Homework #2. MF 501Z.
(Due on Thursday March 27, 2008).

(4b) 1. The weighted Jacobi method for solving the matrix equation $Ax = b$ is given by

$$x^{k+1} = (1-\omega)x^k + \omega X_{j,k}$$

$$X_{j,k} = -D^T(A-D)x^k + D^Tb$$

where $D$ is the diagonal matrix such that $d_{ii} = a_{ii}$ for $i = 1, 2, ..., n$.

Assuming that $A$ is symmetric and positive definite and that $a_{ij} = 0$ for $i \neq j$, for what values of $\omega$ is the method convergent?

(4b) 2. The Kantorovich inequality is equivalent to the following inequality

$$\left( \sum_{i=1}^{n} \frac{\varepsilon_i}{\mu_i} \right) \left( \sum_{i=1}^{n} \frac{\varepsilon_i}{\mu_i}^{-1} \right) \leq \frac{(M+m)^2}{4Mm}$$

where $0 < m \leq \mu_i \leq M$ and $\varepsilon_i > 0$ for $i = 1, ..., n$, and $\sum_{i=1}^{n} \varepsilon_i = 1$.

Prove this inequality.
Hint: First, show that for each \( x_i \), you can find \( p_i \) and \( q_i \) such that

\[
\begin{align*}
\mu_i &= p_i \, \mu + q_i \, \mu' \\
\tilde{\mu}_i &= p_i \, \tilde{\mu} + q_i \, \tilde{\mu}'
\end{align*}
\]

\( p_i, q_i \geq 0 \)

\( p_i + q_i \leq 1 \).

Then, set \( p = \sum_{i=1}^{n} \frac{\xi_i}{\gamma_i} p_i \) and \( q = \sum_{i=1}^{n} \frac{\xi_i}{\gamma_i} q_i \),

show that

\[
\left( \sum_{i=1}^{n} \frac{\xi_i}{\gamma_i} \mu_i \right) \left( \sum_{i=1}^{n} \frac{\xi_i}{\gamma_i} \tilde{\mu}_i \right) = (p \, \mu + q \, \mu') \left( p \, \tilde{\mu} + q \, \tilde{\mu}' \right),
\]

and conclude.

(12.9) 3

Let \( \Omega \) be the unit square and let \( U_{ij} \) denote the classic finite difference approximation to the exact solution of

\[
\begin{cases}
-\Delta u = 1 & \text{in } \Omega, \\
u = 0 & \text{on } \partial \Omega,
\end{cases}
\]

that is

\[
\begin{cases}
-\frac{1}{h^2} \left( U_{i+1,j} + U_{i-1,j} + U_{i,j+1} + U_{i,j-1} - 4 U_{i,j} \right) = 1 \\
U_{i,j} = 0 & \text{if } i,j = 1, \ldots, N-1 \\
\end{cases}
\]

Here \( h = \frac{1}{N} \).
The objective of this exercise is to compare the performance of the Jacobi, Gauss-Seidel, SOR (with the parameter \( w \) of your choice!), steepest descent and conjugate gradient methods.

Given a value for \( R = \frac{1}{n} \), and for each of these methods, compute the number of iterations that reduce the initial error \( (\text{Take } x^0 = 0) \) \( 10^6 \) times. Take \( N = 4, 8, 16, \ldots \)

Fill the following table:

<table>
<thead>
<tr>
<th>( N = \frac{1}{n} )</th>
<th>J</th>
<th>G-S</th>
<th>SOR</th>
<th>SD</th>
<th>CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>\vdots</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Explain the results by using the theory, if possible.
Homework #2. ME 5012: A solution

1. The iteration matrix of the weighted Jacobi method is

\[ G = \text{Id} - \omega \tilde{A}. \]

It can be easily verified that the method is consistent regardless of the value of the parameter \( \omega \).

To find for what values of \( \omega \), \( G \) has a spectrum strictly less than one, we are going to use the theorem on page 4 of the notes on "Classic Iterative Methods." The result states that if we write

\[ A = N - P, \]

and not

\[ G = \tilde{N}^T P, \]

we have that \( P(G) < 1 \) provided \( A \) is symmetric and positive definite and provided

\[ Q = N + \tilde{N}^T - A \]

is positive definite.

