Andrew Cameron

CS 6220

Homework 1 due W 9/24/2008

Problem P10.1.2

Let \( G = \begin{pmatrix} \lambda_1 & m \\ 0 & \lambda_2 \end{pmatrix} \) have \( \rho(G) < 1 \). Obviously the eigenvalues of \( G \) are \( \lambda_1 \) and \( \lambda_2 \).

After writing out the first few \( G^k \), the pattern appeared to be that \( G^k = \begin{pmatrix} \lambda_1^k & m \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \\ 0 & \lambda_2^k \end{pmatrix} \). This formula is obviously right for \( G^1 \). Assuming it is right for some integer \( k \geq 1 \), observe that

\[
G^{k+1} = G^k G = \begin{pmatrix} \lambda_1^k & m \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \\ 0 & \lambda_2^k \end{pmatrix} \begin{pmatrix} \lambda_1 & m \\ 0 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \lambda_1^{k+1} & m \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \lambda_2 \\ 0 & m \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \lambda_2 \end{pmatrix},
\]

so the formula gets \( G^{k+1} \) right. By induction, the formula is valid for all integers \( k \geq 1 \).

The square of the Frobenius norm of \( G^k \) is

\[
\|G^k\|_F^2 = |\lambda_1|^{2k} + |\lambda_2|^{2k} + \left| m \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \right|^2.
\]

Since \( |\lambda_1|, |\lambda_2| \leq \rho(G) < 1 \), certainly \( \lim_{k \to \infty} |\lambda_1|^{2k} = \lim_{k \to \infty} |\lambda_2|^{2k} = 0 \). To handle the third term, notice that

\[
\left| m \sum_{j=0}^{k-1} \lambda_1^j \lambda_2^{k-1-j} \right| \leq |m| \sum_{j=0}^{k-1} |\lambda_1|^j |\lambda_2|^{k-1-j} \leq |m| |k| \rho(G)^{k-1}.
\]

Writing \( k(\rho(G))^{k-1} = \frac{1}{\exp((k-1) \log \rho(G))} \), we see by L’Hopital’s rule that

\[
\lim_{k \to \infty} k(\rho(G))^{k-1} = \lim_{k \to \infty} \frac{1}{\exp((k-1) \log \rho(G))} \log \frac{1}{\rho(G)} = -\frac{1}{\log \rho(G)} \lim_{k \to \infty} (\rho(G))^{k-1}.
\]

Since \( \rho(G) < 1 \), we conclude that \( \lim_{k \to \infty} k(\rho(G))^{k-1} = 0 \).

Having shown that all three terms in the expression for \( \|G^k\|_F^2 \) tend to 0 as \( k \to \infty \), we conclude that \( \lim_{k \to \infty} \|G^k\|_F = 0 \), which implies that \( \lim_{k \to \infty} G^k = 0 \).

Problem P10.1.3

Let \( A \in \mathbb{R}^{2 \times 2} \) be positive definite and symmetric. Since \( A \) is symmetric, we can write it as \( A = \begin{pmatrix} a & b \\ b & d \end{pmatrix} \).

Since \( A \) is positive-definite, we know that, if \( (x, y) \neq (0, 0) \) then

\[
0 < \begin{pmatrix} x \\ y \end{pmatrix}^T \begin{pmatrix} a & b \\ b & d \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = ax^2 + 2bxy + dy^2.
\]
Taking \((x,y) = (1,0)\), we conclude that \(a > 0\). Taking \((x,y) = (0,1)\), we conclude that \(d > 0\). Taking \((x,y) = (\pm \sqrt{\alpha}, \sqrt{\alpha})\), we conclude that \(2\sqrt{ad} (\sqrt{ad} \pm b) > 0\), which implies that \(\sqrt{ad} \pm b > 0\). This means that \(\sqrt{ad} > |b|\).

The Jacobi method takes
\[
M_J = D = \begin{pmatrix} a & 0 \\ 0 & d \end{pmatrix}
\quad \text{and} \quad
N_J = -(L + U) = \begin{pmatrix} 0 & -b \\ -b & 0 \end{pmatrix},
\]
so the iteration matrix is
\[
G_J = M_J^{-1}N_J = \begin{pmatrix} 1/a & 0 \\ 0 & 1/d \end{pmatrix} \begin{pmatrix} 0 & -b \\ -b & 0 \end{pmatrix} = \begin{pmatrix} 0 & -b/a \\ -b/d & 0 \end{pmatrix}.
\]
This has characteristic polynomial
\[
\det(\lambda I - G_J) = \lambda^2 - \frac{b^2}{ad}.
\]
Therefore the eigenvalues of \(G_J\) are \(\pm \frac{b}{\sqrt{ad}}\), so \(\rho(G_J) = \frac{|b|}{\sqrt{ad}} < 1\), so the Jacobi method converges.

**Problem P10.1.5(1)**

Let \(A_1 = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 1 \end{pmatrix}\) and \(A_2 = \begin{pmatrix} 1 & -3/4 \\ -1/12 & 1 \end{pmatrix}\). The associated Jacobi iteration matrices are
\[
J_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 1/2 \\ 1/2 & 0 \end{pmatrix}
\]
and
\[
J_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 3/4 \\ 1/12 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 3/4 \\ 1/12 & 0 \end{pmatrix},
\]
which have characteristic polynomials \(\lambda^2 - 1/4\) and \(\lambda^2 - 1/16\), respectively. Therefore, \(\rho(J_1) = 1/2\) and \(\rho(J_2) = 1/4\), so \(\rho(J_1) > \rho(J_2)\). This means that the Jacobi method converges faster for \(A_2\) than it does for \(A_1\).

In Section 10.1.4, it is shown that diagonal dominance implies convergence of the Jacobi method. Both \(A_1\) and \(A_2\) are diagonally dominant. To me it is not clear that \(A_1\) is “more diagonally dominant” than \(A_2\). Certainly the first row of \(A_1\) is more diagonally dominant than the first row of \(A_2\), but the second row of \(A_2\) is more diagonally dominant than the second row of \(A_1\). What does it mean for a matrix to be “more diagonally dominant” than another matrix?

**Problem P10.1.5(2)**

If the grid point spacing in the horizontal direction is \(h_x\), the second derivative in the horizontal direction of \(u\) at \(P\) is approximated by
\[
\frac{u(E) - u(P) - u(P) - u(W)}{h_x} = \frac{1}{h_x^2} (u(E) + u(W) - 2u(P)).
\]

If the grid point spacing in the vertical direction is \(h_y\), the second derivative in the vertical direction of \(u\) at \(P\) is approximated by
\[
\frac{u(N) - u(P) - u(P) - u(S)}{h_y} = \frac{1}{h_y^2} (u(N) + u(S) - 2u(P)).
\]

If the sum of the approximate second derivatives in the horizontal and vertical directions is to be zero, we have, upon scaling, the equations
\[
4u(P) - \frac{2}{1 + h_x^2} (u(E) + u(W)) - \frac{2}{1 + h_y^2} (u(N) + u(S)) = 0.
\]
As always, we think of ordering the points by looping through the rows and columns, with the inner loop corresponding to going from left to right horizontally across the rows, and the outer loop corresponding to going from the top to the bottom vertically down the columns. The rows each have \(N\) points and the columns each have \(M\) points. Therefore, the index of \(E\) is 1 more than that of \(P\), the index of \(W\) is 1 less than that of \(P\), the index of \(N\) is \(N\) less than that of \(P\), and the index of \(S\) is \(N\) more than that of \(P\). If the index of \(P\) is a multiple of \(N\), it has no neighbour \(E\), if the index of \(P\) is one more than a multiple of \(N\), it has no neighbour \(W\), if the index of \(P\) is less than or equal to \(N\) it has no neighbour \(N\), and if the index of \(P\) is greater than \(N(M - 1)\), it has no neighbour \(S\).

