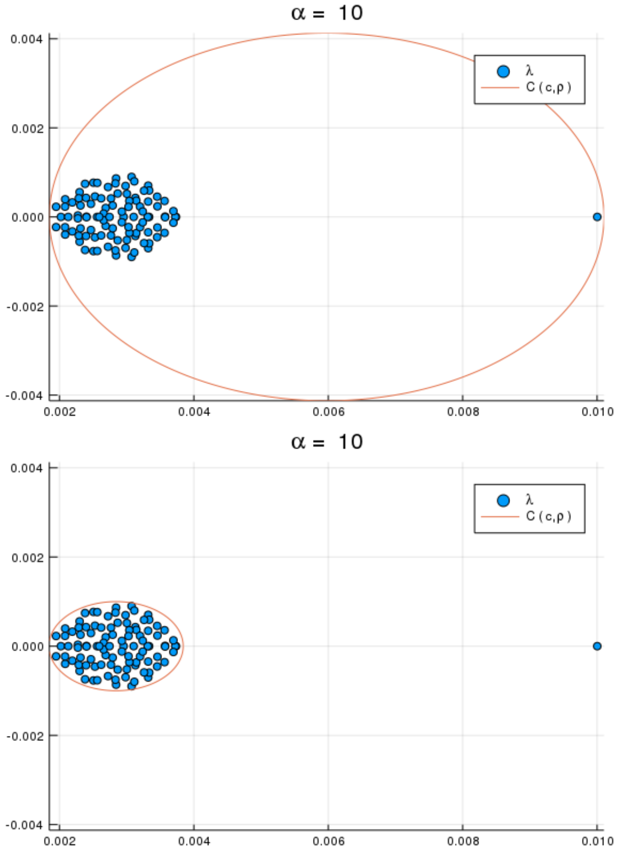


Figure 5: Task 1, (a) & (b): For $\alpha = 10$: The two configurations of discs containing all eigenvalues.

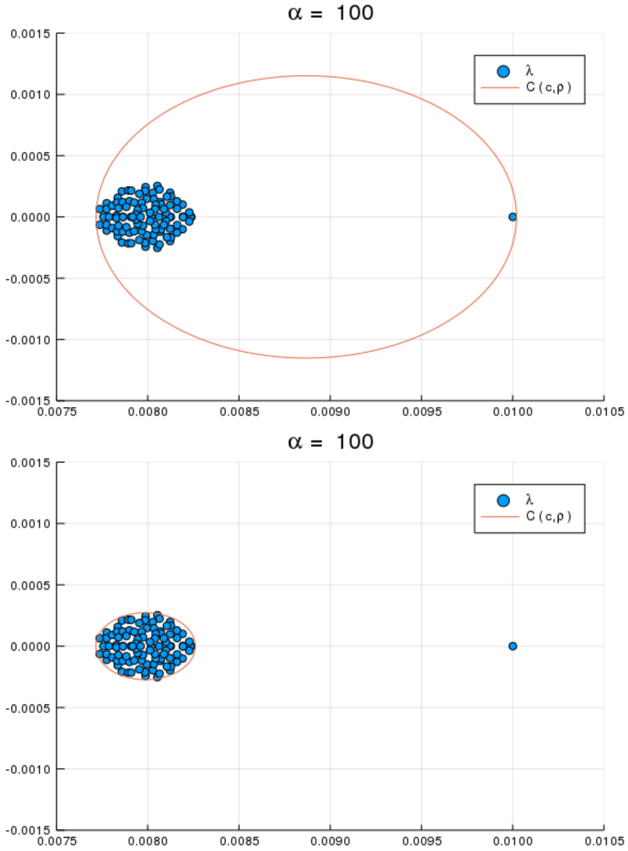


Figure 6: Task 1, (a) & (b): For $\alpha = 100$: The two configurations of discs containing all eigenvalues.

(c)

See the tables below. For gmres we did 1000 samples and one evaluation per sample. For backslash the corresponding digits are 3673 and 1.