In our case, we have that
\[ \text{Id} - \omega \mathbf{D}' \mathbf{A} = \mathbf{G} \]
\[ = \mathbf{N}' \mathbf{P} \]
\[ = \mathbf{N}'(\mathbf{N}' - \mathbf{A}) \]
\[ = \text{Id} - \mathbf{N}' \mathbf{A} \]

and so we must have \( \mathbf{N} = \mathbf{D}/\omega \). This implies that

\[ \mathbf{Q} = \frac{2}{\omega} \mathbf{D} - \mathbf{A}, \]

and it only remains to find for what values of the parameter \( \omega \) this matrix is positive definite.

Since \( \mathbf{A} \) is symmetric and positive definite and since its off-diagonal terms are negative, it can be shown that the matrix

\[ 2\mathbf{D} - \mathbf{A} \]

is positive definite. Using this fact, we see that

\[ \mathbf{Q} = 2\left(1 - \frac{\omega}{\omega}\right)\mathbf{D} + (2\mathbf{D} - \mathbf{A}) \]

is positive definite if \( \left(1 - \frac{\omega}{\omega}\right) > 0 \), that is, if \( \omega \in (0,1) \), the value \( \omega = 0 \), however, must be ruled out since then \( \mathbf{G} = \text{Id} \). Thus the method is stable for \( \omega \in (0,1) \).
If the theorem in the notes escaped your attention reading, let us deal with a simpler case and assume the hypothesis used in class, namely, that the matrix $A$ is strictly diagonal dominant.

So, let $e$ be an eigenvector of $A$ and let $\lambda$ be its corresponding eigenvalue. Then the equation

$$Ge = \lambda e$$

becomes

$$e - \omega \bar{D}^\top A e = \lambda e$$

$$\Rightarrow \quad e (1 - \omega) - \omega \bar{D}^\top (A - D) e = \lambda e$$

$$\Rightarrow \quad (\lambda - 1 + \omega) e = -\omega \bar{D}^\top (A - D) e$$

in equivalently

$$(\lambda - 1 + \omega) e_i = -\sum_{j=1, j \neq i}^{n} \frac{a_{ij}}{a_{ii}} \omega e_j, \quad i = 1, \ldots, n.$$ 

Taking absolute values we get

$$|\lambda - 1 + \omega| |e_i| \leq -\sum_{j=1, j \neq i}^{n} \frac{a_{ij}}{a_{ii}} |\omega| |e_j|, \quad i = 1, \ldots, n.$$ 

If $i$ is such that $|e_j| \leq |e_i|, j = 1, \ldots, n$, we get

$$|\lambda - 1 + \omega| \leq -\sum_{j=1, j \neq i}^{n} \frac{a_{ij}}{a_{ii}} |\omega| =: \Theta.$$
\[-\theta \leq \lambda - 1 + \omega \leq \theta.\]

This implies that

\[1 - \omega - \theta \leq \lambda \leq \theta - \omega + 1\]

and if we want \(|\lambda| < 1\), we must have

\[-1 < 1 - \omega - \theta \quad \text{and} \quad \theta - \omega + 1 < 1.\]

The second inequality implies that \(\omega\) must be nonnegative and that \(\theta < \omega\), which is nothing but our assumption of the matrix being strictly diagonal dominant. The second inequality now holds if and only if

\[\omega < \frac{2}{1 - \sum_{i \neq j} a_{ij} \over a_{ii}}\]

Hence, the method converges if

\[\omega \in (0, \frac{2}{1 - \sum_{i \neq j} a_{ij} \over a_{ii}}).\]
2. By definition of $p_i$ and $q_i$, namely,

\[ \mu_i = p_i \frac{M}{\mu_i} + q_i \frac{M}{\mu_i} \]

\[ \mu_i^{-1} = p_i^{-1} + q_i^{-1} \]

we easily obtain that

\[ p_i = \frac{\mu_i^2 - m^2}{M^2 - m^2} \cdot \frac{M}{\mu_i} \]

\[ q_i = \frac{M^2 - \mu_i^2}{M^2 - m^2} \cdot \frac{M}{\mu_i} \]

We then easily see that $p_i$ and $q_i$ are non-negative since $\mu_i \in [m, M]$. Now

\[ p_i + q_i = \frac{1}{M^2 - m^2} \left[ M \mu_i - \frac{m^2 M}{\mu_i} + \frac{M^2}{\mu_i} - m \mu_i \right] \]

\[ = \frac{1}{M + m} \left[ \frac{\mu_i + m M}{\mu_i} \right] \]

\[ \leq \frac{1}{M + m} \max_{\mu \in [m, M]} \left[ \frac{\mu + m M}{\mu} \right] \]

\[ = 1. \]