The contribution to the equation above from \(P\) corresponds to \(4I_{MN}\), the contribution from \(E\) and \(W\) corresponds to \(-\frac{2}{1 + \frac{h^2}{\pi^2}} (I_M \otimes E_N)\), and the contribution from \(N\) and \(S\) corresponds to \(-\frac{2}{1 + \frac{h^2}{\pi^2}} (E_M \otimes I_N)\). Therefore,

\[
A = 4I_{MN} - \frac{2}{1 + \frac{h^2}{\pi^2}} (I_M \otimes E_N) - \frac{2}{1 + \frac{h^2}{\pi^2}} (E_M \otimes I_N).
\]

With \(S_m\) from Equation 10.1.12 and \(D_m = \text{diag}(\mu_{m,1}, \ldots, \mu_{m,m})\), where \(\mu_{m,k} = 2 \cos\left(\frac{k\pi}{M + 1}\right)\), we know from Section 10.1.7 that \(S_m^T E_m S_m = D_m\). Using this, along with the facts about Kronecker products in Section 10.1.6, we can diagonalise \(A\).

\[
(S_M \otimes S_N)^T A(S_M \otimes S_N) = 4I_{MN} - \frac{2}{1 + \frac{h^2}{\pi^2}} (I_M \otimes S_N^T E_N S_N) - \frac{2}{1 + \frac{h^2}{\pi^2}} (S_M^T E_M S_M \otimes I_N)
\]

\[
= 4I_{MN} - \frac{2}{1 + \frac{h^2}{\pi^2}} (I_M \otimes D_N) - \frac{2}{1 + \frac{h^2}{\pi^2}} (D_M \otimes I_N),
\]

which is diagonal. Therefore the eigenvalues of \(A\) are

\[
4 \left(1 - \frac{\cos\left(\frac{q\pi}{N + 1}\right)}{1 + \frac{h^2}{\pi^2}} - \frac{\cos\left(\frac{p\pi}{M + 1}\right)}{1 + \frac{h^2}{\pi^2}}\right),
\]

for \(p = 1 : M\) and \(q = 1 : N\).

In the case that \(h_x = h_y = h\) this reduces to

\[
4 \left(\frac{1}{2} (1 - \cos\left(\frac{p\pi}{M + 1}\right)) + \frac{1}{2} (1 - \cos\left(\frac{q\pi}{N + 1}\right))\right) = 4 \left(\sin^2\left(\frac{p\pi}{2(M + 1)}\right) + \sin^2\left(\frac{q\pi}{2(N + 1)}\right)\right),
\]

which agrees with the equation after Equation 10.1.13.

**Problem P10.1.6**

The second derivative in the horizontal direction of \(u\) at \(P\) is approximated by

\[
\frac{u(E) - u(P) - u(W) - u(P)}{h} = \frac{1}{h^2} (u(E) + u(W) - 2u(P)).
\]

The second derivative in the vertical direction of \(u\) at \(P\) is approximated by

\[
\frac{u(N) - u(P) - u(S) - u(P)}{h} = \frac{1}{h^2} (u(N) + u(S) - 2u(P)).
\]

The second derivative in the third direction of \(u\) at \(P\) is approximated by

\[
\frac{u(H) - u(P) - u(L) - u(P)}{h} = \frac{1}{h^2} (u(H) + u(L) - 2u(P)).
\]
Think of $H$ standing for “higher” and $L$ (bad notation) standing for “lower.” If the sum of the approximate second derivatives in the horizontal, vertical, and third directions is to be zero, we have, upon scaling, the equations

$$6u(P) - u(E) - u(W) - u(N) - u(S) - u(H) - u(L).$$

We think of ordering the points by looping through the rows, columns, and third dimension, with the innermost loop corresponding to going from left to right horizontally across the rows, the middle loop corresponding to going from the top to the bottom vertically down the columns, and the outermost loop corresponding to going from lower to higher in the third dimension. The rows each have $N$ points, the columns each have $M$ points, and the third direction has $P$ (bad notation) points.

Therefore, the index of $E$ is 1 more than that of $P$, the index of $W$ is 1 less than that of $P$, the index of $N$ is $N$ less than that of $P$, the index of $S$ is $N$ more than that of $P$, the index of $H$ is $MN$ more than that of $P$, and the index of $L$ is $MN$ less than that of $P$. If the index of $P$ is a multiple of $N$, it has no neighbour $E$, if the index of $P$ is one more than a multiple of $N$, it has no neighbour $W$, if the index of $P$ is greater than an integer multiple of $MN$ but no greater than that same multiple of $MN$ plus $N$, it has no neighbour $N$, if the index of $P$ is less than or equal to an integer multiple of $MN$ but greater than that same multiple of $MN$ minus $N$, it has no neighbour $S$, if the index of $P$ is less than or equal to $MN$, it has no neighbour $L$, and if the index of $P$ is greater than $MN(P - 1)$, it has no neighbour $H$.

The contribution to the equation above from $P$ corresponds to $6I_{MNP}$, the contribution from $E$ and $W$ corresponds to $-I_{PM} \otimes E_N$, the contribution from $N$ and $S$ corresponds to $-I_P \otimes E_M \otimes I_N$, and the contribution from $L$ and $H$ corresponds to $-E_P \otimes I_{MN}$.

Therefore,

$$A = 6I_{MNP} - I_{PM} \otimes E_N - I_P \otimes E_M \otimes I_N - E_P \otimes I_{MN}.$$ 

We can diagonalise $A$ using the same notation as in Problem P10.1.5(2).

$$(S_P \otimes S_M \otimes S_N)^T A (S_P \otimes S_M \otimes S_N)
= 6I_{MNP} - I_{PM} \otimes (S_N^T E_N S_N) - I_P \otimes (S_M^T E_M S_M) \otimes I_N - (S_P^T E_P S_P) \otimes I_{MN}
= 6I_{MNP} - I_{PM} \otimes D_N - I_P \otimes D_M \otimes I_N - D_P \otimes I_{MN},$$

which is diagonal. Therefore the eigenvalues of $A$ are

$$2 \left( 3 - \cos \left( \frac{q \pi}{N + 1} \right) - \cos \left( \frac{p \pi}{M + 1} \right) - \cos \left( \frac{r \pi}{P + 1} \right) \right)$$

for $p = 1 : M$, $q = 1 : N$, and $r = 1 : P$.

