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, that is for any $\mathbf{x} \in K$, $\phi(\hat{\mathbf{x}}) \leq \phi(\mathbf{x})$.

Let $\mathbf{y} \in K$, then let $F(t) = \phi(\hat{\mathbf{x}} + t\mathbf{y})$. Note that $F(0) = \phi(\hat{\mathbf{x}})$. Since $\hat{\mathbf{x}}$ minimizes ϕ over K, $F(0) \leq F(t)$ for any t. Thus 0 is a minimizer of F and $F'(0) = 0$. Using vector calculus

$$F'(t) = \mathbf{y}^T \nabla\phi(\hat{\mathbf{x}} + t\mathbf{y})$$

Now using $t = 0$ we see that

$$\begin{aligned}
F'(0) &= 0 \\
\mathbf{y}^T \nabla\phi(\hat{\mathbf{x}}) &= 0
\end{aligned}$$

Thus $\nabla\phi(\hat{\mathbf{x}}) \perp K$, because $\mathbf{y}^T \nabla\phi(\hat{\mathbf{x}}) = 0$ for any vector $\mathbf{y} \in K$. □

3. Show that:

   (a) (Forward error analysis)

   $$\Big|fl(\mathbf{x}^T\mathbf{a}) - \mathbf{x}^T\mathbf{a}\Big| \leq n\epsilon_{machine}|\mathbf{x}|^T|\mathbf{a}| + O(\epsilon_{machine}^2)$$

   where $\mathbf{x}$ and $\mathbf{a}$ are n-dimensional floating point vectors and $fl(\mathbf{x}^T\mathbf{a})$ represents the floating point computation of the dot product.

Proof. I will prove by induction. First consider the case when $n = 1$, then

$$|fl(xa) - xa| = |xa(1+\epsilon) - xa|$$

Where $\epsilon = \epsilon_{machine} + O(\epsilon_{machine}^2)$

$$\begin{aligned} &= \epsilon|x||a| \\ &= 1\epsilon_{machine}|x||a| + O(\epsilon_{machine}^2) \end{aligned}$$

Assume that

$$\left|fl(\mathbf{x}^T\mathbf{a}) - \mathbf{x}^T\mathbf{a}\right| \leq n\epsilon_{machine}|\mathbf{x}|^T|\mathbf{a}| + O(\epsilon_{machine}^2)$$

for $n = 1, 2, \ldots, k$. Now consider the case when $n = k+1$. In this case $\mathbf{x} = [\mathbf{x}_k, x_{k+1}]^T$ and $\mathbf{a} = [\mathbf{a}_k, a_{k+1}]^T$.

$$\begin{aligned} \left|fl(\mathbf{x}^T\mathbf{a}) - \mathbf{x}^T\mathbf{a}\right| &= \left|fl(\mathbf{x}_k^T\mathbf{a}_k + v_{k+1}a_{k+1}) - \mathbf{x}_k^T\mathbf{a}_k - v_{k+1}a_{k+1}\right| \\ &= \left|\Big(fl(\mathbf{x}_k^T\mathbf{a}_k) + fl(v_{k+1}a_{k+1})\Big)(1+\epsilon) - \mathbf{x}_k^T\mathbf{a}_k - v_{k+1}a_{k+1}\right| \\ &= \left|fl(\mathbf{x}_k^T\mathbf{a}_k)(1+\epsilon) - \mathbf{x}_k^T\mathbf{a}_k + fl(v_{k+1}a_{k+1})(1+\epsilon) - v_{k+1}a_{k+1}\right| \\ &\leq \left|fl(\mathbf{x}_k^T\mathbf{a}_k)(1+\epsilon) - \mathbf{x}_k^T\mathbf{a}_k\right| + |fl(v_{k+1}a_{k+1})(1+\epsilon) - v_{k+1}a_{k+1}| \\ &\leq k\epsilon_{machine}|\mathbf{x}_k|^T|\mathbf{a}_k| + O(\epsilon_{machine}^2) + |fl(v_{k+1}a_{k+1})(1+\epsilon) - v_{k+1}a_{k+1}| \\ &\leq k\epsilon_{machine}|\mathbf{x}_k|^T|\mathbf{a}_k| + O(\epsilon_{machine}^2) + (\epsilon_{machine})|v_{k+1}||a_{k+1}| + O(\epsilon_{machine}^2) \\ &\leq (k+1)\epsilon_{machine}|\mathbf{x}|^T|\mathbf{a}| + O(\epsilon_{machine}^2) \end{aligned}$$

