Math 5620 - Introduction to Numerical Analysis - Class Notes

Fernando Guevara Vasquez

Version 1990. Date: January 17, 2012.

Contents

1. Disclaimer 4

Chapter 1. Iterative methods for solving linear systems 5
1. Preliminaries 5
2. Neumann series based methods 7
3. Conjugate gradient and other Krylov subspace methods 15

1. Disclaimer

These notes are work in progress and may be incomplete and contain errors. These notes are based on the following sources:

- Burden and Faires, *Numerical Analysis*, ninth ed.
- Kincaid and Cheney, *Numerical Analysis: Mathematics of scientific computing*
- Stoer and Burlisch, *Introduction to Numerical Analysis*, Springer 1992
- Golub and Van Loan, *Matrix Computations*, John Hopkins 1996

CHAPTER 1

Iterative methods for solving linear systems

1. Preliminaries

1.1. Convergence in $\mathbb{R}^n$. Let $\|\cdot\|$ be a **norm** defined on $\mathbb{R}^n$, i.e. it satisfies the properties:

i. $\forall \mathbf{x} \in \mathbb{R}^n$ $\|\mathbf{x}\| \geq 0$ (non-negativity)
ii. $\|\mathbf{x}\| = 0 \Leftrightarrow \mathbf{x} = \mathbf{0}$ (definiteness)
iii. $\forall \lambda \in \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$ $\|\lambda \mathbf{x}\| = |\lambda| \|\mathbf{x}\|$ (multiplication by a scalar)
iv. $\forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ (triangle inequality)

Some important examples of norms in $\mathbb{R}^n$ are:

(1) $\|\mathbf{x}\|_2 = \left(\sum_{i=1}^n |x_i|^2\right)^{1/2}$ (Euclidean or ℓ_2 norm)
(2) $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$ (ℓ_1 norm)
(3) $\|\mathbf{x}\|_\infty = \max_{i=1,\ldots,n} |x_i|$ (ℓ_∞ or max norm)
(4) $\|\mathbf{x}\|_p = \left(\sum_{i=1}^n |x_i|^p\right)^{1/p}$ (for $p \geq 1$, ℓ_p norm)

A sequence $\{\mathbf{v}^{(k)}\}_{k=1}^\infty$ in $\mathbb{R}^n$ **converges** to $\mathbf{v} \in \mathbb{R}^n$ if and only if

$$\lim_{k\to\infty} \|\mathbf{v}^{(k)} - \mathbf{v}\|. \tag{1}$$

Since $\mathbb{R}^n$ is a finite dimensional vector space the notion of convergence is independent of the norm. This follows from the fact that in a finite dimensional vector space all norms are **equivalent**. Two norms $\|\cdot\|$ and $|||\cdot|||$ are equivalent if there are constants $\alpha, \beta > 0$ such that

$$\forall \mathbf{x}\ \alpha |||\mathbf{x}||| \leq \|\mathbf{x}\| \leq \beta |||\mathbf{x}|||. \tag{2}$$

Another property of $\mathbb{R}^n$ is that it is **complete**, meaning that all Cauchy sequences converge in $\mathbb{R}^n$. A sequence $\{\mathbf{v}^{(k)}\}_{k=1}^\infty$ is said to be a **Cauchy sequence** when

$$\forall \epsilon > 0\ \exists N\ \forall i, j \geq N\ \|\mathbf{v}^{(i)} - \mathbf{v}^{(j)}\| < \epsilon. \tag{3}$$

In English: A Cauchy sequence is a sequence for which any two iterates can be made as close as we want provided that we are far enough in the sequence.

1.2. Induced matrix norms. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a matrix (this notation means that the matrix has n rows and n columns) and let $\|\cdots\|$ be a norm on $\mathbb{R}^n$. The operator or induced matrix norm of $\mathbf{A}$ is defined by:

$$\|\mathbf{A}\| = \sup_{\mathbf{x}\in\mathbb{R}^n, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\|\mathbf{x}\|=1} \|\mathbf{A}\mathbf{x}\|. \tag{4}$$

The induced matrix norm measures what is the maximum dilation of the image of a vector through the matrix $\mathbf{A}$. It is a good exercise to show that it satisfies the axiom of a norm.

Some important examples of induced matrix norms:

(1) $\|\mathbf{A}\|_1 = \max_{j=1,\dots,n} \|\mathbf{c}_j\|_1$, where $\mathbf{c}_j$ is the $j-$th column of $\mathbf{A}$.
(2) $\|\mathbf{A}\|_\infty = \max_{i=1,\dots,n} \|\mathbf{r}_i\|_1$, where $\mathbf{r}_i$ is the $i-$th row of $\mathbf{A}$.
(3) $\|\mathbf{A}\|_2$ = square root of largest eigenvalue of $\mathbf{A}^T\mathbf{A}$. When $\mathbf{A}$ is symmetric we have $\|\mathbf{A}\|_2 = |\lambda_{max}|$, with λ_{max} being the largest eigenvalue of $\mathbf{A}$ in magnitude.