The only term in the above expression for $A$ that has any nonzero entries on the diagonal is $6I_{MNP}$. Therefore, for the Jacobi method, $M_J = 6I_{MNP}$ and

$$N_J = I_{PM} \otimes E_N + I_P \otimes E_M \otimes I_N + E_P \otimes I_{MN},$$

so the iteration matrix is $G_J = M_J^{-1} N_J = \frac{1}{6} N_J$. Observe that

$$(S_P \otimes S_M \otimes S_N)^T G_J (S_P \otimes S_M \otimes S_N) = \frac{1}{6} (I_{PM} \otimes S_N^T E_N S_N + I_P \otimes S_M^T E_M S_M \otimes I_N + S_P^T E_P S_P \otimes I_{MN})$$

$$= \frac{1}{6} (I_{PM} \otimes D_N + I_P \otimes D_M \otimes I_N + D_P \otimes I_{MN}),$$

which is diagonal. Therefore, the eigenvalues of $G_J$ are

$$\frac{1}{3} \left( \cos \left( \frac{q \pi}{N + 1} \right) + \cos \left( \frac{p \pi}{M + 1} \right) + \cos \left( \frac{r \pi}{P + 1} \right) \right)$$

for $q = 1 : N$, $p = 1 : M$, and $r = 1 : P$, so

$$\rho(G_J) = \frac{1}{3} \left( \cos \left( \frac{\pi}{N + 1} \right) + \cos \left( \frac{\pi}{M + 1} \right) + \cos \left( \frac{\pi}{P + 1} \right) \right).$$
This is always less than 1, so the Jacobi method always converges. If $M$, $N$, and $P$ are all large, $\rho(G_J)$ approaches 1.

**Problem P10.1.7**

The setup here is the same as always except we must now deal with $NE$, $NW$, $SW$, and $SE$. The index of $NW$ is $N + 1$ less than that of $P$, the index of $NE$ is $N + 1$ more than that of $P$, and the index of $SE$ is $N + 1$ more than that of $P$. If the index of $P$ is less than or equal to $N$ or one more than a multiple of $N$, it has no neighbour $NW$, if the index of $P$ is less than or equal to $N$ or a multiple of $N$, it has no neighbour $NE$, if the index of $P$ is greater than $N(M - 1)$ or one more than a multiple of $N$, it has no neighbour $SW$, and if the index of $P$ is greater than $N(M - 1)$ or a multiple of $N$, it has no neighbour $SE$.

In the formula

$$20u(P) - 4(u(N) + u(E) + u(S) + u(W)) - (u(NE) + u(SE) + u(SW) + u(NW)) = 0,$$

the contribution from $P$ is $20I_{MN}$. As always, the contribution from $E$ and $W$ is $-4(I_M \otimes E_N)$ and the contribution from $N$ and $S$ is $-(E_M \otimes I_N)$. I am able to convince myself that the contribution from $NE$, $SE$, $SW$, and $NW$ is $-(E_M \otimes E_N)$. This has some nonzeros on the $(N - 1)$st and $(N + 1)$st superdiagonals. These correspond to the contributions from $SW$, $SE$, $NE$, and $NW$, respectively. Therefore,

$$A = 20I_{MN} - 4(I_M \otimes E_N) - 4(E_M \otimes I_N) - (E_M \otimes E_N).$$

Observe that

$$(S_M \otimes S_N)^T A (S_M \otimes S_N)$$

$$= 20I_{MN} - 4(I_M \otimes S_M^T E_N S_N) - 4(S_M^T E_M S_M \otimes I_N) - (S_M^T E_M S_M \otimes S_M^T E_N S_N)$$

$$= 20I_{MN} - 4(I_M \otimes D_N) - 4(D_M \otimes I_N) - (D_M \otimes D_N),$$

which is diagonal. Therefore the eigenvalues of $A$ are

$$20 - 8 \cos\left(\frac{q\pi}{N + 1}\right) - 8 \cos\left(\frac{p\pi}{M + 1}\right) - 4 \cos\left(\frac{q\pi}{N + 1}\right) \cos\left(\frac{p\pi}{M + 1}\right).$$

There are other ways of expressing this.

The only term in the expression for $A$ that has any nonzero entries on the diagonal is $20I_{MN}$. Therefore, for the Jacobi method, $M_J = 20I_{MN}$ and

$$N_J = 4(I_M \otimes E_N) + 4(E_M \otimes I_N) + (E_M \otimes E_N),$$

so the iteration matrix is $G_J = M_J^{-1} N_J = \frac{1}{20} N_J$. Observe that

$$(S_M \otimes S_N)^T G_J (S_M \otimes S_N)$$

$$= \frac{1}{20} (4(I_M \otimes S_M^T E_N S_N) + 4(S_M^T E_M S_M \otimes I_N) + (S_M^T E_M S_M \otimes S_M^T E_N S_N))$$

$$= \frac{1}{20} (4(I_M \otimes D_N) + 4(D_M \otimes I_N) + (D_M \otimes D_N),$$

which is diagonal. Therefore the eigenvalues of $G_J$ are

$$\frac{1}{5} \left(2 \cos\left(\frac{q\pi}{N + 1}\right) + 2 \cos\left(\frac{p\pi}{M + 1}\right) + \cos\left(\frac{q\pi}{N + 1}\right) \cos\left(\frac{p\pi}{M + 1}\right)\right)$$

for $q = 1 : N$ and $p = 1 : M$, so

$$\rho(G_J) = \frac{1}{5} \left(2 \cos\left(\frac{\pi}{N + 1}\right) + 2 \cos\left(\frac{\pi}{M + 1}\right) + \cos\left(\frac{\pi}{N + 1}\right) \cos\left(\frac{\pi}{M + 1}\right)\right).$$
This is always less than 1, so the Jacobi method always converges. If \( M \) and \( N \) are both large, \( \rho(G_J) \) approaches 1.

**Problem P10.1.8**

The matrix \( A = 4I_{MN} - E_M \otimes I_N - I_M \otimes E_N \) is partitioned into a \( M \times M \) matrix of blocks, each of size \( N \times N \). The \( ij \) block is

\[
A_{ij} = \begin{cases} 
4I_N - E_N & \text{if } i = j \\
-I_N & \text{if } |i - j| = 1 \\
0 & \text{otherwise.}
\end{cases}
\]

The block Jacobi method is then used. The block-diagonal of \( A \) is \( M_J = \text{diag}(A_{11}, \ldots, A_{MM}) = 4I_{MN} - I_M \otimes E_N \). Therefore \( N_J = E_M \otimes I_N \).

Observe that

\[
(S_M \otimes S_N)^T M_J (S_M \otimes S_N) = 4I_{MN} - I_M \otimes (S_M^T E_N S_N)
\]

which is a diagonal matrix, say, \( F = \text{diag}(f_1, \ldots, f_{MN}) \), where

\[
f_{(p-1)N+q} = 4 - 2 \cos\left(\frac{q\pi}{N+1}\right)
\]

for \( p = 1: M \) and \( q = 1: N \). Therefore,

\[
(S_M \otimes S_N)^T G_J (S_M \otimes S_N) = (S_M \otimes S_N)^T M_J^{-1} N_J (S_M \otimes S_N)
\]

\[
= (S_M \otimes S_N)^T M_J^{-1} (S_M \otimes S_N) (S_M \otimes S_N)^T (E_M \otimes I_N)(S_M \otimes S_N)
\]

\[
= F^{-1}(S_M^T E_M S_M \otimes I_N)
\]

\[
= F^{-1}(D_M \otimes I_N).
\]

The matrix \( D_M \otimes I_N \) is diagonal, say, \( H = \text{diag}(h_1, \ldots, h_{MN}) \), where

\[
h_{(p-1)N+q} = 2 \cos\left(\frac{p\pi}{M+1}\right)
\]

for \( p = 1: M \), \( q = 1: N \). Therefore, \( F^{-1}H \) is a diagonal matrix, so its diagonal entries are the eigenvalues of \( G_J \), namely

\[
\frac{\cos\left(\frac{p\pi}{M+1}\right)}{2 - \cos\left(\frac{q\pi}{N+1}\right)}
\]

for \( p = 1: M \) and \( q = 1: N \), so

\[
\rho(G_J) = \frac{\cos\left(\frac{p\pi}{M+1}\right)}{2 - \cos\left(\frac{q\pi}{N+1}\right)}.
\]

This is always less than 1, so the block Jacobi method always converges. If \( M \) and \( N \) are large, \( \rho(G_J) \) approaches 1.