Thus

$$\left|fl(\mathbf{x}^T\mathbf{a}) - \mathbf{x}^T\mathbf{a}\right| \leq n\epsilon_{machine}|\mathbf{x}|^T|\mathbf{a}| + O(\epsilon_{machine}^2)$$

for all n. □

(b) Show that

$$\|fl(XA) - XA\|_F \leq n\epsilon_{machine}\|X\|_F\|A\|_F + O(\epsilon_{machine}^2)$$

Proof. Each of the entries of $fl(XA)$ will be of the form $fl(\mathbf{x}^T\mathbf{a})$. Therefore

$$\|fl(XA) - XA\|_F = \sqrt{\sum_{i=1}^{n^2}\Big((fl(\mathbf{x}^T\mathbf{a}) - \mathbf{x}^T\mathbf{a})^2\Big)}$$

From part (a)

$$\begin{aligned}\|fl(XA) - XA\|_F &\leq \sqrt{\sum_{i=1}^{n^2}\left(n^2\epsilon_{machine}^2\left(|\mathbf{x}|^T|\mathbf{a}|\right)^2 + O(\epsilon_{machine}^4)\right)} \\ &\leq n\epsilon_{machine}\sqrt{\sum_{i=1}^{n^2}(x^2)}\sqrt{\sum_{i=1}^{n^2}(a^2)} + O(\epsilon_{machine}^2) \\ &\leq n\epsilon_{machine}\|X\|_F\|A\|_F + O(\epsilon_{machine}^2)\end{aligned}$$

□

(c) Show that the relative backward error $\frac{\|\delta A\|_F}{\|A\|_F} \leq n\kappa(A)O(\epsilon_{machine})$

$$\begin{aligned}\frac{\|\delta A\|_F}{\|A\|_F} &= \frac{\|fl(XA) - XA\|_F}{\|A\|_F} \\ &= \frac{\|X(A+\delta A) - XA\|_F}{\|A\|_F} \\ &\leq n\epsilon_{machine}\|X\|_F\|A\|_F\frac{1}{\|A\|_F} \\ &\leq n\epsilon_{machine}\|X\|_F \\ &\leq n\epsilon_{machine}\kappa(X)\end{aligned}$$

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. First I will prove by induction that $A_{22}^{(k)}$ is symmetric. Consider the base case when $k = 0$, in this case $A_{22}^{(0)} = A^{(0)} = A$, and since A is symmetric $A_{22}^{(0)}$ is symmetric. Now assume that $A_{22}^{(k)}$ is symmetric for $k = 1, 2, \ldots, m-1$ with

$m < n$. Consider $A_{22}^{(m)}$. We have from the Gaussian Elimination algorithm that

$$L^{(m)}A^{(m-1)} = A^{(m)}$$

$$\begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & l_{m+1,m} & \ddots & \\ & & \vdots & & \ddots \\ & & l_{n,m} & & 1 \end{bmatrix} \begin{bmatrix} A_{11}^{(m-1)} & A_{12}^{(m-1)} \\ 0 & A_{22}^{(m-1)} \end{bmatrix} = \begin{bmatrix} A_{11}^{(m)} & A_{12}^{(m)} \\ 0 & A_{22}^{(m)} \end{bmatrix}$$

Looking at just the portion changing over this step. I will relabel the matrices as follows.

$$L = \begin{bmatrix} 1 & & & \\ l_{2,1} & \ddots & & \\ \vdots & & \ddots & \\ l_{n-m-1,1} & & & 1 \end{bmatrix}$$
$$B = A_{22}^{(m-1)} = [b_{ij}]$$
$$C = A_{22}^{(m)} = [c_{ij}]$$

Then the following matrix equation holds

$$LB = \begin{bmatrix} b_{11} & \mathbf{b}_{1,2:n-m-1}^T \\ \mathbf{0} & C \end{bmatrix}$$

Note that according to the Gaussian elimination algorithm $l_{i1} = -b_{i1}/b_{11}$. In order to show that C which is $A_{22}^{(m)}$ is symmetric we must show that $c_{ij} = c_{ji}$ for all i and j. We know that c_{ij} is the $(i+1)$ row of L dotted into the $j+1$ column of B.