A matrix norm that is not an induced matrix norm is the Frobenius norm:

$$\|\mathbf{A}\|_F = \left(\sum_{i,j=1}^{n} |a_{ij}|^2\right)^{1/2}. \tag{5}$$

Some properties of induced matrix norms:

(1) $\|\mathbf{A}\mathbf{x}\| \leq \|\mathbf{A}\|\,\|\mathbf{x}\|$.
(2) $\|\mathbf{A}\mathbf{B}\| \leq \|\mathbf{A}\|\,\|\mathbf{B}\|$.
(3) $\|\mathbf{A}^k\| \leq \|\mathbf{A}\|^k$.

1.3. Eigenvalues. The **eigenvalues** of a $n\times n$ matrix $\mathbf{A}$ are the roots of the characteristic polynomial

$$p(\lambda) = \det(\lambda\mathbf{I} - \mathbf{A}). \tag{6}$$

This is a polynomial of degree n, so it has at most n complex roots. An **eigenvector v** associated with an eigenvalue λ is a nonzero vector such that $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$. Sometimes it is convenient to refer to an eigenvalue λ and corresponding eigenvector $\mathbf{v} \neq \mathbf{0}$ as an **eigenpair** of $\mathbf{A}$.

The **spectral radius** $\rho(\mathbf{A})$ is defined as the magnitude of the largest eigenvalue in magnitude of a matrix $\mathbf{A}$ i.e.

$$\rho(\mathbf{A}) = \max\{|\lambda| \mid \det(\lambda\mathbf{I} - \mathbf{A}) = 0\}. \tag{7}$$

The spectral radius is the radius of the smallest circle in $\mathbb{C}$ containing all eigenvalues of $\mathbf{A}$.

Two matrices $\mathbf{A}$ and $\mathbf{B}$ are said to be similar if there is an invertible matrix $\mathbf{X}$ such that

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{B}. \tag{8}$$

Similar matrices have the same characteristic polynomial and thus the same eigenvalues. This follows from the properties of the determinant since

$$\det(\lambda\mathbf{I}-\mathbf{A}) = \det(\mathbf{X}^{-1})\det(\lambda\mathbf{I}-\mathbf{A})\det(\mathbf{X}) = \det(\lambda\mathbf{I}-\mathbf{X}^{-1}\mathbf{A}\mathbf{X}) = \det(\lambda\mathbf{I}-\mathbf{B}). \tag{9}$$

2. Neumann series based methods

2.1. Neumann series. We start by a fundamental theorem that is the theoretical basis for several methods for solving the linear system $\mathbf{Ax} = \mathbf{b}$. This theorem can be seen as a generalization to matrices of the geometric series identity $(1-x)^{-1} = 1 + x + x^2 + \dots$ for $|x| < 1$.

THEOREM 1. *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be such that* $\|\mathbf{A}\| < 1$ *for some induced matrix norm* $\|\cdot\|$, *then:*

i. $\mathbf{I}-\mathbf{A}$ *is invertible*
ii. $(\mathbf{I}-\mathbf{A})^{-1} = \mathbf{I}+\mathbf{A}+\mathbf{A}^2+\cdots = \sum_{k=0}^{\infty}\mathbf{A}^k$.

PROOF. Assume for contradiction that $\mathbf{I}-\mathbf{A}$ is singular. This means there is a $\mathbf{x} \neq \mathbf{0}$ such that $(\mathbf{I}-\mathbf{A})\mathbf{x} = \mathbf{0}$. Taking $\mathbf{x}$ such that $\|\mathbf{x}\| = 1$, we have

$$1 = \|\mathbf{x}\| = \|\mathbf{Ax}\| \le \|\mathbf{A}\|\,\|\mathbf{x}\| = \|\mathbf{A}\|, \tag{10}$$

which contradicts the hypothesis $\|\mathbf{A}\| < 1$.

We now need to show convergence to $(\mathbf{I}-\mathbf{A})^{-1}$ of the partial series

$$\sum_{k=0}^{m}\mathbf{A}^k. \tag{11}$$

Observe that:

$$(\mathbf{I}-\mathbf{A})\sum_{k=0}^{m}\mathbf{A}^k = \sum_{k=0}^{m}\mathbf{A}^k - \mathbf{A}^{k+1} = \mathbf{A}^0 - \mathbf{A}^{m+1}. \tag{12}$$

Therefore

$$\|(\mathbf{I}-\mathbf{A})\sum_{k=0}^{m}\mathbf{A}^k - \mathbf{I}\| = \|\mathbf{A}^{m+1}\| \le \|\mathbf{A}\|^{m+1} \to 0 \text{ as } m\to\infty. \tag{13}$$

□

Here is an application of this theorem to estimate the norm

$$\|(\mathbf{I}-\mathbf{A})^{-1}\| \le \sum_{k=0}^{\infty}\|\mathbf{A}^k\| \le \sum_{k=0}^{\infty}\|\mathbf{A}\|^k = \frac{1}{1-\|\mathbf{A}\|}. \tag{14}$$

Here is a generalization of the Neumann series theorem.