**Problem P10.1.9**

We have

\[
M_{ADI} = \frac{1}{4}(4I_{MN} - I_M \otimes E_N)(4I_{MN} - E_M \otimes I_N)
\]

\[
= \frac{1}{4}(16I_{MN} - 4(I_M \otimes E_N) - 4(E_M \otimes I_N) + (E_M \otimes E_N))
\]
and \(N_{\text{ADI}} = \frac{1}{4}(E_M \otimes E_N)\). Observe that

\[
(S_M \otimes S_N)^T M_{\text{ADI}}(S_M \otimes S_N)
\]
\[
= \frac{1}{4} \left( 16 I_{MN} - 4(I_M \otimes S_N^T E_N S_N) - 4(S_M^T E_M S_M \otimes I_N) + (S_M^T E_M S_M \otimes S_N^T E_N S_N) \right)
\]
\[
= \frac{1}{4} \left( 16 I_{MN} - 4(I_M \otimes D_N) - 4(D_M \otimes I_N) + (D_M \otimes D_N) \right),
\]

which is a diagonal matrix, say, \(F = \text{diag}(f_1, \ldots, f_{MN})\), where

\[
f_{(p-1)N+q} = 4 - 2\cos\left(\frac{q\pi}{N+1}\right) - 2\cos\left(\frac{p\pi}{M+1}\right) + \cos\left(\frac{p\pi}{M+1}\right)\cos\left(\frac{q\pi}{N+1}\right)
\]

\[
= \left(2 - \cos\left(\frac{q\pi}{N+1}\right)\right)\left(2 - \cos\left(\frac{p\pi}{M+1}\right)\right)
\]

for \(p = 1 : M\) and \(q = 1 : N\). Therefore,

\[
(S_M \otimes S_N)^T G_{\text{ADI}}(S_M \otimes S_N) = (S_M \otimes S_N)^T M_{\text{ADI}}^{-1} N_{\text{ADI}}(S_M \otimes S_N)
\]
\[
= (S_M \otimes S_N)^T M_{\text{ADI}}^{-1}(S_M \otimes S_N)(S_M \otimes S_N)^T \frac{1}{4}(E_M \otimes E_N)(S_M \otimes S_N)
\]
\[
= F^{-1} \frac{1}{4}(S_M^T E_M S_M \otimes S_N^T E_N S_N)
\]
\[
= F^{-1} \frac{1}{4}(D_M \otimes D_N).
\]

The matrix \(\frac{1}{4}(D_M \otimes D_N)\) is diagonal, say, \(H = \text{diag}(h_1, \ldots, h_{MN})\), where

\[
h_{(p-1)N+q} = \cos\left(\frac{p\pi}{M+1}\right)\cos\left(\frac{q\pi}{N+1}\right)
\]

for \(p = 1 : M\) and \(q = 1 : N\). Therefore, \(F^{-1} H\) is a diagonal matrix, so its diagonal entries are the eigenvalues of \(G_{\text{ADI}}\), namely

\[
\rho(G_{\text{ADI}}) = \frac{\cos\left(\frac{p\pi}{M+1}\right)\cos\left(\frac{q\pi}{N+1}\right)}{\left(2 - \cos\left(\frac{q\pi}{N+1}\right)\right)\left(2 - \cos\left(\frac{p\pi}{M+1}\right)\right)}
\]

for \(p = 1 : M\) and \(q = 1 : N\), so

\[
\rho(G_{\text{ADI}}) = \frac{\cos\left(\frac{p\pi}{M+1}\right)\cos\left(\frac{q\pi}{N+1}\right)}{\left(2 - \cos\left(\frac{q\pi}{N+1}\right)\right)\left(2 - \cos\left(\frac{p\pi}{M+1}\right)\right)}
\]

This is always less than 1, so the alternating direction method always converges. If \(M\) and \(N\) are large, \(\rho(G_{\text{ADI}})\) approaches 1.

**Problem P10.1.10**

For \(\lambda > 0\) and \(k \geq 0\) an integer, define \(c_k(\lambda) = \cosh(k \cosh^{-1}(\lambda))\). Fix \(\rho \in (0, 1)\). For \(k \geq 0\) an integer, define

\[
w_{k+1} = \frac{2c_k(1/\rho)}{\rho c_{k+1}(1/\rho)}.
\]

Certainly, we can define, for all \(x \geq 0\), \(C(x) = \cosh(x \cosh^{-1}(1/\rho))\), \(S(x) = \sinh(x \cosh^{-1}(1/\rho))\), \(T(x) = \tanh(x \cosh^{-1}(1/\rho))\), and

\[
W(x + 1) = \frac{2C(x)}{\rho C(x + 1)}.
\]

If \(k \geq 1\) is an integer then \(w_k = W(k)\).
First observe that, if \( x > 0 \) then
\[
C'(x) = \sinh(x \cosh^{-1}(1/\rho)) \cosh^{-1}(1/\rho) = S(x) \cosh^{-1}(1/\rho).
\]

Using two hyperbolic trigonometric identities, we see that, if \( x > 0 \) then
\[
W'(x + 1) = \frac{2}{\rho(C(x + 1))^2} (C(x + 1)C'(x) - C(x)C'(x + 1))
= \frac{2 \cosh^{-1}(1/\rho)}{\rho(C(x + 1))^2} (C(x + 1)S(x) - C(x)S(x + 1))
= \frac{2 \cosh^{-1}(1/\rho)}{\rho(C(x + 1))^2} S(-1)
= \frac{-2 \cosh^{-1}(1/\rho)}{\rho(C(x + 1))^2} \sqrt{(1/\rho)^2 - 1}
< 0.
\]

Therefore, \( W \) is decreasing on \((1, \infty)\).

(ii)

If \( k \geq 1 \) is an integer then \( w_{k+1} - w_k = W(k + 1) - W(k) < 0 \), because \( W \) is decreasing on \((k, k + 1)\), so \( w_{k+1} < w_k \).

(iii)

Using several hyperbolic trigonometric identities, if \( x \geq 0 \) then
\[
W(x + 1) = \frac{2C(x)}{\rho(C(x)C(1) + S(x)S(1))}
= \frac{2}{\rho(\frac{1}{\rho} + T(x)\sqrt{(1/\rho)^2 - 1})}
= \frac{2}{1 + T(x)\sqrt{1 - \rho^2}}.
\]

As \( x \to \infty \), \( x \cos^{-1}(1/\rho) \to \infty \), so \( T(x) \to 1 \). Therefore,
\[
\lim_{k \to \infty} w_k = \lim_{x \to \infty} W(x) = \lim_{x \to \infty} W(x + 1) = \frac{2}{1 + \sqrt{1 - \rho^2}}.
\]

(i)

First observe that
\[
W(1) = \frac{2C(0)}{\rho C(1)} = 2.
\]

Since \( W \) is decreasing on \((1, \infty)\) and continuous on \([1, \infty)\), it is never as large as \( W(1) \) nor as small as \( \lim_{x \to \infty} W(x) \) on \((1, \infty)\). That is, \( 1 < W(x) < 2 \) for \( x \in (1, \infty) \), so, if \( k > 1 \) is an integer, \( 1 < w_k = W(k) < 2 \).