$$\begin{aligned} c_{ij} &= -l_{i+1,1}b_{1,j+1} + l_{i+1,i+1}b_{i+1,j+1} \\ &= -\frac{b_{i+1,1}}{b_{11}}b_{1,j+1} + 1b_{i+1,j+1} \end{aligned}$$

From our inductive hypothesis B is symmetric so $b_{i,j} = b_{j,i}$

$$\begin{aligned} &= -\frac{b_{1,i+1}}{b_{11}}b_{j+1,1} + 1b_{j+1,i+1} \\ &= -\frac{b_{j+1,1}}{b_{11}}b_{1,i+1} + l_{j+1,j+1}b_{j+1,i+1} \\ &= c_{ji} \end{aligned}$$

Therefore $c_{ij} = c_{ji}$ for all i and j and $C = A_{22}^{(m)}$ is symmetric. Thus by mathematical induction $A_{22}^{(k)}$ is symmetric for all k.

Next I will prove that $A_{22}^{(k)}$ is positive definite. Assume to the contradiction that $A_{22}^{(k)}$ is not positive definite. This implies that there exists a vector $\mathbf{x}$ such that $\mathbf{x}^T A_{22}^{(k)} \mathbf{x} \leq 0$. Construct $\mathbf{y}$ such that $\mathbf{y} \in \mathbb{R}^m$ and the last entries are $\mathbf{x}$ and the first entries are 0. In this case because Gaussian elimination doesn't affect the determinant

$$\begin{aligned} \mathbf{y}^T A \mathbf{y} &= \mathbf{y}^T \begin{bmatrix} A_{11}^{(k)} & A_{12}^{(k)} \\ 0 & A_{22}^{(k)} \end{bmatrix} \mathbf{y} \\ &= \mathbf{x}^T A_{22}^{(k)} \mathbf{x} \\ &\leq 0 \end{aligned}$$

This contradicts the fact that A is positive definite, therefore $A_{22}^{(k)}$ must be positive definite for all k. □

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

Proof. In order to show this we must first find a formula for $a_{ii}^{(k)}$. By the Gaussian elimination algorithm without pivoting

$$a_{ii}^{(k)} = a_{ii}^{(k-1)} - \frac{a_{i1}^{(k-1)}}{a_{11}^{(k-1)}} a_{1i}^{(k-1)}$$

For a symmetric positive definite matrix $a_{i1}^{(k-1)} = a_{1i}^{(k-1)}$, so this can be rewritten as

$$a_{ii}^{(k)} = a_{ii}^{(k-1)} - \frac{\left(a_{i1}^{(k-1)}\right)^2}{a_{11}^{(k-1)}}$$

It is known that $a_{11}^{(k-1)}$ is an eigenvalue of A and since A is positive definite $a_{11}^{(k-1)} > 0$. Also $\left(a_{i1}^{(k-1)}\right)^2 > 0$, therefore $\frac{\left(a_{i1}^{(k-1)}\right)^2}{a_{11}^{(k-1)}} > 0$. This implies that $a_{ii}^{(k)} \leq a_{ii}^{(k-1)}$, as subtracting a positive number makes a number lower. □

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

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

(b) Prove that $\|A^+\|_2 \leq \left\|A_1^{-1}\right\|_2$.

Proof.

$$\left\|A^+\right\|_2 = \sup_{\mathbf{b}} \frac{\|A^+\mathbf{b}\|_2}{\|\mathbf{b}\|_2}$$

Let P be the orthogonal projector onto the range of A, then because P is unitary and the 2-norm is unitarily invariant, $\|\mathbf{b}\|_2 = \|P\mathbf{b}\|_2$. Therefore

$$\left\|A^+\right\|_2 = \sup_{\mathbf{b}} \frac{\|A^+\mathbf{b}\|_2}{\|P\mathbf{b}\|_2}.$$

Now considering the least squares problem $\min_{\mathbf{x}}\|A\mathbf{x} - b\|_2$, we know that $A^+\mathbf{b} = \mathbf{x}$ and $P\mathbf{b} = A\mathbf{x}$. Therefore

$$\left\|A^+\right\|_2 = \sup_{\mathbf{x}} \frac{\|\mathbf{x}\|_2}{\|A\mathbf{x}\|_2}.$$