Table 1: $\alpha = 1$

	gmres							
	$m = 100$		$m = 200$		$m = 500$		$m = 1000$	
	resnorm	time	resnorm	time	resnorm	time	resnorm	time
n = 5	2.583	162.117 μs	3.848	304.663 μs	6.378	1.511 ms	9.0192	3.915 ms
n = 10	2.548	363.392 μs	3.826	591.597 μs	6.326	2.231 ms	9.007	8.300 ms
n = 20	2.428	1.055 ms	3.744	1.439 ms	6.307	5.081 ms	8.967	16.586 ms
n = 50	1.874	5.246 ms	3.415	8.117 ms	6.194	17.627 ms	8.796	48.643 ms
n = 100	$2.212e-12$	28.858 ms	2.794	32.087 ms	5.718	59.167 ms	8.555	133.731 ms
	Backslash							
	$m = 100$		$m = 200$		$m = 500$		$m = 1000$	
	resnorm	time	resnorm	time	resnorm	time	resnorm	time
	$1.056e-12$	1.350 ms	$4.632e-14$	6.207 ms	$1.868e-13$	45.407 ms	$7.074e-13$	170.036 ms

Table 2: $\alpha = 100$

	gmres							
	$m = 100$		$m = 200$		$m = 500$		$m = 1000$	
	resnorm	time	resnorm	time	resnorm	time	resnorm	time
n = 5	$6.545e-7$	159.159 μs	$5.636e-6$	291.925 μs	$9.712e-5$	962.453 μs	$6.893e-4$	3.790 ms
n = 10	$1.773e-14$	362.841 μs	$7.944e-13$	564.961 μs	$1.667e-10$	1.905 ms	$7.337e-9$	7.722 ms
n = 20	$2.262e-15$	1.014 ms	$4.983e-15$	1.712 ms	$9.328e-15$	6.269 ms	$2.400e-14$	16.591 ms
n = 50	$2.269e-15$	5.625 ms	$5.044e-15$	8.152 ms	$9.291e-15$	18.087 ms	$2.402e-14$	58.155 ms
n = 100	$2.224e-15$	30.860 ms	$5.050e-15$	32.769 ms	$1.030e-14$	58.455 ms	$1.030e-14$	61.989 ms
	Backslash							
	$m = 100$		$m = 200$		$m = 500$		$m = 1000$	
	resnorm	time	resnorm	time	resnorm	time	resnorm	time
	$1.662e-15$	1.313 ms	$3.893e-15$	5.170 ms	$7.977e-15$	35.644 ms	$1.565e-14$	159.975 ms

(d)

For $\alpha = 1$ we conclude that the backslash operator is always superior to the gmres method. Based on the results from (b), the convergence is close to none until $n = m$. Then we reach about $1e - 14$, thus finding an n such that the relative norm of the residual is about $1e - 5$ is not feasible.

For $\alpha = 100$ the situation is different. To obtain the relative residual norm the errors in the table should be scaled by 6, 8, 13 and 18. Then all errors, except for the one corresponding to $m = 1000$ is below $1e - 5$. However, setting $n = 6$ gives a relative residual norm of $1e - 6$ for $m = 1000$ as well. Moreover, gmres is faster: for $n = 5$ the realtive timings $t_{\text{gmres}}/t_{\text{backslash}}$ are: 0.159, 0.0582, 0.0274857 and 0.0238365.

Task 4

The linear system of equations

$$\begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 \end{bmatrix} x = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{1}$$

is considered.

(a)

We determine the constants α, β and γ such that for the iterates $x_0, \dots x_3$ of the conjugate gradient method we obtain $\text{span}(x_0, x_1, x_2, x_3) = \text{span}(c_0, c_1, c_2, c_3)$, where

$$C = [c_0, c_1, c_2, c_3] = \begin{bmatrix} 1 & \alpha & 0 & \gamma \\ 1 & 0 & \beta & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 7 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}. \tag{2}$$

For this purpose, we use Lemma 2.2.4 from the lecture notes, stating that

$$\text{span}(b, Ab, \dots A^{m-1}b) = \text{span}(x_0, x_1, \dots, x_m). \tag{3}$$

Thus, we want to compute α, β and γ such that

$$\text{span}(b, Ab, A^2b, A^3b) = \text{span}(c_0, c_1, c_2, c_3). \tag{4}$$

We can directly conclude that $\text{span}(b) = \text{span}(c_0)$. Next, we want to make sure that $\text{span}(b, Ab) = \text{span}(c_0, c_1)$. By column reduction we thus find that $\alpha = 0$. Using that $\text{span}(b, Ab, A^2b) = \text{span}(c_0, c_1, c_2)$, we can similarly find that $\beta = -1$ and finally, using (4), and again reduce columns, we identify that $\gamma = 6$.

(b)

See the implemented Julia code. We have replaced `???` with $\|Ax-b\|_{A^{-1}}$. Comparing $\texttt{x}_{\texttt{opt}}$ and $\texttt{x}_{\texttt{cg}}$, we obtain a difference of $1.51 \cdot 10^{-11}$.