**Problem P10.1.11**

Assume that \( \rho \in \mathbb{R} \) and let \( A = \begin{pmatrix} 1 & \rho \\ -\rho & 1 \end{pmatrix} \). Gauss-Seidel takes \( M_G = \begin{pmatrix} 1 & 0 \\ -\rho & 1 \end{pmatrix} \) and \( N_G = \begin{pmatrix} 0 & -\rho \\ 0 & 0 \end{pmatrix} \), so the iteration matrix is
\[
G_G = M_G^{-1} N_G = \begin{pmatrix} 1 & 0 \\ \rho & 1 \end{pmatrix} \begin{pmatrix} 0 & -\rho \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -\rho \\ 0 & -\rho^2 \end{pmatrix}.
\]
This has characteristic polynomial \( \det(\lambda I_n - G) = \lambda(\lambda - \rho^2) \), so it has eigenvalues 0 and \(-\rho^2\) and spectral radius (bad notation - both \(G\) and \(\rho\) stand for two different things!) \(\rho(G) = \rho^2\). Therefore, the Gauss-Seidel method will converge if and only if \(|\rho| < 1\).

(b)

For SOR, \(M_{\text{SOR}} = \begin{pmatrix} 1 & -\omega \rho \\ -\omega \rho & 1 \end{pmatrix} \) and \(N_{\text{SOR}} = \begin{pmatrix} 1 & -\omega \rho \\ 0 & 1 - \omega \end{pmatrix} \), so the iteration matrix is

\[
G_{\text{SOR}} = M_{\text{SOR}}^{-1}N_{\text{SOR}} = \begin{pmatrix} 1 & 0 \\ \omega \rho & 1 - \omega \end{pmatrix} \begin{pmatrix} 1 & -\omega \rho \\ 0 & 1 - \omega \end{pmatrix} = \begin{pmatrix} 1 - \omega & -\omega \rho \\ \omega(1 - \omega)\rho & -\omega^2 \rho^2 + (1 - \omega) \end{pmatrix}.
\]

This has characteristic polynomial

\[
\det(\lambda I_n - G_{\text{SOR}}) = \lambda^2 + \lambda(\omega^2 \rho^2 - 2(1 - \omega) + (1 - \omega)^2),
\]

so it has eigenvalues

\[
\lambda_{\pm} = \frac{1}{2} \left( 2(1 - \omega) - \omega^2 \rho^2 \pm |\omega\rho| \sqrt{\omega^2 \rho^2 - 4(1 - \omega)} \right).
\]

Suppose first that \(\omega^2 \rho^2 - 4(1 - \omega) < 0\). Then both eigenvalues have nonzero imaginary parts and both have the same magnitude. In fact,

\[
|\lambda_{\pm}|^2 = \left( \frac{2(1 - \omega) - \omega^2 \rho^2}{2} \right)^2 + \left( \frac{|\omega\rho|}{2} \right)^2 (4(1 - \omega) - \omega^2 \rho^2) = |1 - \omega|^2.
\]

Therefore, \(\rho(G_{\text{SOR}}) = |1 - \omega|\), so the SOR method converges if and only if \(0 < \omega < 2\).

Suppose next that \(\omega^2 \rho^2 - 4(1 - \omega) = 0\). Then both eigenvalues are the same, so

\[
\rho(G_{\text{SOR}}) = |\lambda_{\pm}| = \frac{|2(1 - \omega) - \omega^2 \rho^2|}{2} = |1 - \omega|.
\]

Therefore the SOR method converges if and only if \(0 < \omega < 2\).

Suppose third that \(\omega^2 \rho^2 - 4(1 - \omega) > 0\). If \(2(1 - \omega) - \omega^2 \rho^2 \geq 0\) then we would have that \(1 - \omega < 0\), which then implies that \(\omega^2 \rho^2 < 0\), a contradiction. Therefore, we can assume here that \(2(1 - \omega) - \omega^2 \rho^2 < 0\). Then

\[
|\lambda_{-}| > |\lambda_{+}|, \quad \rho(G_{\text{SOR}}) = |\lambda_{-}| = \frac{1}{2} \left( \omega^2 \rho^2 - 2(1 - \omega) + |\omega\rho| \sqrt{\omega^2 \rho^2 - 4(1 - \omega)} \right).
\]

Therefore, the SOR method converges if and only if

\[
\omega^2 \rho^2 - 2(1 - \omega) + |\omega\rho| \sqrt{\omega^2 \rho^2 - 4(1 - \omega)} < 2.
\]

We do some manipulations to simplify this. It is straightforward to arrive at

\[
|\omega\rho| \sqrt{\omega^2 \rho^2 - 4(1 - \omega)} < 2(2 - \omega) - \omega^2 \rho^2.
\]

The obvious thing to do now is to square both sides, but we must make sure that the right hand side is positive. That is, \(\omega^2 \rho^2 - 2(2 - \omega) < 0\). After squaring both sides, we arrive at the condition that \(\omega^2 \rho^2 - (2 - \omega)^2 < 0\). It is easy to see that these two conditions can be replaced by the equivalent simpler conditions

\[
0 < \omega < 2 \quad \text{and} \quad \omega^2 \rho^2 - (2 - \omega)^2 < 0.
\]

At this point, I have an explicit formula for \(\rho(G_{\text{SOR}})\), valid for all \(\omega, \rho \in \mathbb{R}\), namely

\[
\rho(G_{\text{SOR}}) = \begin{cases} 
\frac{|1 - \omega|}{\frac{1}{2} \left( \omega^2 \rho^2 - 2(1 - \omega) + |\omega\rho| \sqrt{\omega^2 \rho^2 - 4(1 - \omega)} \right)} & \text{if } \omega^2 \rho^2 - 4(1 - \omega) < 0 \\
\frac{1}{2} \left( \omega^2 \rho^2 - 2(1 - \omega) + |\omega\rho| \sqrt{\omega^2 \rho^2 - 4(1 - \omega)} \right) & \text{otherwise.}
\end{cases}
\]
We can now write

\[ \omega^2 \rho^2 - 4(1 - \omega) \leq 0 \quad \text{and} \quad 0 < \omega < 2 \]

or

\[ \omega^2 \rho^2 - 4(1 - \omega) > 0 \quad \text{and} \quad 0 < \omega < 2 \quad \text{and} \quad \omega^2 \rho^2 - (2 - \omega)^2 < 0. \]

Evidently the level-0 sets of \( f_1(\rho, \omega) = \omega^2 \rho^2 - 4(1 - \omega) \) and \( f_2(\rho, \omega) = \omega^2 \rho^2 - (2 - \omega)^2 \) are central to this problem. For \( \rho \) fixed, \( f_1(\rho, \cdot) \) is quadratic except when \( \rho = 0 \), when it is linear. When \( \rho \neq 0 \), its leading coefficient is positive. Let

\[ r_{11}(\rho) = \frac{2}{1 + \sqrt{1 + \rho^2}} \quad \text{and} \quad r_{12}(\rho) = -2\left(\frac{1 + \sqrt{1 + \rho^2}}{\rho^2}\right). \]

Then \( r_{11}(\rho) \) is always a root of \( f_1(\rho, \cdot) \) and \( r_{12}(\rho) \) is a root of \( f_1(\rho, \cdot) \), provided that \( \rho \neq 0 \). Observe that \( r_{12} < 0 < r_{11} \leq 1 \).

