Project 1 in TMA4205

723818, 724041

September 28, 2014

a)

Discretization of the differential equation

$$-U_{xx} + aU_x = f \tag{1}$$

with boundary conditions

$$u(0) = U_0 = 1, u(1) = U_n = -1 \tag{2}$$

using difference methods. Where $a = a(x)$ and $f = f(x)$, $h = 1/N$. Approximation for second order term:

$$U_{xx} = \frac{u_{j+1} - 2u_j + u_{j-1}}{h^2}$$

Approximations for first order term:

$$U_x = \frac{u_{j+1} - u_{j-1}}{2h}$$

$$U_x = \frac{u_j - u_{j-1}}{h}$$

$$U_x = \frac{u_{j+1} - u_j}{h}$$

The equation can now be written as

$$AU = b$$

Using the notation $a(x_j) = a_j$ and $f(x_j) = f_j$, the goal is now want to find

$$A = \begin{pmatrix} \alpha_1 & \delta_1 & & & \\ \gamma_2 & \alpha_2 & \delta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \gamma_{n-2} & \alpha_{n-2} & \delta_{n-2} \\ & & & \gamma_{n-1} & \alpha_{n-1} \end{pmatrix}, b = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{j-2} \\ \beta_{j-1} \end{pmatrix}$$

Observing that $b_j = f_j + \tau_j$, τ nonzero only for $j = 1$ and $j = n - 1$.
for each of the different discretizations of U_x.

Central difference

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + a_j \frac{u_{j+1} - u_{j-1}}{2h} = u_{j+1}\left(\frac{1}{h^2} + a_j \frac{1}{2h}\right) + u_j\left(-\frac{2}{h^2}\right) + u_{j-1}\left(\frac{1}{h^2} - a_j \frac{1}{2h}\right)$$

And so

$$\gamma_j = \frac{1}{h^2} - a_j \frac{1}{2h}, \alpha_j = -\frac{2}{h^2}, \delta_j = \frac{1}{h^2} + a_j \frac{1}{2h}$$

.
And

$$\tau_i = \begin{cases} -(\frac{1}{h^2} - a_1\frac{1}{2h}) & i = 1 \\ (\frac{1}{h^2} + a_{n-1}\frac{1}{2h}) & i = n-1 \\ 0 & else \end{cases}$$

Backward difference

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + a_j\frac{u_j - u_{j-1}}{h} = u_{j+1}(\frac{1}{h^2}) + u_j(-\frac{2}{h^2} + a_j\frac{1}{h}) + u_{j-1}(\frac{1}{h^2} - a_j\frac{1}{h})$$

So

$$\gamma_j = \frac{1}{h^2} - a_j\frac{1}{h}, \alpha_j = -\frac{2}{h^2} + a_j\frac{1}{h}, \delta_j = \frac{1}{h^2}$$

.
And

$$\tau_i = \begin{cases} -(\frac{1}{h^2} - a_1\frac{1}{h}) & i = 1 \\ (\frac{1}{h^2}) & i = n-1 \\ 0 & else \end{cases}$$

Forward differences

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} + a_j\frac{u_{j+1} - u_j}{h} = u_{j+1}(\frac{1}{h^2} + a_j\frac{1}{h}) + u_j(-\frac{2}{h^2} - a_j\frac{1}{h}) + u_{j-1}(\frac{1}{h^2})$$

So $\gamma_j = \frac{1}{h^2}\alpha_j = -(\frac{2}{h^2} + a_j\frac{1}{h})$, $\delta_j = \frac{1}{h^2} + a_j\frac{1}{h}$.
And

$$\tau_i = \begin{cases} -(\frac{1}{h^2}) & i = 1 \\ (\frac{1}{h^2} + a_{n-1}\frac{1}{h}) & i = n-1 \\ 0 & else \end{cases}$$

b) A should be irreducable diagonaly dominant. A diagonal dominant matrix has the following property:

$$|a_{ii}| \geq \sum_{j\neq} |a_{ij}|, \forall i.$$

Central differences

$$|\gamma_j| + |\delta_j| = \frac{|(ah-2)|+|-(ah+2)|}{2h^2} = \begin{cases} \frac{a_j}{h}, & a_j > \frac{2}{h} \\ \frac{2}{h^2}, & \frac{-2}{h} \leq a_j \leq \frac{2}{h} \\ \frac{-a_j}{h}, & a_j < \frac{-2}{h} \end{cases}$$

$\leq \frac{2}{h^2} = |\alpha_j|$ as long as $\frac{-2}{h^2} \leq a_j \leq \frac{2}{h^2}$.