0.1 (c)

For GMRES, `???` is replaced by $\|Ax-b\|_2$. Now, comparing $\texttt{x}_{\texttt{opt}}$ with $\texttt{x}_{\texttt{gmres}}$, we obtain the difference $1.06 \cdot 10^{-11}$.

Task 5

Given a real symmetric matrix A with eigenvalues 10, 10.5 and 100 eigenvalues in the interval $[2,3]$, we prove a bound for then number of steps needed for CG to reduce the error measured in $\|Ax_n - b\|_{A^{-1}} = \|x_n - x_*\|$ by a factor 10^7. We assume exact arithmetic and no breakdown.

Proof:
Let $\kappa = \lambda_{\max}/\lambda_{\min}$. By Theorem 38.5. in T.B. we have

$$\frac{\|e_n\|}{\|e_0\|} = 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n. \tag{5}$$

We seek n such that the error has been reduced by a factor 10^7, i.e. n such that (5) is about $1e-7$. For the matrix A the largest lower bound for $\lambda_{\min}$ is 2. The largest eigenvalue $\lambda_{\max} = 10.5$. We now solve the following problem for n,

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^n \leq 10^{-7}$$

$$\Leftrightarrow n \leq \frac{\log\left(0.5\,10^{-7}\right)}{\log\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)}$$

$$\Leftrightarrow n \geq 17.967661670561174$$

$$\Leftrightarrow n \geq 18.$$

We confirm this numerically, $n = 17$ gives

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{18} \approx 2.473e-7$$

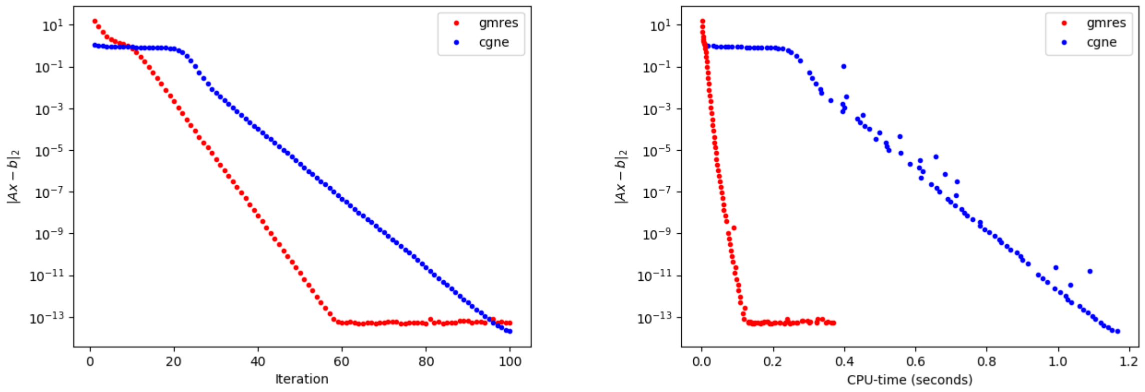
and $n = 18$ gives

$$2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{18} \approx 9.702e-8.$$

Task 6

(a)

We compare GMRES and CGN for a given matrix B and right hand side b. The result of the comparison is visualised in Figure 7. The iterates of CGN span a different Krylov subspace than gmres and it is



(a) Error versus number of iterations required for gmres and cgn for the given system.

(b) Error versus CPU-time required for gmres and cgn for the given system.

Figure 7

mentioned in the lecture notes on this topic that in most cases, this subspace has worse approximation properties than the usual Krylov subspace used for the gmres iterates. This correspond with what can be observed in Figure 7a.

(b)

The result can be related to the convergence theory for CGN and GMRES. Investigating the eigenvalues of the matrix B, it is clear that all eigenvalues are contained in a disk, but for one isolated eigenvalue. However, trying to apply the convergence theory with two discs from the lecture notes is not possible in this case, as the obtained convergence factor is larger than 1. Therefore, one large disc is used and we apply Corollary 2.1.5, choosing $r = 28$ and $c = 30$. The resulting disc and bound is found in Figure 8.

Note that for both methods, the bound for the convergence factor is heavily overestimated. For gmres, a better approach would be to prove wiki problem 2.24 and use this estimate for an isolated eigenvalue instead.