Since

\[ \sum_{i=1}^{n} \varepsilon_i \mu_i = p M + q \frac{M}{\mu_i} \]

\[ \sum_{i=1}^{n} \varepsilon_i \mu_i^{-1} = p^{-1} + q \frac{M}{\mu_i} \]

we readily get that

\[ \left( \sum_{i=1}^{n} \varepsilon_i \mu_i \right) \left( \sum_{i=1}^{n} \varepsilon_i \mu_i^{-1} \right) = (p M + q \frac{M}{\mu_i})(p^{-1} + q \frac{M}{\mu_i}) \]
Since $p_i, q_i$ are non-negative and $p_i + q_i \leq 1$, we have that $p, q$ are non-negative and $p + q \leq 1$. Then

\[
\left( \sum_{i=1}^{n} \frac{p_i}{\alpha_i} \right) \left( \sum_{i=1}^{n} \frac{q_i}{\alpha_i} \right) \leq \max_{\substack{p, q \geq 0 \\ p + q \leq 1}} \Phi(p, q),
\]

where

\[
\Phi(p, q) = (pH + qM) \left( pH' + qM' \right).
\]

But

\[
\Phi(p, q) = p^2 + q^2 + pq \left( \frac{M}{M} + \frac{M}{M} \right) = (p + q)^2 + pq \left( \frac{M}{M} + \frac{M}{M} - 2 \right)
\]

and since $\frac{M}{M} + \frac{M}{M} > 2$,

\[
\Phi(p, q) \leq \Phi \left( \frac{1}{2}, \frac{1}{2} \right) = 1 + \frac{1}{4} \left( \frac{M}{M} + \frac{M}{M} - 2 \right) = \frac{1}{4} \left( \frac{M}{M} + \frac{M}{M} + 2 \right) = \frac{(M + M)^2}{4HM}.
\]

Hence

\[
\left( \sum_{i=1}^{n} \frac{p_i}{\alpha_i} \right) \left( \sum_{i=1}^{n} \frac{q_i}{\alpha_i} \right) \leq \frac{(M + M)^2}{4HM},
\]

as wanted.
In this result, let us see if we have a similar result for the eigenfunctions \( \{ U_{ij} \} \) of the discrete version of the above problem given by \((P)\). So, take

\[
U_{ij} = \sin \left( \pi \frac{i}{N} \right) \sin \left( \pi \frac{j}{N} \right)
\]

for \( i = 0 \to N \) and \( j = 0 \to N \). First, we see that \( U_{ij} \) satisfies the boundary conditions since \( U_{ij} = 0 \) for \( i = 0, i = N, j = 0, \) or \( j = N \). Note also that both \( n \) and \( m \) are integers bigger than 0 and smaller than \( N \).
Now set

\[ \Theta_{x,j} = -\frac{1}{\theta^2} \left( \frac{\partial}{\partial x} \left( \frac{\partial}{\partial x} \right) \right) \psi^{k,j} \psi^{k,j-1} - 4 \psi^{k,j} \right) . \]

Then, noting that

\[ \psi^{k,j} + \psi^{k,j-1} = \left( \sin \left( \frac{\pi h}{N} \right) + \sin \left( \frac{\pi m}{N} \right) \right) \sin \left( \frac{\pi \theta}{N} \right) \]

\[ = 2 \cos \left( \frac{\pi h}{N} \right) \sin \left( \frac{\pi h}{N} \right) \sin \left( \frac{\pi \theta}{N} \right) \]

\[ = 2 \cos \left( \frac{\pi h}{N} \right) \psi^{k,j} , \]

and that, similarly,

\[ \psi^{k,j-1} + \psi^{k,j-1} = 2 \cos \left( \frac{\pi m}{N} \right) \psi^{k,j} , \]

we obtain that

\[ \Theta_{x,j} = \left[ -\frac{1}{\theta^2} \left( 2 \cos \left( \frac{\pi m}{N} \right) + 2 \cos \left( \frac{\pi h}{N} \right) \right) - 4 \right] \psi^{k,j} \]

\[ = \left[ +\frac{1}{\theta^2} \left( +4 \sin^2 \left( \frac{\pi h}{2N} \right) + 4 \sin^2 \left( \frac{\pi m}{2N} \right) \right) \right] \psi^{k,j} \]