Backwards differences

$$|\gamma_j| + |\delta_j| = |-\frac{1}{h^2} - \frac{a_j}{h}| + |\frac{1}{h^2}| = \frac{1}{h^2}|ha_j + 1| + \frac{1}{h^2} = \begin{cases} \frac{ha_j+2}{h^2}, & a_j > -\frac{1}{h} \\ -\frac{a}{h}, & a_j \leq -\frac{1}{h} \end{cases}$$

$$\leq \frac{1}{h^2}|2 + ha_j| = \begin{cases} \frac{ha_j+2}{h^2}, & a_j > -\frac{2}{h} \\ -\frac{ha_j+2}{h^2}, & a_j \leq -\frac{2}{h} \end{cases} = |\alpha_j| \text{ as long as } a_j \geq -\frac{1}{h}.$$

Forward differences

$$|\gamma_j| + |\delta_j| = \frac{1}{h^2} + \frac{1}{h^2}|ha_j - 1| = \begin{cases} \frac{a_j}{h}, & a_j > \frac{1}{h} \\ \frac{2-ha_j}{h^2}, & a_j \leq \frac{1}{h} \end{cases}$$

$$\leq \frac{1}{h^2}|2 - ha_j| = \begin{cases} \frac{ha_j-2}{h^2}, & a_j > \frac{2}{h} \\ \frac{2-ha_j}{h^2}, & a_j \leq \frac{2}{h} \end{cases} = |\alpha_j| \text{ as long as } a_j \leq \frac{1}{h}.$$

We see that this method will be diagonally dominant for all the methods as long as $-\frac{1}{h} \leq a_j \leq \frac{1}{h}$, which for small steps h gives us a lot of room to choose the $a_j's$.

If we make an adjacency matrix out of our A matrix, (that is replace the α_j, γ_j and δ_j with 1). The directed graph produced by this new matrix can reach any other point, since you can move from every single point to its two neighbours (along the x axis) and itself, (as long as you are not on the endpoints, then you can only move to the single neighbour of the point and itself). Such an a directed graph is strongly connected, as a matrix is irreducible if the directed graph of its adjacency matrix is strongly connected. Therefore A is irreducible.

Theorem 4.9 in Saad holds for our matrix A.

c) Assuming now that a is independent of x.The eigenvalues of such a tri-diagonal matrix have a general form

$$\lambda_m = \alpha + 2\sqrt{\delta\gamma}\cos(\frac{m\pi}{n+1})$$

For the estimation of U_x using backward differences, the eigenvalues for A will be

$$\lambda_m = \frac{2h-2}{h^2} + \frac{2}{h^2}\sqrt{1-2h}\cos(\frac{m\pi}{n+1})$$

d) The Eigenvalues of $G_J = D^{-1}(D - A)$, where D is the diagonal of A can be found by som calculations.

$$D^{-1}(D-A) = \begin{pmatrix} \alpha^{-1} & 0 & \cdots & & \\ 0 & \alpha^{-1} & 0 & \cdots & \\ 0 & 0 & \alpha^{-1} & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} 0 & -\delta & 0 & \cdots & & \\ -\gamma & 0 & -\delta & 0 & \cdots & \\ 0 & -\gamma & 0 & -\delta & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \end{pmatrix}$$

$$= \begin{pmatrix} 0 & -\delta\alpha^{-1} & 0 & \cdots & & \\ -\gamma\alpha^{-1} & 0 & -\delta\alpha^{-1} & 0 & \cdots & \\ 0 & -\gamma\alpha^{-1} & 0 & -\delta\alpha^{-1} & 0 & \cdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \end{pmatrix}$$

This result and the formula from part c) gives

$$\lambda_m = 2\sqrt{\gamma\delta\alpha^{-2}}\cos(\frac{m\pi}{n+1}), m = 1, \cdots, n$$

Substituting for U_x from equation (5) gives

$$\lambda_m = \sqrt{\frac{1-2h}{(h-1)^2}}\cos(\frac{m\pi}{n+1})$$

The spectral radius of G_J is

$$\rho(G_J) = \max_m(|\lambda_m|) = \sqrt{\frac{1-2h}{(h-1)^2}}\cos(\frac{\pi}{n+1}) < 1$$