For CGN, we use the condition number bound for the error in iteration m,

$$\frac{\|e_m\|_2}{\|e_0\|_2} \leq 2\left(\frac{\sqrt{K(B^TB)}-1}{\sqrt{K(B^TB)}+1}\right)^m \tag{6}$$

Task 7

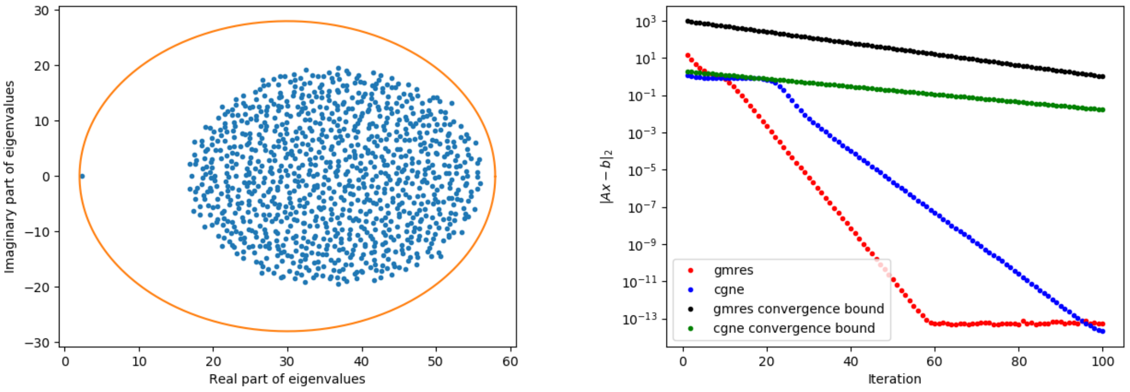
(a)

Given that $A = V^{-1}\Lambda V$ we want to show $A^k = A = V^{-1}\Lambda^k V$, which is done by induction. The initial stage is

$$A^2 = V^{-1}\Lambda V V^{-1}\Lambda V = V^{-1}\Lambda^2 V. \tag{7}$$

Assume $A^k = V^{-1}\Lambda^k V$ for some nonzero k, then

$$A^{k+1} = (V^{-1}\Lambda V)^k (V^{-1}\Lambda V) = V^{-1}\Lambda^k V\, V^{-1}\Lambda V = V^{-1}\Lambda^{k+1} V. \tag{8}$$



(a) Eigenvalues of the matrix B visualised together with disc of convergence used for gmres convergence estimate.

(b) Convergence for gmres and CGN along with estimates of their convergence factors.

Figure 8

Thus $A^k = V^{-1}\Lambda^k V$. A simple consequence is that for $p \in P_n^0$ one has

$$p(A) = \sum_{k=1}^{n} a_k A^k = \sum_{k=1}^{n} a_k V^{-1}\Lambda^{k+1} V = V^{-1}\left(\sum_{k=1}^{n} a_k \Lambda^{k+1}\right) V = V^{-1} p\left(\Lambda^{k+1}\right) V \tag{9}$$

with $a_0 = 1$ for $p \in P_n^0$ and knowing that $A^0 = I$. We have

$$\min_{p\in P_n^0} \|p(A)\| \leq \|V\|\|V^{-1}\| \min_{p\in P_n^0} \|p(\Lambda)\| \tag{10}$$

as a consequence of norms being submultiplicative.

(b)

First we show by induction that

$$\begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}^k = \begin{pmatrix} \lambda_1^k & k\lambda_1^{k-1} \\ 0 & \lambda_1^k \end{pmatrix}. \tag{11}$$

The initial stage is for $k = 2$:

$$\begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}^2 = \begin{pmatrix} \lambda_1^2 & 2\lambda_1 \\ 0 & \lambda_1^2 \end{pmatrix}. \tag{12}$$

Assume (11) holds for k, then

$$\begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}^{k+1} = \begin{pmatrix} \lambda_1^k & k\lambda_1^{k-1} \\ 0 & \lambda_1^k \end{pmatrix} \begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix} = \begin{pmatrix} \lambda_1^{k+1} & (k+1)\lambda_1^k \\ 0 & \lambda_1^{k+1} \end{pmatrix}, \tag{13}$$

i.e. the proposition (11) holds for all nonzero k.