\[ = \left[ \left( \frac{\sin \left( \frac{\pi h}{2N} \right)}{\frac{\pi h}{2N}} \right)^2 \frac{1}{\pi h^2} + \left( \frac{\sin \left( \frac{\pi m}{2N} \right)}{\frac{\pi m}{2N}} \right)^2 \frac{1}{\pi m^2} \right] \psi^{k,j} \]

given that \( h = \frac{1}{N} \). We thus see that \( \psi^{k,j} \) is indeed an eigenfunction of the discrete operator defined
by \((P)\), whose corresponding eigenvalue is

$$\lambda_{nm} = \left(\frac{\sin \left(\frac{\pi n}{2N}\right)}{\frac{\pi n}{2N}}\right)^2 \pi^2 n^2 + \left(\frac{\sin \left(\frac{\pi m}{2N}\right)}{\frac{\pi m}{2N}}\right)^2 \pi^2 m^2.$$ 

Since the function \(\frac{1}{\sin \theta}\) goes to one as \(\theta\) goes to zero, we see that \(\lambda_{nm}\) goes to \(\pi(n^2+m^2)\) as \(N\) goes to infinity, as expected!

With this information, we can obtain the spectral radius of the iteration matrix of the Jacobi method, \(I - D^{-1} A\). Since \(D = \frac{4}{h^2} I_d\), we get that the eigenvalues of \(I - D^{-1} A\) are

$$\lambda_{nm}^J := 1 - \frac{h^2}{4} \lambda_{nm} \quad 1 \leq n, m \leq N-1$$

Hence

$$\lambda_{nm}^J = 1 - \sin^2 \left(\frac{\pi n}{2N}\right) - \sin^2 \left(\frac{\pi m}{2N}\right) \quad 1 \leq n, m \leq N-1.$$ 

and the spectral radius is then

$$\rho_J := \max_{1 \leq n, m \leq N-1} \left| \lambda_{nm}^J \right| = \left| \lambda_{11}^J \right|$$

$$= \left| 1 - 2 \sin^2 \left(\frac{\pi (N-1)}{2N}\right) \right|$$

$$= \left| 1 - 2 \sin^2 \left(\frac{\pi}{2N}\right) \right|$$
Since our matrix is symmetric, positive definite matrix who is also block-tridiagonal, we have

\[ \rho_{\text{SOR}} = \omega_{\text{opt}} - 1 < \rho_{\text{GS}} = \rho_j^2 \]

where

\[ \omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho_j^2}} \]

Hence

\[ \rho_{\text{GS}} = \left| 1 - 2 \sin^2 \left( \frac{\pi}{2N} \right) \right|^2 \]

\[ \rho_{\text{SOR}} = \frac{2}{1 + \sqrt{4 \sin^2 \left( \frac{\pi}{2N} \right) - 4 \sin^4 \left( \frac{\pi}{2N} \right)}} - 1 \]

So, for \( N \geq 2 \),

\[ \rho_j = 1 - 2 \sin^2 \left( \frac{\pi}{2N} \right), \]

\[ \rho_{\text{GS}} = 1 - \sin^2 \left( \frac{\pi}{N} \right), \]

\[ \rho_{\text{SOR}} = 1 - \frac{2 \sin \left( \frac{\pi}{N} \right)}{1 + \sin \left( \frac{\pi}{N} \right)} \]

and for big \( N \),
\[
\rho_j \sim 1 - \frac{\pi^2}{2N^2}
\]
\[
\rho_i \sim 1 - \frac{\pi^2}{N^2}
\]
\[
\rho_{\text{SOR}} \sim 1 - 2\frac{\pi}{N}
\]

To reduce the residual error 10^6 times, we thus need that
\[
\rho^k = 10^{-6}
\]
\[
\Rightarrow k \log \rho = -6 \log 10
\]
\[
\Rightarrow k = -6 \log 10 \div \log \rho
\]

Hence, since \(\log_e (1 - x) \sim -x\) for small \(x > 0\),
\[
k_j = -6 \log_{10_0} \div \log \rho_j \sim \frac{12}{\pi^2} \log_{10_0} N^2
\]
\[
k_i \sim \frac{6}{\pi^2} \log_{10_0} N^2
\]
\[
k_{\text{SOR}} \sim \frac{3}{\pi} \log_{10_0} N
\]