Note that

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

Therefore $\|A\mathbf{x}\|_2 \geq \|A_1\mathbf{x}\|_2$, so

$$\left\|A^+\right\|_2 \leq \sup_{\mathbf{x}} \frac{\|\mathbf{x}\|_2}{\|A_1\mathbf{x}\|_2}.$$

If we let $\mathbf{y} = A_1\mathbf{x}$, then $\mathbf{x} = A_1^{-1}\mathbf{y}$ and

$$\begin{aligned} \left\|A^+\right\|_2 &\leq \sup_{\mathbf{y}} \frac{\left\|A_1^{-1}\mathbf{y}\right\|_2}{\|\mathbf{y}\|_2} \\ &= \left\|A_1^{-1}\right\|_2 \end{aligned}$$

Thus $\|A^+\|_2 \leq \left\|A_1^{-1}\right\|_2$. □

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) Suppose algorithm 33.1 is executed for a particular A and $\mathbf{b}$ until at some step n, an entry $h_{n+1,n} = 0$ is encountered.

   (a) Show how (33.13) can be simplified in this case. What does this imply about the structure of a full $m \times m$ Hessenberg reduction $A = QHQ^*$ of A?
   This implies that H is
   $$H = \begin{bmatrix} H_n & X \end{bmatrix}$$
   where X can be anything and Q is Q_n extended to an orthonormal basis.

   (b) Show that K_n is an invariant subspace of A, i.e., $AK_n \subseteq K_n$.

   Proof. Let $\mathbf{x} \in AK_n$, then $\mathbf{x} = A\mathbf{y}$ where $\mathbf{y} \in K_n$. If $\mathbf{y} \in K_n$, then $\mathbf{y} = \sum_{i=1}^{n} (y_i \mathbf{q}_i)$. This implies that
   $$\mathbf{x} = \sum_{i=1}^{n} (y_i A\mathbf{q}_i)$$
   For $1 \leq i \leq n-1$, $A\mathbf{q}_i \in K_n$. Consider $A\mathbf{q}_n$, equation 33.4 states
   $$A\mathbf{q}_n = h_{1n}\mathbf{q}_1 + \cdots + h_{nn}\mathbf{q}_n + h_{n+1,n}\mathbf{q}_{n+1}$$
   Since $h_{n+1,n} = 0$
   $$A\mathbf{q}_n = h_{1n}\mathbf{q}_1 + \cdots + h_{nn}\mathbf{q}_n$$
   Thus $A\mathbf{q}_n \in K_n$ as well. This implies that $x = A\mathbf{y} \in K_n$. Thus $AK_n \subseteq K_n$. □

   (c) Show that if the Krylov subspaces of A generated by $\mathbf{b}$ are defined by $K_k = \langle \mathbf{b}, A\mathbf{b}, \ldots, A^{k-1}\mathbf{b} \rangle$, then $K_n = K_{n+1} = K_{n+2}$.

   Proof. In part (b) we showed that $AK_n \subseteq K_n$. Obviously $K_{n+1} = AK_n$, so $K_{n+1} \subseteq K_n$. Clearly from the definition $K_n \subseteq K_{n+1}$ as any vector in $\langle \mathbf{b}, A\mathbf{b}, \ldots, A^{n-1}\mathbf{b} \rangle$ will also be in $\langle \mathbf{b}, A\mathbf{b}, \ldots, A^{n}\mathbf{b} \rangle$. Thus $K_n = K_{n+1}$. This can also be applied for any $m > n$, $K_m = A^{m-n}K_n$, and therefore $K_m \subseteq K_n$. So clearly $K_n = K_m$ for any $m > n$. □

   (d) Show that each eigenvalue of H_n is an eigenvalue of A.
   Let λ be an eigenvalue of H_n, then there exists a nonzero eigenvector $\mathbf{x}$ such that $H_n\mathbf{x} = \lambda\mathbf{x}$. From equation 33.12 we know that $H_n = Q_n^* A Q_n$. This implies that
   $$\begin{aligned} Q_n^* A Q_n \mathbf{x} &= \lambda \mathbf{x} \\ A Q_n \mathbf{x} &= \lambda Q_n \mathbf{x} \end{aligned}$$

Let $\mathbf{y} = Q_n\mathbf{x}$

$$A\mathbf{y} = \lambda\mathbf{y}$$

Thus λ is also an eigenvalue of A.

