Caleb Logemann
MATH 562 Numerical Analysis II
Final Exam

1. Let $A \in \mathbb{R}^{m\times m}$ be written in the form $A = L + D + U$, where L is strictly lower triangular, D is the diagonal of A, and U is the strictly upper triangular part of A. Assuming D is invertible, $A\mathbf{x} = \mathbf{b}$ is equivalent to $\mathbf{x} = -D^{-1}(L+U)\mathbf{x} + D^{-1}\mathbf{b}$. The Jacobi iteration method for solving $A\mathbf{x} = \mathbf{b}$ is defined by

$$\mathbf{x}^{(n+1)} = -D^{-1}(L+U)\mathbf{x}^{(n)} + D^{-1}\mathbf{b}$$

Show that if A is nonsingular and strictly row diagonally dominant:

$$0 < \sum_{j\neq i} (|a_{ij}|) < |a_{ii}|$$

then the Jacobi iteration converges to $\mathbf{x}_* = A^{-1}\mathbf{b}$ for each fixed $\mathbf{b} \in \mathbb{R}^m$.

Proof. Let $\mathbf{e}_n$ be the error of the nth iteration of the Jacobi iteration from the actual solution, that is let

$$\mathbf{e}_n = \mathbf{x}^{(n)} - \mathbf{x}_*.$$

The Jacobi iteration converges to the real solution if

$$\lim_{n\to\infty} (\|\mathbf{e}_n\|_\infty) = 0$$

The error vector can be expressed recursively by noting that $\mathbf{x}^{(n)}$ is the Jacobi iteration evaluated on $\mathbf{x}^{(n-1)}$ and that $\mathbf{x}_*$ is a fixed point of the Jacobi iteration as it is the true solution to the linear system. This means that

$$\begin{aligned}\mathbf{x}^{(n)} &= -D^{-1}(L+U)\mathbf{x}^{(n-1)} + D^{-1}\mathbf{b}\\ \mathbf{x}_* &= -D^{-1}(L+U)\mathbf{x}_* + D^{-1}\mathbf{b}.\end{aligned}$$

Therefore we can express the error recursively as

$$\begin{aligned}\mathbf{e}_n &= \mathbf{x}^{(n)} - \mathbf{x}_*\\ \mathbf{e}_n &= \left(-D^{-1}(L+U)\mathbf{x}^{(n-1)} + D^{-1}\mathbf{b}\right) - \left(-D^{-1}(L+U)\mathbf{x}_* + D^{-1}\mathbf{b}\right)\\ \mathbf{e}_n &= -D^{-1}(L+U)\mathbf{x}^{(n-1)} + D^{-1}(L+U)\mathbf{x}_*\\ \mathbf{e}_n &= -D^{-1}(L+U)\left(\mathbf{x}^{(n-1)} - \mathbf{x}_*\right)\\ \mathbf{e}_n &= -D^{-1}(L+U)\mathbf{e}_{n-1}.\end{aligned}$$

Extrapolating this backwards we see that $\mathbf{e}_n$ can be expressed in terms of $\mathbf{e}_0$

$$\mathbf{e}_n = \left(-D^{-1}(L+U)\right)^n \mathbf{e}_0.$$

Now we can consider the limit of $\|\mathbf{e}_n\|_\infty$ as n goes to infinity.

$$\lim_{n\to\infty}(\|\mathbf{e}_n\|_\infty) = \lim_{n\to\infty}\left(\left\|\left(-D^{-1}(L+U)\right)^n \mathbf{e}_0\right\|_\infty\right)$$
$$\lim_{n\to\infty}(\|\mathbf{e}_n\|_\infty) \leq \|\mathbf{e}_0\|_\infty \lim_{n\to\infty}\left(\left\|D^{-1}(L+U)\right\|_\infty^n\right)$$