Gershgorin's theorem states that every eigenvalue of a matrix will lie inside at least one Gershgorin disc, $D(a_{ii}, R_i)$, where $R_i = \sum_{i \neq j} a_{ij}$. In this case $a_{ii} = 0$ for all i, and $R_i \leq \alpha^{-1}(\gamma + \delta) = 1$ for all i, so all eigenvalues should be inside $D(0, 1)$. This clearly holds since $\rho(G_J) \to 1$, as $n \to \infty$ and $m \to 0$. For the first and the last row, the disc will be slightly smaller.

e) Error estimate

$$
\begin{aligned}
e^{(k)} &= u - u^{(k)} \\
&= u - G_J u^{(k-1)} + D^{-1}A, \qquad\qquad\qquad\qquad Au = b \\
&= u - G_J u^{(k-1)} - D^{-1}Au \\
&= (I - D^{-1}A)u - G_J u^{(k-1)}, \qquad\qquad G_J = D^{-1}(D-A) \\
&= G_J(u - u^{(k-1)}) \\
\vdots &= \vdots \\
e^{(k)} &= G_J^k(u - u^{(0)}) = G_J^k e^{(0)}
\end{aligned}
$$

Observing that G_J has $n-1$ distinct eigenvalues, we get that we can diagonalize the matrix into $G_J = P\Lambda P^{-1}$. We get

$$\|e^{(k)}\|_2 \leq \|P\|_2 \cdot \|\Lambda\|_2^k \cdot \|P^{-1}\|_2 \cdot \|e^{(0)}\|_2 \tag{3}$$

where $\|\Lambda\|_2 = \max|\lambda| = \sqrt{\frac{1-2h}{(h-1)^2}}\cos(\frac{\pi}{n+1}) = \sqrt{\frac{n(n-2)}{(n-1)^2}}\cos(\frac{\pi}{n+1})$. Further by assuming equation(3) is an equality we take the logarithm and get the following result:

$$\log\frac{\|e^{(k)}\|_2}{\|P\|_2 \cdot \|P^{-1}\|_2 \cdot \|e^{(0)}\|_2} = k\log(\sqrt{\frac{n(n-2)}{(n-1)^2}}\cos(\frac{\pi}{n+1})) =$$

$$k[\frac{1}{2}\log n + \frac{1}{2}(n-2) - \log(n-1) + \log(\cos(\frac{\pi}{n+1}))] \approx k[\log(\cos(\frac{\pi}{n+1}))] \approx$$

$$k[\log(1 - \frac{\pi^2}{2(n+1)^2})] \approx -k\frac{\pi^2}{2(n+1)^2}$$

We now have the formula

$$k \approx \frac{2(n+1)^2}{\pi^2}\log\frac{\|P\|_2 \cdot \|P^{-1}\|_2 \cdot \|e^{(0)}\|_2}{\|e^{(k)}\|_2}$$

Several of the approximations requires n to be of a certain size, say $n \geq 10$. Now we would like to see how k behaves when we double n, assuming that

$$\log\frac{\|e_n^{(0)}\|_2 \cdot \|P_n^{-1}\|_2 \cdot \|P_n\|_2}{\|e_n^{(k_n)}\|_2} = \log\frac{\|e_{2n}^{(0)}\|_2 \cdot \|P_{2n}^{-1}\|_2 \cdot \|P_{2n}\|_2}{\|e_{2n}^{(k_{2n})}\|_2}$$

$$\frac{k_{2n}}{k_n} = \frac{\frac{2(2n+1)^2}{\pi^2}}{\frac{2(n+1)^2}{\pi^2}} = \frac{(2n+1)^2}{(n+1)^2} \approx 4$$

f) If U has the value $U = \cos(\pi x)$ Then f can be calculated

$$f = -U_{xx} + 2U_x = \pi^2 \cos(\pi x) - 2\pi \sin(\pi x)$$

Let now U_* be the vector $U(x_j)$ of the real solution, defined only in the discrete points x_j. The error in the discrete points are then

$$e_*^{(k)} = U_* - u^{(k)} = U_* - U + U - u^{(k)} = U - u^{(k)}$$

In the figure below the logarithm of the error, $\log e_*^{(k)}$, is plottet against the number of iterations k, for different discretization points n.