Introduce the monomial $p_k(z) = z^k$, then

$$p_k\left(\begin{pmatrix} \lambda_1 & 1 \\ 0 & \lambda_1 \end{pmatrix}\right) = \begin{pmatrix} \lambda_1^k & k\lambda_1^{k-1} \\ 0 & \lambda_1^k \end{pmatrix} = \begin{pmatrix} p_k(\lambda_1) & p_k'(\lambda_1) \\ 0 & p_k(\lambda_1) \end{pmatrix} \tag{14}$$

which holds for all nonzero k from the induction proof above. We now have

$$p\left(\begin{pmatrix}\lambda_1 & 1\\ 0 & \lambda_1\end{pmatrix}\right)=\sum_{k=1}^{n}a_kp_k\left(\begin{pmatrix}\lambda_1 & 1\\ 0 & \lambda_1\end{pmatrix}\right)=\sum_{k=1}^{n}\begin{pmatrix}a_kp_k(\lambda_1) & a_kp_k'(\lambda_1)\\ 0 & a_kp_k(\lambda_1)\end{pmatrix}=\begin{pmatrix}p(\lambda_1) & p'(\lambda_1)\\ 0 & p(\lambda_1)\end{pmatrix}. \tag{15}$$

(c)

Let A be a block diagonal matrix, such that

$$A=\begin{pmatrix}A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_m\end{pmatrix} \tag{16}$$

where A_i are Jordan block matrices. Due to the block structure we have

$$p(A)=\begin{pmatrix}p(A_1) & & & \\ & p(A_2) & & \\ & & \ddots & \\ & & & p(A_m)\end{pmatrix}. \tag{17}$$

Each block $p(A_i)$ has a singular value decomposition $p(A_i)=U_iS_iV_H^*$, where U_i and V_i are unitary matrices. S_i is a diagonal matrix with the singular values σ as elements. We can now write $p(A)$ as follows.

$$p(A)=\underbrace{\begin{pmatrix}U_1 & & & \\ & U_2 & & \\ & & \ddots & \\ & & & U_m\end{pmatrix}}_{U:=}\underbrace{\begin{pmatrix}S_1 & & & \\ & S_2 & & \\ & & \ddots & \\ & & & S_m\end{pmatrix}}_{S:=}\underbrace{\begin{pmatrix}V_1^H & & & \\ & V_2^H & & \\ & & \ddots & \\ & & & V_m^H\end{pmatrix}}_{V^H:=} \tag{18}$$

due to the rules of multiplication for block diagonal matrices. The final result follows from the definition of the operator norm $\|\cdot\|_2$:

$$\begin{aligned}\|p(A)\|_2 &= \sigma_{\max}(p(A))=\max S=\max_{i=1,\dots,m}(\max S_i))\\ &=\max_{i=1,\dots,m}(\sigma_{\max}(p(A_i)))=\max_{i=1,\dots,m}(\|p(A_i)\|_2)\\ &=\max\left(\left\|\begin{pmatrix}p(\lambda_1) & p'(\lambda_1)\\ 0 & p(\lambda_1)\end{pmatrix}\right\|_2, |p(\lambda_3)|,\dots,|p(\lambda_m)|\right)\end{aligned}$$

(d)

It is clear that

$$p(z)=(\alpha_n+\beta_nz)\frac{(c-z)^{n-1}}{c^{n-1}} \tag{19}$$

satisfies $p\in P_n$. This immediately gives $\alpha_n=1$. We now study

$$p'(z)=\frac{c\left(1-\frac{z}{c}\right)^n(\alpha_n-\alpha_nn+\beta_n(c-nz))}{(c-z)^2}=\frac{c\left(1-\frac{z}{c}\right)^n(1-n+\beta_n(c-nz))}{(c-z)^2}. \tag{20}$$

Thus

$$p'(\lambda_1)=0\Leftrightarrow\frac{c\left(1-\frac{\lambda_1}{c}\right)^n(1-n+\beta_n(c-n\lambda_1))}{(c-\lambda_1)^2}=0\Leftrightarrow(1-n+\beta_n(c-n\lambda_1))=0, \tag{21}$$

that is

$$\beta_n = \frac{n-1}{c-n\lambda_1}. \tag{22}$$

In turn this assumes that $c \neq n\lambda_1$ for $n > 1$.