(e) Show that if A is nonsingular, then the solution $\mathbf{x}$ to the system of equations $A\mathbf{x} = \mathbf{b}$ lies in K_n.

Since the eigenvalues of A are the same as the eigenvalues of H_n, if A is nonsingular than H_n is also nonsingular. Furthermore

$$\begin{aligned} A^{-1} &= (Q_n H_n Q_n^*)^{-1} \\ A^{-1} &= Q_n H_n^{-1} Q_n^* \end{aligned}$$

Therefore

$$\begin{aligned} \mathbf{x} &= A^{-1}\mathbf{b} \\ &= Q_n H_n^{-1} Q_n^* \mathbf{b} \\ &= Q_n\Big(H_n^{-1} Q_n^* \mathbf{b}\Big) \end{aligned}$$

This shows that $\mathbf{x}$ is a linear combination of the columns of Q_n which form a basis of K_n. Therefore $\mathbf{x} \in K_n$.

8. Problem 36.1 (Page 283) In Lecture 27 it was pointed out that the eigenvalues of a symmetric matrix $A \in \mathbb{R}^{m\times m}$ are the stationary values of the Rayleigh quotient $r(\mathbf{x}) = (\mathbf{x}^T A\mathbf{x})/(\mathbf{x}^T\mathbf{x})$ for $\mathbf{x} \in \mathbb{R}^m$. Show that the Ritz values at step n of the Lanczos iteration are the stationary values of $r(\mathbf{x})$ if $\mathbf{x}$ is resticted to K_n.

Proof. The Ritz values are the eigenvalues of T_n the tridiagonal matrix. In this case $A = Q_n^* T_n Q_n$. So if $\mathbf{y} = Q_n\mathbf{x}$ is an eigenvector of T_n, with a corresponding Ritz value, then

$$\begin{aligned} r(\mathbf{x}) &= \frac{\mathbf{x}^T A\mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\ &= \frac{\mathbf{x}^T Q_n^* T_n Q_n \mathbf{x}}{\mathbf{x}^T\mathbf{x}} \\ &= \frac{\mathbf{x}^T Q_n^* T_n Q_n \mathbf{x}}{\mathbf{x}^T Q_n^* Q_n\mathbf{x}} \\ &= \frac{\mathbf{y}^T T_n \mathbf{y}^T}{\mathbf{y}^T\mathbf{y}} \end{aligned}$$

This is the Ritz value associated with $\mathbf{y}$ at the nth step, so the Ritz values are stationary for the Rayleigh quotient. □

9. Let $f : \mathbb{R} \to \mathbb{R}$ be twice continuously differentiable for all x in the neighborhood $\{x \in \mathbb{R} | |x - \xi| < r\}$ of a simple zero ξ of f such that $f(\xi) = 0$. Consider the two-step Newton method:

$$y_k = x_k - f(x_k)/f'(x_k), \quad x_{k+1} = y_k - f(y_k)/f'(x_k).$$

(a) Show that if the method converges, then

$$\lim_{k\to\infty}\left(\frac{x_{k+1} - \xi}{(y_k - \xi)(x_k - \xi)}\right) = \frac{f''(\xi)}{f'(\xi)}$$

Proof. First we assume that the method converges, this implies that

$$\lim_{k\to\infty}(x_k - \xi) = 0$$

and

$$\lim_{k\to\infty}(y_k - \xi) = 0$$

Also the problem will use the Taylor expansion of $f(y_k)$ and $f'(x_k)$ about ξ, so these will be stated here for later use.

$$\begin{aligned} f(y_k) &= f(\xi) + (y_k - \xi)f'(\xi) + O((y_k - \xi)^2) \\ &= (y_k - \xi)(f'(\xi) + O(y_k - \xi)) \\ f'(x_k) &= f'(\xi) + (x_k - \xi)f''(\xi) + O((x_k - \xi)^2) \end{aligned}$$