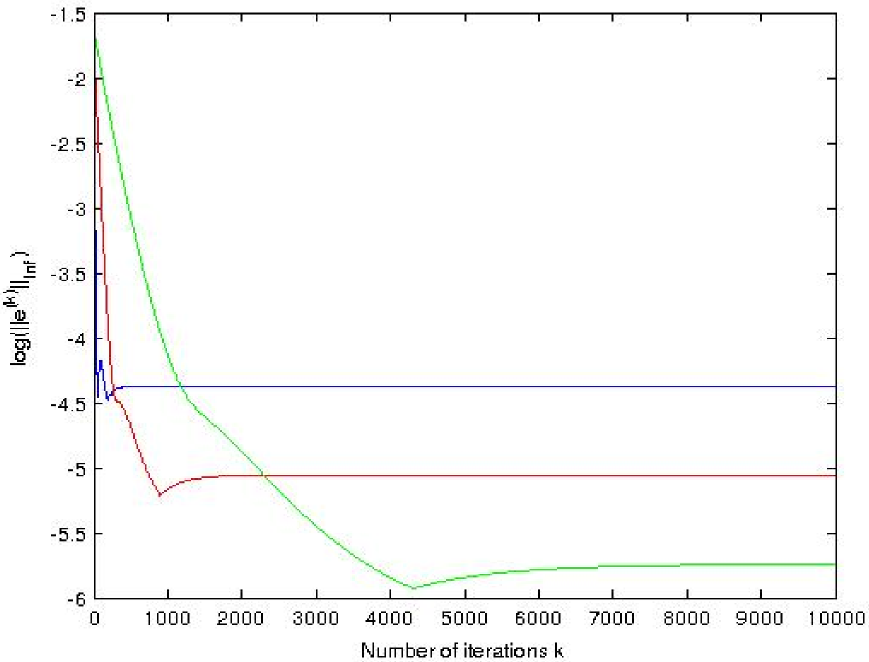

Figure 1: Blue: $n = 20$, red: $n = 40$, green: $n = 80$.

As the figure shows, convergence is slower with higher n, this phenomenon is explained in e), and also shows that the method stabilizes approximately for 4 times greater k when you double n. As known from difference methods, the approximate solution will converge slowly toward the true solution when $h \to 0$, that is why the error is smaller for higher n. Another phenomenon worth noticing is how the error is not strictly decreasing. That is because the error is taken from the difference between the real solution U and the iterated solution $u^{(k)}$, and not the solution to $A^{-1}b$ as is being approximated by iterative methods. This means that the iterative solution closes inn on the real solution, because the real solution is between our guess, $u^{(0)}$, and $A^{-1}b$.

g) In the plot below, Jacobi, backwards and forwards Gauss-Seidel iterations are compared to each other. Each of the methods have the logarithm of the error, $\log e_*^{(0)}$, plotted against the number of iterations k, for $n = 40$.

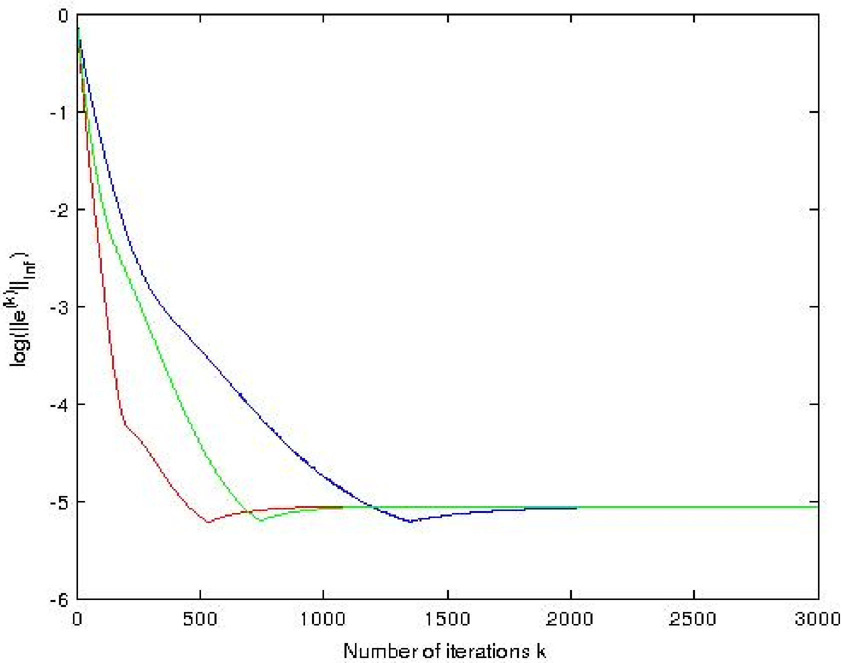

Figure 2: Blue: Jacobi, red: backward Gauss Seidel, green: forward Gauss-Seidel.