To see how many iterations we must do for the SD and CG methods, let us recall that
\[
\| x^k - x \|_A \leq \left( \frac{k-1}{k+1} \right)^k \| x^0 - x \|_A \quad (SD)
\]
\[
\| x^k - x \|_A \leq 2 \left( \frac{1}{\sqrt{k+1}} \right)^k \| x^0 - x \|_A \quad (CG)
\]

where \(k\) is the current number of \(A\).
\[ k_{3D} = -6 \log_{10} \left[ \frac{\log\left( \frac{k-1}{k+1} \right)}{\log\left( \frac{1}{k+1} \right)} \right], \]
\[ k_{CG} = \log(10^{6/2}) / \log \left( \frac{\sqrt{k-1}}{\sqrt{k+1}} \right). \]

To evaluate \( k \) it is easy since \( A \) being symmetric and positive definite,

\[ k = \max_{1 \leq n,m \leq N-1} \frac{\lambda_{nm}}{\max_{1 \leq n,m \leq N-1} \lambda_{nm}} \]

\[ = \frac{\lambda_{N+1,N-1}}{\lambda_{11}} \]

\[ = \frac{\sin^2 \left( \frac{\pi}{2N} \right)}{\sin^2 \left( \frac{\pi}{2N} \right)} \]

\[ = \frac{1}{\sin^2 \left( \frac{\pi}{2N} \right)} - 1 \]

\[ \approx \frac{4N^2}{\pi^2} \]

for large \( N \). So, if \( N \) is large, \( \alpha \approx k \) and

\[ k_{3D} \approx \frac{12}{\pi^2} \log_{10} 10 \cdot N^2, \]

\[ k_{CG} \approx \left[ 6 \log_{10} 10 + \log_2 2 \right] \frac{N}{\pi}. \]
Hence, for $N$ big, we have

\[ k_J \sim 2.8N^2 \]
\[ k_{GS} \sim 1.4N^2 \]
\[ k_{SOR} \sim 2.2N \]
\[ k_{SD} \sim 2.8N^2 \]
\[ k_{CG} \sim 4.6N \]

So, we have the following theoretical result:

<table>
<thead>
<tr>
<th>$N = \frac{1}{h}$</th>
<th>J</th>
<th>GS</th>
<th>SOR</th>
<th>SD</th>
<th>CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>45</td>
<td>22</td>
<td>9</td>
<td>22</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>179</td>
<td>90</td>
<td>18</td>
<td>90</td>
<td>37</td>
</tr>
<tr>
<td>16</td>
<td>717</td>
<td>358</td>
<td>35</td>
<td>358</td>
<td>73</td>
</tr>
<tr>
<td>32</td>
<td>2,867</td>
<td>1433</td>
<td>70</td>
<td>1433</td>
<td>147</td>
</tr>
</tbody>
</table>

(You can now compare your results with these theoretical predictions!)

Now, if you did not used the SOR method with optimal $\omega$, you have to proceed as follows:

First, (see page 9 of the notes on "Classical Iterative Methods"), note that in our case the eigenvalues of the iteration matrix of the SOR method are

\[ \lambda_{nm}^{\text{SOR}} = \frac{1}{\omega} \frac{1}{\lambda_{nm}^{\text{mm}}} + \frac{1}{\lambda_{nm}^{\text{mm}}} - \frac{1}{\lambda_{nm}^{\text{mm}}} \quad 1 \leq n, m \leq N-1. \]
Hence

\[
\lambda_{hm}^{\text{SOR}} = \frac{\omega - \sin^2 \left( \frac{\pi m}{2N} \right) - \sin^2 \left( \frac{\pi n}{2N} \right)}{\omega \sqrt{1 - \sin^2 \left( \frac{\pi m}{2N} \right) - \sin^2 \left( \frac{\pi n}{2N} \right)}}
\]

and, assuming that \( \omega > 1 \),

\[
\lambda_{hm}^{\text{SOR}} \leq \lambda_{11}^{\text{SOR}} \quad (=: \rho_{\text{SOR}})
\]

\[
\sim \frac{\omega - \left( \frac{\pi}{2N} \right)^2}{\omega}
\]

\[
\sim 1 - \frac{\pi^2}{2\omega N^2}
\]

Hence

\[
k^{\text{SOR}} \cong -6 \log_{10} e \rho_{\text{SOR}} \sim \frac{120}{\pi^2} \log_{10} e \cdot N^2.
\]

We then see that the behavior of SOR in this case is by far inferior to that of SOR with optimal choice of \( \omega \).