Now consider $\left\|D^{-1}(L+U)\right\|_\infty$. The infinity norm is the max row sum of the matrix, that is

$$\left\|D^{-1}(L+U)\right\|_\infty = \max *_{1\leq i\leq m} \sum_{j=1}^{m}\left(\left|\left(D^{-1}(L+U)\right)_{ij}\right|\right)$$

Because $L+U = A-D$, $(L+U)_{ij} = a_{ij}$ if $i \neq j$ and $(L+U)_{ii} = 0$. Also D^{-1} is diagonal with $\left(D^{-1}\right)_{ii} = \frac{1}{D_{ii}} = \frac{1}{a_{ii}}$. Therefore the matrix product $D^{-1}(L+U)$ has entries $\left(D^{-1}(L+U)\right)_{ij} = \frac{a_{ij}}{a_{ii}}$ if $i \neq j$ or if $i = j$, then $\left(D^{-1}(L+U)\right)_{ii} = 0$. We can now say that

$$\left\|D^{-1}(L+U)\right\|_\infty = \max *_{1\leq i\leq m} \sum_{j\neq k}\left(\left|\frac{a_{ij}}{a_{ii}}\right|\right)$$
$$\left\|D^{-1}(L+U)\right\|_\infty = \max *_{1\leq i\leq m} \frac{1}{|a_{ii}|}\sum_{j\neq k}\left(|a_{ij}|\right)$$

However since A is strictly row diagonally dominant $|a_{ii}| > \sum_{j\neq k}\left(|a_{ij}|\right)$, we can conclude that $\frac{1}{|a_{ii}|}\sum_{j\neq k}\left(|a_{ij}|\right) < 1$. Therefore

$$\left\|D^{-1}(L+U)\right\|_\infty < 1$$

Since $\left\|D^{-1}(L+U)\right\|_\infty < 1$, it is true that $\lim_{n\to\infty}\left(\|D^{-1}(L+U)\|_\infty^n\right) = 0$. Thus

$$\lim_{n\to\infty}(\|\mathbf{e}_n\|_\infty) \leq \|\mathbf{e}_0\|_\infty \lim_{n\to\infty}\left(\left\|D^{-1}(L+U)\right\|_\infty^n\right)$$
$$\lim_{n\to\infty}(\|\mathbf{e}_n\|_\infty) \leq 0$$

This shows that the error converges to zero, and this proves that the Jacobi iteration does converge to the true solution if A is strictly row diagonally dominant. □

2. Let $A \in \mathbb{R}^{m\times m}$ be symmetric positive definite (SPD), $\mathbf{b} \in \mathbb{R}^m$ and define $\phi : \mathbb{R}^m \to \mathbb{R}$ by

$$\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}$$

Suppose K is a subspace of $\mathbb{R}^m$. Show that $\hat{\mathbf{x}} \in K$ minimizes $\phi(\mathbf{x})$ over K if and only if $\nabla\phi(\hat{\mathbf{x}}) \perp K$.

Proof. First let me describe $\nabla\phi(\mathbf{x})$.

$$\begin{aligned}
\nabla\phi(\mathbf{x}) &= \begin{bmatrix} \frac{\partial\phi}{\partial x_1} \\ \cdots \\ \frac{\partial\phi}{\partial x_m} \end{bmatrix} \\
\frac{\partial\phi}{\partial x_i} &= \frac{\partial}{\partial x_i}\Big(\frac{1}{2}\mathbf{x}^T A\mathbf{x} - \mathbf{x}^T\mathbf{b}\Big) \\
&= \frac{\partial}{\partial x_i}\left(\frac{1}{2}\sum_{j=1}^{m}\left(x_j\sum_{k=1}^{m}(a_{jk}x_k)\right) - \sum_{j=1}^{m}(x_j b_j)\right) \\
&= \frac{1}{2}\frac{\partial}{\partial x_i}\sum_{j=1}^{m}\left(x_j\sum_{k=1}^{m}(a_{jk}x_k)\right) - b_i \\
&= \frac{1}{2}\frac{\partial}{\partial x_i}\left(x_i\sum_{k=1}^{m}(a_{ik}x_k) + \sum_{j\neq i}\left(x_j\sum_{k=1}^{m}(a_{jk}x_k)\right)\right) - b_i \\
&= \frac{1}{2}\left(\sum_{k\neq i}(a_{ik}x_k) + 2a_{ii}x_i + \sum_{j\neq i}(a_{ji}x_j)\right) - b_i \\
&= \frac{1}{2}\left(\sum_{k=1}^{m}(a_{ik}x_k) + \sum_{j=1}^{m}(a_{ji}x_j)\right) - b_i \\
&= \frac{1}{2}\Big((A\mathbf{x})_i + \Big(A^T\mathbf{x}\Big)_i\Big) - b_i
\end{aligned}$$