It seems that backwards Gauss-Seidel outpreforms the forward Gauss-Seidel. My guess would be that since backwards Gauss-Seidel uses more information per iteration, it converges faster than forwards Gauss-Seidel.

h) The eigenvalues for the $G_{\omega J}$ are as follows:

$$G_{\omega J} \cdot x = [I - \omega D^{-1} A] \cdot x = (1 - \frac{\omega \lambda_m}{\alpha}) \cdot x = (1 - \omega(1 - \frac{2\sqrt{\gamma\delta}}{\alpha} \cos(\frac{m\pi}{n+1}))) \cdot x$$

With backward euler and $n = \frac{1}{h}$ turns out as

$$G_{\omega J} \cdot x = (1 - \omega(1 - \sqrt{\frac{n(n-2)}{(n-1)^2}} \cos(\frac{m\pi}{n+1}))) \cdot x = \lambda_{\omega J} \cdot x$$

Here x is an eigenvector for A, since $D^{-1} = \frac{1}{\alpha} I$ this is also a eigenvector of D^{-1} and therefore of the entire $G_{\omega J}$ matrix. We see that the vector is also diagonizable. Obserbing $\lambda_{\omega J}$ we see that $0 < (1 - \sqrt{\frac{n(n-2)}{(n-1)^2}} \cos \frac{m\pi}{n+1}) < 2$, which for big n can get really close to the upper limit if $m = n$, and really close to the lower limit if $m = 1$. We can divide $\max |\lambda_{\omega J}|$ into two parts one of which $0 \le \omega \le 1$ where

$$\max \|\lambda_{\omega J}\| = 1 - \omega(1 - \sqrt{\frac{n(n-2)}{(n-1)^2}} \cos \frac{\pi}{n+1})$$

is the maximum of the absolute value of something positive, and another $1 \le \omega \le 2$ where

$$\max \|\lambda_{\omega J}\| = \max \|1 - \omega(1 - \sqrt{\frac{n(n-2)}{(n-1)^2}} \cos \frac{n\pi}{n+1})\| = \omega(1 + \sqrt{\frac{n(n-2)}{(n-1)^2}} \cos \frac{\pi}{n+1}) - 1$$

is the maximum of the absolute value of something negative.
For $0 \le \omega \le 1$, we have that $\max |\lambda_{\omega J}|$ is smallest for $\omega = 1$, which is also true for $1 \le \omega \le 2$. Here we see that $\rho(G_{\omega J})$ is minimized for $\omega = 1$ where

$$\lambda_{\max}(G_{\omega J}) = -\lambda_{\min}(G_{\omega J})$$

We want to find an optimal ω by finding the smallest $\|e^{(k)}\|$ possible:

$$\min \|e^{(k)}\|_2 = \min \|b - Au^{(k)}\|_2 = \min \|b - Au^{(k-1)} + \omega A D^{-1} A u^{(k-1)} - \omega A D^{-1} b\|_2$$
$$= \min \|(I - \omega A D^{-1})(b - Au^{(k-1)})\|_2 = \min \|G_{\omega J} e^{(k-1)}\|$$

We thereby get that

$$\min \|e^{(k)}\|_2 \le \min \|G_{\omega J}\|_2^k \|e^{(0)}\|_2$$

since $e^{(0)}$ is chosen, we basicly look for the smallest $G_{\omega J}$. As the matrix is diagonizable $G_{\omega J} = P \Lambda_{\omega J} P^{-1}$, and since the eigenvectors are the same for A and $G_{\omega J}$, we get P and P^{-1} independant of ω. The minimum of $\|G_{\omega J}\|$ is found when $\|\Lambda_{\omega J}\|_2 = \max |\lambda_{\omega J}|$ is at its smallest, which is the case for $\omega = 1$. In other words, the relaxation does not change anything!!!