For \( \rho \) fixed, \( f_2(\rho, \cdot) \) is quadratic except when \( |\rho| = 1 \), when it is linear. When \( |\rho| > 1 \), its leading coefficient is positive. When \( |\rho| < 1 \), its leading coefficient is negative. Let

\[ r_{21}(\rho) = \frac{2}{|\rho| + 1} \quad \text{and} \quad r_{22}(\rho) = \frac{2}{1 - |\rho|}. \]

Then \( r_{21}(\rho) \) is always a root of \( f_2(\rho, \cdot) \) and \( r_{22}(\rho) \) is a root of \( f_2(\rho, \cdot) \), provided that \( |\rho| \neq 1 \). Observe that \( 0 < r_{21} \leq 2 \) and that \( r_{22}(\rho) > 2 \) if \( |\rho| < 1 \) and \( r_{22}(\rho) < 0 \) if \( |\rho| > 1 \).

We can use these facts to get explicit analytic formulas for the set on which \( \rho(G_{SOR}) < 1 \) and, for each \( \rho \in \mathbb{R} \), the value of \( \omega \) which minimises \( \rho(G_{SOR}) \).

First we simplify the condition \( C_1 \) that

\[ f_1(\rho, \omega) \leq 0 \quad \text{and} \quad 0 < \omega < 2 \]

The first of these conditions is equivalent to \( r_{12}(\rho) \leq \omega \leq r_{11}(\rho) \). Since \( r_{12} < 0 \) and \( r_{11} \leq 2 \), we see that \( C_1 \) is equivalent to \( 0 < \omega \leq r_{11}(\rho) \).

Finding a way to simplify the condition \( C_2 \) that

\[ f_1(\rho, \omega) > 0 \quad \text{and} \quad 0 < \omega < 2 \quad \text{and} \quad f_2(\rho, \omega) < 0 \]

is more complicated. The first of these conditions is equivalent to \( \omega < r_{12}(\rho) \) or \( \omega > r_{11}(\rho) \). Since \( r_{12} < 0 \), the first and second conditions can be replaced by \( r_{11}(\rho) < \omega < 2 \). The third condition can take several forms. If \( |\rho| > 1 \), it reads that \( r_{22}(\rho) < \omega < r_{21}(\rho) \). Since \( r_{22}(\rho) < 0 \) and \( r_{11} > 0 \), when we combine this with the first two conditions, we get that \( r_{11}(\rho) < \omega < r_{21}(\rho) \). If \( |\rho| = 1 \), the third condition reads that \( \omega < r_{21}(\rho) \). Since \( r_{21}(\rho) = 1 \), when we combine this with the first two conditions, we get that \( r_{11}(\rho) < \omega < r_{21}(\rho) \). If \( |\rho| < 1 \), it reads that \( \omega < r_{21}(\rho) \) or \( \omega > r_{22}(\rho) \). Since \( r_{22}(\rho) > 2 \), when we combine this with the first two conditions, we get that \( r_{11}(\rho) < \omega < r_{21}(\rho) \). We have shown that \( C_2 \) is equivalent to the condition that \( r_{11}(\rho) < \omega < r_{21}(\rho) \). Since \( \sqrt{1 + \rho^2} \leq 1 + |\rho| \), we see that \( r_{11}(\rho) \leq r_{21}(\rho) \).

Now we see that the SOR method converges if and only if \( C_1 \) or \( C_2 \) is true, which is the case if and only if \( 0 < \omega < r_{21}(\rho) \).

We can now write

\[ \rho(G_{SOR}) = \begin{cases} 1 - \omega & \text{if } 0 < \omega < r_{11}(\rho) \\ 1 - r_{11}(\rho) & \text{if } \omega = r_{11}(\rho) \\ \frac{1}{2} \left( \omega^2 \rho^2 - 2(1 - \omega) + \omega |\rho| \sqrt{\omega^2 \rho^2 - 4(1 - \omega)} \right) & \text{if } r_{11}(\rho) < \omega < r_{21}(\rho) \\ \text{something} \geq 1 & \text{otherwise.} \end{cases} \]
Therefore \[
\frac{\partial}{\partial \omega} \rho(G_{SOR}) = \begin{cases} 
-1 & \text{if } 0 < \omega < r_{11}(\rho) \\
\omega \rho^2 + 1 + \frac{\rho}{2} \sqrt{f_1(\rho, \omega)} + \frac{\omega |\rho|}{\sqrt{f_1(\rho, \omega)}} \left( \frac{1}{2} \omega \rho^2 + \omega \right) & \text{if } r_{11}(\rho) < \omega < r_{21}(\rho) \\
\text{irrelevant} & \text{otherwise}
\end{cases}
\]

In particular, for fixed $\rho \in \mathbb{R}$, $\rho(G_{SOR})$ is decreasing for $0 < \omega < r_{11}(\rho)$ and increasing for $r_{11}(\rho) < \omega < r_{21}(\rho)$. Therefore, since $\rho(G_{SOR})$ is continuous on $\mathbb{R}^2$, for each fixed $\rho \in \mathbb{R}$, $\rho(G_{SOR})$ attains its minimum at $\omega = r_{11}(\rho)$.

Below are some of the numerical experiments that I did that led me to these results. At first it was not at all clear to me that there would be an analytic solution to this problem. In this MATLAB code, I make a contour plot of $\rho(G_{SOR})$, ignoring where it is $\geq 1$. I then ask MATLAB to find where $\rho(G_{SOR})$ attains its minimum, using the 2-output form of the \texttt{min} command. This optimal value of $\omega$ is plotted in the black dashed line against the contour plot.

I also plotted $r_{11}$ and $r_{21}$ and these seemed to line up perfectly with the optimal value of $\omega$ and the maximal value of $\omega$ allowable to keep $\rho(G_{SOR}) < 1$, respectively. These would obviously not show up on the first plot, so they are on the second one.

```matlab
rho=linspace(-10,10,1000);
omega=linspace(-1,3,1000);
[rho_mat,omega_mat]=meshgrid(rho,omega);
a=1-omega_mat-1/2*omega_mat.^2.*rho_mat.^2;
b=1/2*abs(omega_mat.*rho_mat).*(omega_mat.^2.*rho_mat.^2-4*(1-omega_mat)).^(1/2);
lambda_plus=a+b;
lambda_minus=a-b;
r_11=2./(1+(1+rho.^2).^(1/2));
r_21=2./(abs(rho)+1);
rho_of_G_SOR=max(abs(lambda_plus),abs(lambda_minus));
[smallest_rho_of_G_SOR,omega_opt_indices]=min(rho_of_G_SOR,[],1);
figure; hold on;
plot(rho,omega(omega_opt_indices),'--k','LineWidth',2);
xlabel('\rho');ylabel('\omega');ylim([-0.5,2.5]);
legend('\rho(G_{SOR})','\omega_{opt}');colorbar;
figure; hold on;
plot(rho,r_11,':');
plot(rho,r_21,'--k');
xlabel('\rho');ylabel('\omega');ylim([-0.5,2.5]);
legend('r_{11}(\rho)','r_{21}(\rho)');
```