This is one entry of the vector $\nabla\phi(\mathbf{x})$ therefore we can write the entire vector as

$$\nabla\phi(\mathbf{x}) = \frac{1}{2}\Big(A\mathbf{x} + A^T\mathbf{x}\Big) - \mathbf{b}$$

Since A is symmetric, $A = A^T$, this symplifies to

$$\nabla\phi(\mathbf{x}) = A\mathbf{x} - \mathbf{b}$$

Now assume that $\nabla\phi(\hat{\mathbf{x}}) \perp K$, therefore $\mathbf{x}\cdot(A\hat{\mathbf{x}} - \mathbf{b}) = 0$ for any $\mathbf{x} \in K$. Let $\mathbf{x} \in K$, then $\mathbf{x} = \hat{\mathbf{x}} + \mathbf{y}$ for some $\mathbf{y} \in K$.

$$\begin{aligned}
\phi(\mathbf{x}) &= \phi(\hat{\mathbf{x}} + \mathbf{y}) \\
&= \frac{1}{2}(\hat{\mathbf{x}} + \mathbf{y})^T A(\hat{\mathbf{x}} + \mathbf{y}) - (\hat{\mathbf{x}} + \mathbf{y})^T\mathbf{b} \\
&= \frac{1}{2}\Big(\hat{\mathbf{x}}^T + \mathbf{y}^T\Big)A(\hat{\mathbf{x}} + \mathbf{y}) - \hat{\mathbf{x}}^T\mathbf{b} - \mathbf{y}^T\mathbf{b} \\
&= \frac{1}{2}\Big(\hat{\mathbf{x}}^T A + \mathbf{y}^T A\Big)(\hat{\mathbf{x}} + \mathbf{y}) - \hat{\mathbf{x}}^T\mathbf{b} - \mathbf{y}^T\mathbf{b} \\
&= \frac{1}{2}\Big(\hat{\mathbf{x}}^T A\hat{\mathbf{x}} + \hat{\mathbf{x}}^T A\mathbf{y} + \mathbf{y}^T A\hat{\mathbf{x}} + \mathbf{y}^T A\mathbf{y}\Big) - \hat{\mathbf{x}}^T\mathbf{b} - \mathbf{y}^T\mathbf{b}
\end{aligned}$$

Note that $\hat{\mathbf{x}}^T A\mathbf{y} = \mathbf{y}^T A^T \hat{\mathbf{x}} = \mathbf{y}^T A\hat{\mathbf{x}}$ because A is symmetric

$$\begin{aligned}
&= \frac{1}{2}\Big(\hat{\mathbf{x}}^T A\hat{\mathbf{x}} + 2\mathbf{y}^T A\hat{\mathbf{x}} + \mathbf{y}^T A\mathbf{y}\Big) - \hat{\mathbf{x}}^T\mathbf{b} - \mathbf{y}^T\mathbf{b} \\
&= \frac{1}{2}\hat{\mathbf{x}}^T A\hat{\mathbf{x}} - \hat{\mathbf{x}}^T\mathbf{b} + \mathbf{y}^T A\hat{\mathbf{x}} - \mathbf{y}^T\mathbf{b} + \frac{1}{2}\mathbf{y}^T A\mathbf{y} \\
&= \frac{1}{2}\hat{\mathbf{x}}^T A\hat{\mathbf{x}} - \hat{\mathbf{x}}^T\mathbf{b} + \mathbf{y}^T(A\hat{\mathbf{x}} - \mathbf{b}) + \frac{1}{2}\mathbf{y}^T A\mathbf{y}
\end{aligned}$$