Now I will consider the limit

$$\begin{aligned}
&\lim_{k\to\infty}\left(\frac{1}{(y_k-\xi)(x_k-\xi)}(x_{k+1}-\xi)\right)\\
&=\lim_{k\to\infty}\left(\frac{1}{(y_k-\xi)(x_k-\xi)}\left(y_k-\xi-\frac{f(y_k)}{f'(x_k)}\right)\right)\\
&=\lim_{k\to\infty}\left(\frac{1}{(y_k-\xi)(x_k-\xi)}\left(y_k-\xi-\frac{(y_k-\xi)(f'(\xi)+O(y_k-\xi))}{f'(x_k)}\right)\right)\\
&=\lim_{k\to\infty}\left(\frac{1}{(x_k-\xi)}\left(1-\frac{f'(\xi)+O(y_k-\xi)}{f'(x_k)}\right)\right)\\
&=\lim_{k\to\infty}\left(\frac{1}{(x_k-\xi)}\left(1-\frac{f'(\xi)+O(y_k-\xi)}{f'(\xi)+(x_k-\xi)f''(\xi)+O((x_k-\xi)^2)}\right)\right)\\
&=\lim_{k\to\infty}\left(\frac{1}{(x_k-\xi)}\left(\frac{f'(\xi)+(x_k-\xi)f''(\xi)+O((x_k-\xi)^2)-f'(\xi)-O(y_k-\xi)}{f'(\xi)+(x_k-\xi)f''(\xi)+O((x_k-\xi)^2)}\right)\right)\\
&=\lim_{k\to\infty}\left(\frac{1}{(x_k-\xi)}\left(\frac{(x_k-\xi)f''(\xi)+O((x_k-\xi)^2)-O(y_k-\xi)}{f'(\xi)+(x_k-\xi)f''(\xi)+O((x_k-\xi)^2)}\right)\right)\\
&=\lim_{k\to\infty}\left(\frac{f''(\xi)+O((x_k-\xi))-O((y_k-\xi)/(x_k-\xi))}{f'(\xi)+(x_k-\xi)f''(\xi)+O((x_k-\xi)^2)}\right)
\end{aligned}$$

Taking the limit results in

$$=\frac{f''(\xi)}{f'(\xi)}$$

□

(b) The convergence is cubic:

$$\lim_{k\to\infty}\left(\frac{x_{k+1}-\xi}{(x_k-\xi)^3}\right)=\frac{1}{2}\left(\frac{f''(\xi)}{f'(\xi)}\right)^2$$

Proof. In this proof I will use the Taylor expansions of $f(x_k)$ and $f'(x_k)$ around ξ, so I will state them here

$$\begin{aligned}
f(x_k)&=f(\xi)+(x_k-\xi)f'(\xi)+O((x_k-\xi)^2)\\
&=(x_k-\xi)(f'(\xi)+O(x_k-\xi))\\
f'(x_k)&=f'(\xi)+(x_k-\xi)f''(\xi)+O((x_k-\xi)^2)
\end{aligned}$$

Now consider the limit

$$\lim_{k\to\infty}\left(\frac{x_{k+1}-\xi}{(x_k-\xi)^3}\right)=\lim_{k\to\infty}\left(\frac{x_{k+1}-\xi}{(y_k-\xi)(x_k-\xi)}\frac{y_k-\xi}{(x_k-\xi)^2}\right)$$

From part (a)

$$= \frac{f''(\xi)}{f'(\xi)} \lim_{k\to\infty} \left(\frac{y_k - \xi}{(x_k - \xi)^2} \right)$$

Now consider this limit individually

$$\lim_{k\to\infty} \left(\frac{y_k - \xi}{(x_k - \xi)^2} \right)$$
$$= \lim_{k\to\infty} \left(\frac{1}{(x_k - \xi)^2} \left(x_k - \xi - \frac{f(x_k)}{f'(x_k)} \right) \right)$$
$$= \lim_{k\to\infty} \left(\frac{1}{(x_k - \xi)^2} \left(x_k - \xi - \frac{(x_k - \xi)\left(f'(\xi) + \frac{1}{2}(x_k - \xi) f''(\xi) + O((x_k - \xi)^2) \right)}{f'(x_k)} \right) \right)$$
$$= \lim_{k\to\infty} \left(\frac{1}{x_k - \xi} \left(1 - \frac{f'(\xi) + \frac{1}{2}(x_k - \xi) f''(\xi) + O((x_k - \xi)^2)}{f'(x_k)} \right) \right)$$
$$= \lim_{k\to\infty} \left(\frac{1}{x_k - \xi} \left(1 - \frac{f'(\xi) + \frac{1}{2}(x_k - \xi) f''(\xi) + O((x_k - \xi)^2)}{f'(\xi) + (x_k - \xi) f''(\xi) + O((x_k - \xi)^2)} \right) \right)$$