11
Let $S \in \mathbb{R}^{n \times n}$ have SVD $S = U\Sigma V^T$, where $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$, where $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$. Let $A = \begin{pmatrix} I_n & S \\ -S^T & I_n \end{pmatrix}$. For Gauss-Seidel, $M_G = \begin{pmatrix} I_n & 0 \\ -S^T & I_n \end{pmatrix}$ and $N_G = \begin{pmatrix} 0 & -S \\ 0 & 0 \end{pmatrix}$, so the iteration matrix is

$$G_G = M_G^{-1} N_G = \begin{pmatrix} I_n & 0 \\ -S^T & I_n \end{pmatrix} \begin{pmatrix} 0 & -S \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -S \\ 0 & -S^T S \end{pmatrix}.$$

Observe that

$$\begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^T G_G \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} = \begin{pmatrix} 0 & -U^T SV \\ 0 & -V^T S^T SV \end{pmatrix} = \begin{pmatrix} 0 & -\Sigma \\ 0 & -\Sigma^2 \end{pmatrix}.$$

Therefore the eigenvalues of $G_G$ are $\{0, -\sigma_1^2, \ldots, -\sigma_n^2\}$, so $\rho(G_G) = \sigma_1^2$, and the Gauss-Seidel method converges if and only if the largest singular value of $S$ is less than 1.
For SOR, $M_{\text{SOR}} = \begin{pmatrix} I_n & 0 \\ -\omega S^T & I_n \end{pmatrix}$ and $N_{\text{SOR}} = \begin{pmatrix} (1 - \omega)I_n & -\omega S \\ 0 & (1 - \omega)I_n \end{pmatrix}$, so the iteration matrix is

$$G_{\text{SOR}} = M_{\text{SOR}}^{-1} N_{\text{SOR}} = \begin{pmatrix} I_n & 0 \\ \omega S^T & I_n \end{pmatrix} \begin{pmatrix} (1 - \omega)I_n & -\omega S \\ 0 & (1 - \omega)I_n \end{pmatrix} = \begin{pmatrix} (1 - \omega)I_n & -\omega S \\ \omega(1 - \omega)S^T & -\omega^2 S^T S + (1 - \omega)I_n \end{pmatrix}.$$  

Observe that

$$\begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^T G_{\text{SOR}} \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} = \begin{pmatrix} (1 - \omega)I_n & -\omega U^T S V \\ \omega(1 - \omega)V^T S U & -\omega^2 V^T S^T S V + (1 - \omega)I_n \end{pmatrix} = \begin{pmatrix} (1 - \omega)I_n & -\omega \Sigma \\ \omega(1 - \omega)\Sigma & -\omega^2 \Sigma^2 + (1 - \omega)I_n \end{pmatrix} = \begin{pmatrix} (1 - \omega)I_n & -\omega \sigma_1, \ldots, \sigma_n \\ \omega(1 - \omega)\sigma_1, \ldots, \sigma_n & -\omega^2 \sigma_1^2, \ldots, \sigma_n^2 + (1 - \omega)I_n \end{pmatrix}.$$  

We can’t just read off eigenvalues quite yet. However, we can transform the $2 \times 2$ block matrix of diagonal $n \times n$ matrices into an $n \times n$ block diagonal matrix of $2 \times 2$ matrices just by renumbering variables. Let $P$ be a $2n \times 2n$ permutation matrix with entries

$$p_{ij} = \begin{cases} 1 & \text{if } j \leq n \text{ and } i = 2j - 1 \\ 1 & \text{if } j \geq n + 1 \text{ and } i = 2(j - n) \\ 0 & \text{otherwise.} \end{cases}$$  

Since $P$ has exactly one 1 in each row and each column, it is orthogonal. Multiplication by $P$ on the left reorders the rows. For $i \leq n$, the $i$th row of $X$ is the $(2i - 1)$st row of $PX$. For $i \geq n + 1$, the $i$th row of $X$ is the $2(i - n)$th row of $PX$. Multiplication by $P^T$ on the right reorders the columns. For $j \leq n$, the $j$th column of $X$ is the $(2j - 1)$st column of $XP^T$. For $j \geq n + 1$, the $j$th column of $X$ is the $2(j - n)$th column of $XP^T$. Therefore,

$$P \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix}^T G_{\text{SOR}} \begin{pmatrix} U & 0 \\ 0 & V \end{pmatrix} P^T = \text{diag}(W_1, \ldots, W_n),$$

where

$$W_i = \begin{pmatrix} 1 - \omega & -\omega \sigma_i \\ \omega(1 - \omega)\sigma_i & -\omega^2 \sigma_i^2 + (1 - \omega) \end{pmatrix}.$$  

We have reduced this problem to the block version of part (b). The SOR method will converge if and only if $0 < \omega < r_{21}(\sigma_1)$ for all $i = 1 : n$. Obviously $r_{21}(\rho)$ is decreasing for $\sigma_n < \rho < \sigma_1$. Therefore the SOR method will converge if and only if $0 < \omega < r_{21}(\sigma_1)$.  

To get the optimal $\omega$, we use the fact that the spectral radius of $G_{\text{SOR}}$ is the maximum of the spectral radii of the $W_i$. Obviously $r_{11}(\rho)$ is decreasing for $\rho > 0$, so we have that $0 < r_{11}(\sigma_1) \leq r_{11}(\sigma_i)$ for all $i = 1 : n$. Therefore, using our spectral radius formula from part (b), if $\omega = r_{11}(\sigma_1)$ then $\rho(W_i) = 1 - r_{11}(\sigma_1)$ for all $i = 1 : n$, which means that $\rho(G_{\text{SOR}}) = 1 - r_{11}(\sigma_1)$. However, if $\omega \neq r_{11}(\sigma_1)$, then $\rho(W_1) > 1 - r_{11}(\sigma_1)$ because $\rho(W_1)$ attains its minimum when $\omega = r_{11}(\sigma_1)$. In this case, we would have that $\rho(G_{\text{SOR}}) > 1 - r_{11}(\sigma_1)$. Therefore, we conclude that the optimal value of $\omega$ is $r_{11}(\sigma_1)$.

**Problem P10.1.13**

(a)

Using the iteration formula,

$$e^{(k+1)} = y^{(k+1)} - x = \omega(By^{(k)} + d - y^{(k-1)}) + (y^{(k-1)} - x) = \omega(By^{(k)} + d - y^{(k-1)}) + e^{(k-1)}.$$
This equation still involves \( y^{(k)} \) and \( y^{(k-1)} \). To get rid of these, we use the assumption that \( x = Bx + d \). Then,

\[
By^{(k)} + d - y^{(k-1)} = By^{(k)} - x + d - (y^{(k-1)} - x) = B(y^{(k)} - x) - (y^{(k-1)} - x) = Be^{(k)} - e^{(k-1)}.
\]

Therefore,

\[
e^{(k+1)} = \omega Be^{(k)} + (1 - \omega)e^{(k-1)}.
\]

(b)

We already have that \( e^{(0)} = p_0(B)e^{(0)} \), where \( p_0(\lambda) = 1 \), so \( p_0 \) is an even polynomial. We are assuming that

\[
e^{(1)} = y^{(1)} - x = (By^{(0)} + d) - x = (By^{(0)} + d) - (Bx + d) = B(y^{(0)} - x) = Be^{(0)} = p_1(B)e^{(0)},
\]

where \( p_1(\lambda) = \lambda \), so \( p_1 \) is an odd polynomial.