We know that $\nabla\phi(\hat{\mathbf{x}}) \perp K$, therefore $\mathbf{y}^T(A\hat{\mathbf{x}} - \mathbf{b}) = 0$

$$\begin{aligned}
&= \frac{1}{2}\hat{\mathbf{x}}^T A\hat{\mathbf{x}} - \hat{\mathbf{x}}^T\mathbf{b} + \frac{1}{2}\mathbf{y}^T A\mathbf{y} \\
&= \phi(\hat{\mathbf{x}}) + \frac{1}{2}\mathbf{y}^T A\mathbf{y}
\end{aligned}$$

Since A is positive definite $\mathbf{y}^T A\mathbf{y} \geq 0$ and therefore

$$\phi(\mathbf{x}) \geq \phi(\hat{\mathbf{x}})$$

This $\hat{\mathbf{x}}$ minimizes $\phi(\mathbf{x})$ over K.

Now assume that $\hat{\mathbf{x}}$ minimizes $\phi(\mathbf{x})$ over K. □

3.

4. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. Let Gaussian elimination be carried out on A without pivoting. After k steps, A will be reduced to the form

$$A^{(k)} = \begin{pmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{pmatrix}$$

where $A_{22}^{(k)}$ is an $(n-k)\times(n-k)$ matrix. Show by induction

(a) $A_{22}^{(k)}$ is symmetric positive definite.

Proof. □

(b) $a_{ii}^{(k)} \leq a_{ii}^{(k-1)}$ for all $k \leq i \leq n$, $k = 1, \cdots, n-1$.

Proof. □

5. Let $A \in \mathbb{R}^{m\times n}$ with $m > n$ and

$$A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}$$

where A_1 is a nonsingular $n \times n$ matrix, and A_2 is an $(m-n)\times n$ arbitrary matrix.

(a) What is the pseudo-inverse A^+ of A such that $A^+A = I_n$? Express it explicitly in terms of A_1 and A_2.

The pseudo-inverse of A is defined as

$$A^+ = (A^TA)^{-1}A^T$$

Writing this in terms of A_1 and A_2 results in

$$A^+ = \left(\begin{pmatrix} A_1^T & A_2^T \end{pmatrix}\begin{pmatrix} A_1 \\ A_2 \end{pmatrix}\right)^{-1}\begin{pmatrix} A_1^T & A_2^T \end{pmatrix}$$

$$\left(A_1^TA_1 + A_2^TA_2\right)^{-1}\begin{pmatrix} A_1^T & A_2^T \end{pmatrix}$$

(b)

6. Let $A \in \mathbb{C}^{m\times m}$ with $\text{rank}(A) = r$. Suppose an SVD of A is given by $A = U\Sigma V^*$, where $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m$ denote the columns of U and $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m$ denote the columns of V. Prove that $\langle \mathbf{v}_{r+1}, \ldots, \mathbf{v}_m\rangle = (A)$.

Proof. First let $\mathbf{x} \in \langle \mathbf{v}_{r+1}, \ldots, \mathbf{v}_m\rangle$, then $\mathbf{x} = \sum_{i=r+1}^{m}(b_i\mathbf{v}_i)$. If we let $b_i = 0$ for $i = 1, 2, \ldots, r$ and $\mathbf{b} = [b_i]$, then $\mathbf{x} = V\mathbf{b}$. Now consider $A\mathbf{x}$.

$$\begin{aligned} A\mathbf{x} &= U\Sigma V^*V\mathbf{b} \\ &= U\Sigma\mathbf{b} \end{aligned}$$