(e)

Assuming x_n is the n:th iterate generated by gmres, we have by lemma 2.1.3 from that lecture notes that

$$\begin{aligned}
&\|Ax_n - b\|_2 = \min_{x\in\mathcal{K}n(A,b)} \|Ax-b\|_2 = \min_{p\in P_n^0} \|p(A)b\| \leq \|V\|\|V^{-1}\| \min_{p\in P_n^0} \|p(\Lambda)\|\|b\| \\
\Leftrightarrow &\frac{\|Ax_n - b\|_2}{\|b\|} \leq \|V\|\|V^{-1}\| \min_{p\in P_n^0} \|p(\Lambda)\|
\end{aligned}$$

due to the result in 7 (a). Let

$$q(z) = \left(1 + z\frac{n-1}{c-n\lambda_1}\right)\frac{(c-z)^{n-1}}{c^{n-1}}. \tag{23}$$

from the previous task, which by construction is an element of P_n^0. Thus

$$\min_{p\in P_n^0} \|p(\Lambda)\| \leq \|q(\Lambda)\| = \max\left(\left\|\begin{pmatrix} q(\lambda_1) & q'(\lambda_1) \\ 0 & q(\lambda_1)\end{pmatrix}\right\|_2, |q(\lambda_3)|, \ldots, |q(\lambda_m)|\right). \tag{24}$$

Recall that $q'(\lambda_1) = 0$ and that the matrix 2-norm of diagonal matrix is the largest element in modulus. The expression above can be simplified as

$$\min_{p\in P_n^0} \|p(\Lambda)\| \leq \max\left(|q(\lambda_1)|, |q(\lambda_3)|, \ldots, |q(\lambda_m)|\right) = \max_{\lambda_i}\left(1 + \lambda_i\frac{n-1}{c-n\lambda_1}\right)\frac{(c-\lambda_i)^{n-1}}{c^{n-1}}. \tag{25}$$

It is given that all eigenvalues are contained in the disc centered at c with radius ρ. By taking the modulus the inequality (25) and assuming $\lambda_1 \neq 0$ we get

$$\min_{p\in P_n^0} \|p(\Lambda)\| \leq \max_{\lambda_i}\left|1 + \lambda_i\frac{n-1}{c-n\lambda_1}\right|\frac{\rho^{n-1}}{|c^{n-1}|} \leq \max_{\lambda_i}\frac{\overbrace{|c-\lambda_i|}^{\leq\rho} + n\overbrace{|\lambda_i - \lambda_1|}^{\leq 2\rho}}{|c-n\lambda_1|}\frac{\rho^{n-1}}{|c^{n-1}|} \leq \gamma_n\frac{\rho^n}{|c^n|}$$

with

$$\gamma_n = \frac{\frac{1}{n}+2}{\left||\frac{1}{n}| - |\frac{\lambda_1}{c}|\right|}. \tag{26}$$

We already claimed that $c \neq n\lambda_1$, thus the denominator is nonzero for all n. In the limit we have

$$\lim_{n\to\infty} \gamma_n = 2\frac{|c|}{|\lambda_1|} \tag{27}$$

which is bounded. Combining all the results above gives

$$\frac{\|Ax_n - b\|_2}{\|b\|} \leq \|V\|\|V^{-1}\|\gamma_n\frac{\rho^n}{|c^n|}.$$

If $\lambda_1 = 0$ then $\beta_n = (n-1)/c$ and the corresponding bound for (25) is

$$\min_{p\in P_n^0} \|p(\Lambda)\| \leq \max_{\lambda_i}\left|1 + \lambda_i\frac{n-1}{c}\right|\frac{\rho^{n-1}}{|c^{n-1}|} \leq \max_{\lambda_i}\left(\overbrace{|c-\lambda_i|}^{\leq\rho} + n|\lambda_i|\right)\frac{\rho^{n-1}}{|c^n|} \leq \gamma_n\frac{\rho^n}{|c^n|}.$$

However, now

$$\gamma_n = \max_{\lambda_i} \left(1 + n\frac{|\lambda_i|}{\rho}\right), \tag{28}$$

which is not a bounded sequence.

(f)

For nonzero λ_1 we have convergence, but the speed is influenced by γ_n. Roughly, the further the centre c is from λ_1 the better. For many iterations we approximately get

$$\gamma_n \frac{\rho^n}{|c^n|} \approx \frac{2\rho}{|\lambda_1|} \frac{\rho^{n-1}}{|c^{n-1}|}. \tag{29}$$

Thus the rate of convergence is the same, but the factor $\frac{2\rho}{|\lambda_1|}$ may be large. So if the double eigenvalues lie close to zero and the other eigenvalues lie far away from the origin, then the factor will be large.

For $\lambda_1 = 0$ the sequence γ_n is not bounded. Note that this does not mean that GMRES will diverge, only that the estimate gives no information.

(e)

We discussed with Aku Kammonen and Parikshit Upadhyaya.