Now suppose that \( k \geq 1 \) is an integer and that, for all \( j = 1 : k \), \( e^{(j)} = p_j(B)e^{(0)} \), where \( p_j \) is a polynomial of the same parity as \( j \). Then, using the result of part (a),

\[
e^{(k+1)} = \omega Be^{(k)} + (1 - \omega)e^{(k-1)}
\]

\[
= \omega Bp_k(B)e^{(k)} + (1 - \omega)p_{k-1}(B)e^{(0)}
\]

\[
= (\omega Bp_k(B) + (1 - \omega)p_{k-1}(B))e^{(0)}.
\]

Let \( p_{k+1}(\lambda) = \omega \lambda p_k(\lambda) + (1 - \omega)p_{k-1}(\lambda) \), so that \( e^{(k+1)} = p_{k+1}(B)e^{(0)} \). Since \( k - 1 \) has the same parity as \( k + 1 \), \( p_{k-1} \) is a polynomial of the same parity as \( k + 1 \), so \( \lambda \mapsto (1 - \omega)p_{k-1}(\lambda) \) is a polynomial of the same parity as \( k + 1 \). Since \( k \) has the opposite parity as \( k + 1 \), \( p_k \) is a polynomial of the opposite parity as \( k + 1 \), so \( \lambda \mapsto \omega \lambda p_k(\lambda) \) is a polynomial of the same parity as \( k + 1 \). Therefore, \( p_{k+1} \) is a polynomial of the same parity as \( k + 1 \). By induction, for all integers \( k \geq 0 \), \( e^{(k)} = p_k(B)e^{(0)} \), where \( p_k \) is a polynomial of the same parity as \( k \).

(c)

If \( f^{(k)} = Q^T e^{(k)} \) then the result of part (a) reads

\[
f^{(k+1)} = Q^T e^{(k+1)}
\]

\[
= \omega Q^T Be^{(k)} + (1 - \omega)Q^T e^{(k-1)}
\]

\[
= \omega (Q^T BQ)(Q^T e^{(k)}) + (1 - \omega)f^{(k-1)}
\]

\[
= QDf^{(k)} + (1 - \omega)f^{(k-1)},
\]

where \( D = \text{diag}(\lambda_1, \ldots, \lambda_n) \), where \( \lambda_1 \geq \cdots \geq \lambda_n \). Component-wise, this reads

\[
f_j^{k+1} = \omega \lambda_j f_j^{(k)} + (1 - \omega)f_j^{(k-1)}.
\]

For each \( j \), this is a 3-term linear recurrence relation with characteristic polynomial

\[
q_j(r) = r^2 - \omega \lambda_j - (1 - \omega),
\]

which has roots

\[
r_j^\pm = \frac{1}{2} \left( \omega \lambda_j \pm \sqrt{\omega^2 \lambda_j^2 + 4(1 - \omega)} \right).
\]

If \( \omega^2 \lambda_j^2 + 4(1 - \omega) \neq 0 \) then \( r_j^- \neq r_j^+ \) and the solution is

\[
f_j^{(k)} = c_- (r_j^-)^k + c_+ (r_j^+)^k
\]

for all integers \( k \geq 0 \), where

\[
\alpha_j = \frac{r_j^+ f_j^{(0)} - f_j^{(1)}}{r_j^+ - r_j^-}.
\]
\[
\beta_j = \frac{f_j^{(1)} - r_j f_j^{(0)}}{r_j^+ - r_j^-}.
\]

If \( \omega^2 \lambda_j^2 + 4(1 - \omega) = 0 \) and \( \lambda \neq 0 \) then it turns out that \( \omega \neq 0 \), so \( r_j^- = r_j^+ = r_j = \omega \lambda_j / 2 \neq 0 \), and the solution is
\[
f_j^{(k)} = c_- r_j^k + c_+ k r_j^k
\]
for all integers \( k \geq 0 \), where
\[
\alpha_j = f_j^{(0)}
\]
and
\[
\beta_j = \frac{f_j^{(1)} - r_j f_j^{(0)}}{r_j}.
\]

If \( \omega^2 \lambda_j^2 + 4(1 - \omega) = 0 \) and \( \lambda = 0 \) then it turns out \( \omega = 1 \), and the solution is
\[
f_j^{(k)} = 0
\]
for all integers \( k \geq 2 \).

(d) The method converges if and only if \( \lim_{k \to \infty} e_j^{(k)} = 0 \). Since \( Q \) is orthogonal, \( \|e_j^{(k)}\|_2 = \|Qf_j^{(k)}\|_2 = \|f_j^{(k)}\|_2 \), so the method converges if and only if \( \lim_{k \to \infty} f_j^{(k)} = 0 \). An optimal \( \omega \) would send \( f_j^{(k)} \to 0 \) fastest in some reasonable sense, because this would send \( e_j^{(k)} \to 0 \) fastest in some other reasonable sense. By part (c), we see that \( f_j^{(k)} \) grows like \( (r_j^\pm)^k \) or \( k(r_j^\pm)^k \). Certainly the method converges if and only if all the \( |r_j^\pm| < 1 \). I will declare \( \omega \) to be optimal if it minimises \( \max_j \left( \max(|r_j^-|, |r_j^+|) \right) \). This is arbitrary yet reasonable. There are many other reasonable measures of “optimality,” and they are probably not all equivalent.

This is very similar to Problem P10.1.11(b). Hence I suspect there is something special about the solution \( g(\lambda_j) = 2 - \sqrt{1 - \lambda_j^2} \) of \( \omega^2 \lambda_j^2 + 4(1 - \omega) = 0 \). This is confirmed with the plots generated by the following MATLAB code.

```matlab
lambda=linspace(-2,2,1000);
omega=linspace(-1,3,1000);
[lambda_mat,omega_mat]=meshgrid(lambda,omega);
a=1/2*omega_mat.*lambda_mat;
b=1/2*(omega_mat.^2.*lambda_mat.^2+4*(1-omega_mat)).^(1/2);
r_plus=a+b;
r_minus=a-b;
g=2-(1-lambda.^2).^(1/2);
max_r=max(abs(r_plus),abs(r_minus));
[smallest_lambda_of_max_r,omega_opt_indices]=min(max_r,[],1);
figure; hold on;
contour(lambda,omega,max_r.*max_r<1+max_r>1);
plot(lambda,omega,omega_opt_indices,'--k','LineWidth',2);
xlabel('\lambda');ylabel('\omega');ylim([-0.5,2.5]);
legend('\text{max(|r_-|,|r_+|)}','\omega_{opt}');colorbar;
figure; hold on;
plot(lambda,g,'k');
xlabel('\lambda');ylabel('\omega');ylim([-0.5,2.5]);
legend('g(\lambda)');
```
Evidently, the method converges if and only if $0 < \omega < 2$ and $-1 < \lambda_j < 1$ for all $j = 1 : n$. The overall optimal $\omega$ is $g(\max(|\lambda_1|,|\lambda_n|))$. The reasoning is exactly the same as in Problem P10.1.11(c).

Having already done Problem P10.1.11, I am not going to rigorously justify the statements in the last paragraph.