However Σ is a diagonal matrix with σ_i along the diagonal, so $(\Sigma\mathbf{b})_i = \sigma_i b_i$. Since A has $\text{rank}(A) = r$, we know that $\sigma_i = 0$ for $i \geq r+1$. Therefore if $1 \leq i \leq r$, then $\sigma_i b_i = 0$ because $b_i = 0$. If $r+1 \leq i \leq m$, then $\sigma_i b_i = 0$ because $\sigma_i = 0$. Therefore we can conclude that $\Sigma\mathbf{b} = \mathbf{0}$ Thus $A\mathbf{x} = \mathbf{0}$ and $\mathbf{x} \in (A)$.

Now assume that $\mathbf{x} \in (A)$, that is $A\mathbf{x} = \mathbf{0}$.

$$\begin{aligned} A\mathbf{x} &= \mathbf{0} \\ U\Sigma V * \mathbf{x} &= \mathbf{0} \\ U^*U\Sigma V * \mathbf{x} &= U^*\mathbf{0} \\ \Sigma V * \mathbf{x} &= \mathbf{0} \\ \begin{bmatrix} \sigma_1\mathbf{v}_1^*\mathbf{x} \\ \cdots \\ \sigma_r\mathbf{v}_r^*\mathbf{x} \\ \sigma_{r+1}\mathbf{v}_{r+1}^*\mathbf{x} \\ \cdots \\ \sigma_m\mathbf{v}_m^*\mathbf{x} \end{bmatrix} &= \mathbf{0} \end{aligned}$$

Since $\sigma_i = 0$ for $r+1 \leq i \leq m$

$$\begin{bmatrix} \sigma_1 \mathbf{v}_1^* \mathbf{x} \\ \cdots \\ \sigma_r \mathbf{v}_r^* \mathbf{x} \\ 0 \cdots \\ 0 \end{bmatrix} = \mathbf{0}$$

This implies that $\mathbf{v}_i^* \mathbf{x} = 0$ for $1 \leq i \leq r$ since $\sigma_i > 0$ for $1 \leq i \leq r$. This is equivalent to $\mathbf{x} \perp \mathbf{v}_i$ for $1 \leq i \leq r$. Hence $\mathbf{x} \in \langle \mathbf{v}_1, \cdots, \mathbf{v}_r \rangle^\perp$. Since V is unitary

$$\langle \mathbf{v}_1, \cdots, \mathbf{v}_r \rangle^\perp = \langle \mathbf{v}_{r+1}, \ldots, \mathbf{v}_m \rangle$$

Thus $\mathbf{x} \in \langle \mathbf{v}_{r+1}, \ldots, \mathbf{v}_m \rangle$. □

7. Problem 33.2 (Page 255)
8. Problem 36.1 (Page 283)
9.
10. MATLAB project

    Below are the function for performing the Jacobi method, the Gauss-Seidel Method, and the Conjugate Gradient method.

```
function [x0, k, r] = Jacobi(A, b, tol, maxIter)
    M = diag(diag(A));
    N = M - A;

    x0 = zeros(size(b));
    k = 0;
    r = norm(b - A*x0, inf);
    while(r(end) > tol && k < maxIter)
        k = k + 1;
        x = M\(N*x0) + M\b;
        x0 = x;
        r = [r, norm(b - A*x0, inf)];
    end
end
```

```
function [x0, k, r] = GaussSeidel(A, b, tol, maxIter)
    M = tril(A);
    N = M - A;
```

```
    x0 = zeros(size(b));
    k = 0;
    r = norm(b - A*x0, inf);
    while(r(end) > tol && k < maxIter)
        k = k + 1;
        x = M\(N*x0) + M\b;
        x0 = x;
        r = [r, norm(b - A*x0, inf)];
    end
end
```