By finding a common denominator and adding the fractions

$$= \lim_{k\to\infty} \left(\frac{1}{x_k - \xi} \left(\frac{\frac{1}{2}(x_k - \xi) f''(\xi) + O((x_k - \xi)^2)}{f'(\xi) + (x_k - \xi) f''(\xi) + O((x_k - \xi)^2)} \right) \right)$$
$$= \lim_{k\to\infty} \left(\frac{\frac{1}{2} f''(\xi) + O(x_k - \xi)}{f'(\xi) + (x_k - \xi) f''(\xi) + O((x_k - \xi)^2)} \right)$$

Taking the limit

$$= \frac{1}{2} \frac{f''(\xi)}{f'(\xi)}$$

Now going back to the original limit

$$\lim_{k\to\infty} \left(\frac{x_{k+1} - \xi}{(x_k - \xi)^3} \right) = \frac{f''(\xi)}{f'(\xi)} \lim_{k\to\infty} \left(\frac{y_k - \xi}{(x_k - \xi)^2} \right)$$
$$\lim_{k\to\infty} \left(\frac{x_{k+1} - \xi}{(x_k - \xi)^3} \right) = \frac{1}{2} \left(\frac{f''(\xi)}{f'(\xi)} \right)^2$$

□

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
```

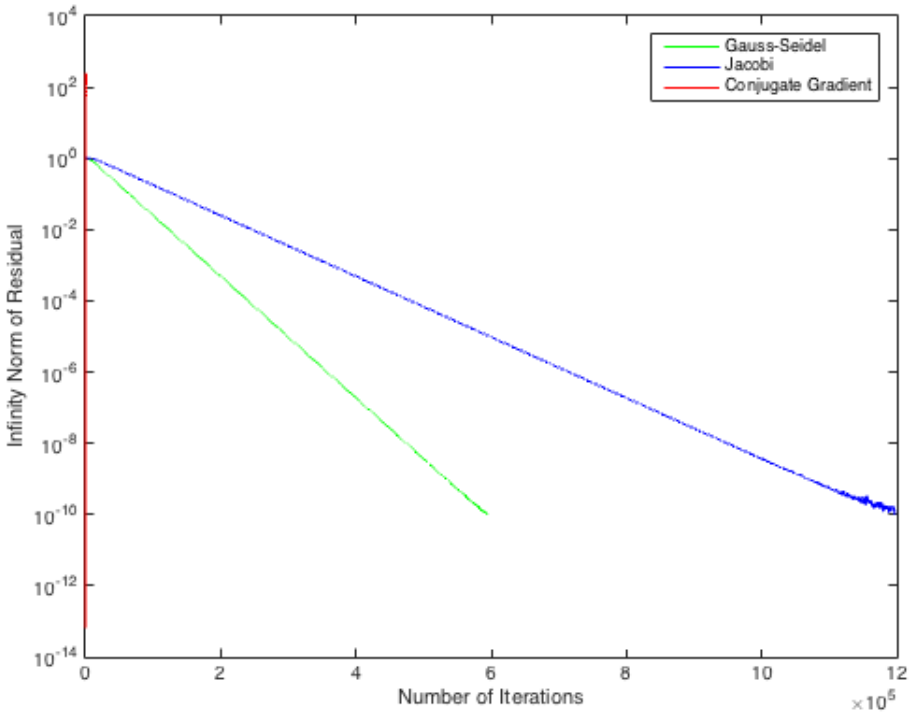
```
function [x0, k, r] = ConjugateGradient(A, b, tol, maxIter)
    x0 = zeros(size(b));
    k = 0;
    r0 = b;
    p0 = r0;
    r = norm(r0, inf);
    while(r(end) > tol && k < maxIter)
        a = (r0'*r0)/(r0'*A*p0);
        x = x0 + a*p0;
        r1 = r0 - a*A*p0;
        bn = (r1'*r1)/(r0'*r0);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