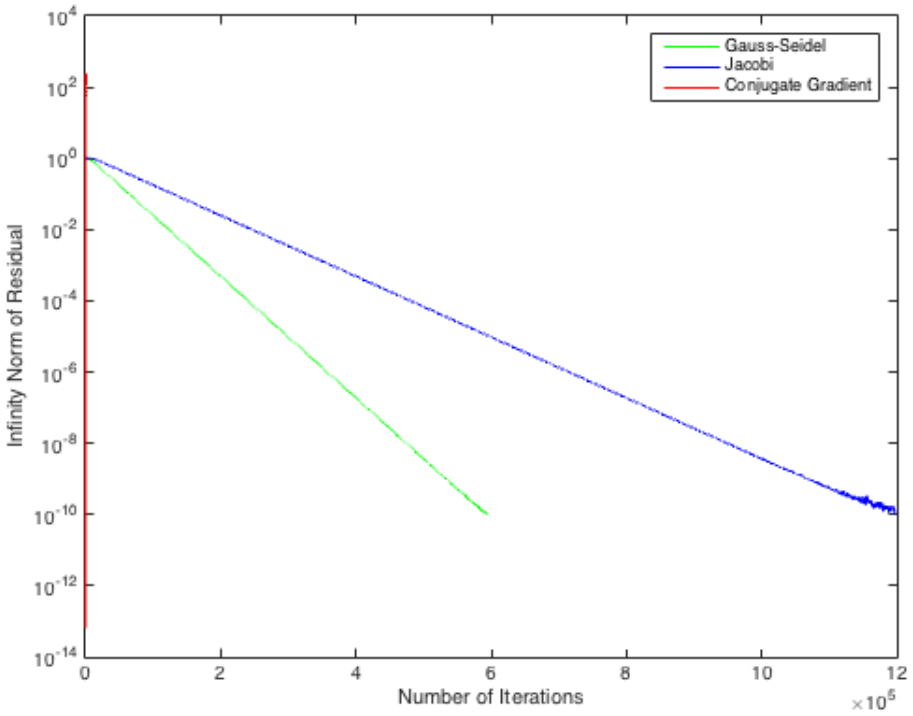
```
function [x0, k, r] = ConjugateGradient(A, b, tol, maxIter)
    x0 = zeros(size(b));
    k = 0;
    r0 = b;
    p0 = r0;
    r = norm(r0, inf);
    while(r(end) > tol && k < maxIter)
        a = (r0'*r0)/(r0'*A*p0);
        x = x0 + a*p0;
        r1 = r0 - a*A*p0;
        bn = (r1'*r1)/(r0'*r0);
        p0 = r1 + bn*p0;

        % move to next step
        k = k+1;
        x0 = x;
        r0 = r1;
        r = [r, norm(r0,inf)];
    end
end
```

The following script uses these three methods to solve the diffusion equation $-u_{xx} = 1$ on $x \in (0, 1)$. It also plots the residual against the number of iterations.

```
%% Problem 10
% Initial matrix
tol = 1e-10;
maxIter = 2e6;
m = 500;
h = 1/(m+1);
e = ones(m, 1);
A = (1/h^2)*spdiags([-e, 2*e, -e], -1:1, m,m);
[xGS, kGS, rGS] = GaussSeidel(A, e, tol, maxIter);
[xJ, kJ, rJ] = Jacobi(A, e, tol, maxIter);
[xCG, kCG, rCG] = ConjugateGradient(A, e, tol, maxIter);
figure;
```

```
semilogy(0:kGS, rGS, 'g', 0:kJ, rJ, 'b', 0:kCG, rCG, 'r');
xlabel('Number of Iterations');
ylabel('Infinity Norm of Residual');
legend('Gauss-Seidel', 'Jacobi', 'Conjugate Gradient');
```



We see in this plot that the Gauss-Seidel method converges much faster than the Jacobi method. Both of these experience linear convergence, but Gauss-Seidel is a faster linear convergence. The Conjugate Gradient method converges much faster than either of the stationary methods. The Conjugate Gradient method converges in only 250 steps which is less than $m = 500$, the size of the matrix.