THEOREM 2. *If* $\mathbf{A}$ *and* $\mathbf{B}$ *are* $n\times n$ *matrices such that* $\|\mathbf{I}-\mathbf{AB}\| < 1$ *for some induced matrix norm then*

i. $\mathbf{A}$ *and* $\mathbf{B}$ *are invertible.*

ii. The inverses are:

$$\begin{aligned} \mathbf{A}^{-1} &= \mathbf{B}\sum_{k=0}^{\infty}(\mathbf{I}-\mathbf{AB})^k, \\ \mathbf{B}^{-1} &= \left[\sum_{k=0}^{\infty}(\mathbf{I}-\mathbf{AB})^k\right]\mathbf{A}. \end{aligned} \tag{15}$$

PROOF. By using the Neumann series theorem, $\mathbf{AB} = \mathbf{I} - (\mathbf{I} - \mathbf{AB})$ is invertible and

$$(\mathbf{AB})^{-1} = \sum_{k=0}^{\infty}(\mathbf{I}-\mathbf{AB})^k. \tag{16}$$

Therefore

$$\begin{aligned} \mathbf{A}^{-1} &= \mathbf{B}(\mathbf{AB})^{-1} = \mathbf{B}\sum_{k=0}^{\infty}(\mathbf{I}-\mathbf{AB})^k, \\ \mathbf{B}^{-1} &= (\mathbf{AB})^{-1}\mathbf{A} = \left[\sum_{k=0}^{\infty}(\mathbf{I}-\mathbf{AB})^k\right]\mathbf{A}. \end{aligned} \tag{17}$$

$\square$

2.2. Iterative refinement. Let $\mathbf{A}$ be an invertible matrix. The iterative refinement method is a method for generating successively better approximations to the solution of the linear system $\mathbf{Ax} = \mathbf{b}$. Assume we have an invertible matrix $\mathbf{B}$ such that $\mathbf{x} = \mathbf{Bb} \approx \mathbf{A}^{-1}\mathbf{b}$ and applying $\mathbf{B}$ is much cheaper than applying solving a system with the matrix $\mathbf{A}$ (we shall see how good the approximation needs to be later). This approximate inverse $\mathbf{B}$ may come for example from an incomplete LU factorization or from running a few steps of an iterative method to solve $\mathbf{Ax} = \mathbf{b}$. Can we use successively refine the approximations given by this method? The idea is to look at the iteration

$$\begin{aligned} \mathbf{x}^{(0)} &= \mathbf{Bb} \\ \mathbf{x}^{(k)} &= \mathbf{x}^{(k-1)} + \mathbf{B}(\mathbf{b} - \mathbf{Ax}^{(k-1)}). \end{aligned} \tag{18}$$

If this iteration converges, then the limit must satisfy

$$\mathbf{x} = \mathbf{x} + \mathbf{B}(\mathbf{b} - \mathbf{Ax}), \tag{19}$$

i.e. if the method converges it converges to a solution of $\mathbf{Ax} = \mathbf{b}$.

THEOREM 3. *The iterative refinement method* (18) *generates iterates of the form*

$$\mathbf{x}^{(m)} = \mathbf{B}\sum_{k=0}^{m}(\mathbf{I}-\mathbf{AB})^k\mathbf{b},\ m \ge 0. \tag{20}$$

Thus by the generalized Neumann series theorem, the method converges to the solution of $\mathbf{Ax} = \mathbf{b}$ *provided* $\|\mathbf{I}-\mathbf{AB}\| < 1$ *for some induced matrix norm (i.e. provided* $\mathbf{B}$ *is sufficiently close to being an inverse of* $\mathbf{A}$*).*

PROOF. We show the form of the iterates in the iterative refinement method by induction on m. First the case $m = 0$ is trivial since $\mathbf{x}^(0) = \mathbf{Bb}$. Assuming the $m-$th case holds:

$$\begin{aligned}
\mathbf{x}^{(m+1)} &= \mathbf{x}^{(m)} + \mathbf{B}(\mathbf{b} - \mathbf{Ax}^{(m)}) \\
&= \mathbf{B}\sum_{k=0}^{m}(\mathbf{I}-\mathbf{AB})^k\mathbf{b} + \mathbf{Bb} - \mathbf{AB}\sum_{k=0}^{m}(\mathbf{I}-\mathbf{AB})^k\mathbf{b} \\
&= \mathbf{B}\left[\mathbf{b} + (\mathbf{I}-\mathbf{AB})\sum_{k=0}^{m}(\mathbf{I}-\mathbf{AB})^k\mathbf{b}\right] \\
&= \mathbf{B}\sum_{k=0}^{m+1}(\mathbf{I}-\mathbf{AB})^k\mathbf{b}.
\end{aligned} \tag{21}$$

□

2.3. Matrix splitting methods. In order to solve the linear system $\mathbf{Ax} = \mathbf{b}$, we introduce a **splitting matrix Q** and use it to define the iteration:

$$\begin{aligned}
\mathbf{x}^{(0)} &= \text{given}, \\
\mathbf{Qx}^{(k)} &= (\mathbf{Q}-\mathbf{A})\mathbf{x}^{(k-1)} + \mathbf{b},\ k \geq 1.
\end{aligned} \tag{22}$$

Since we need to solve for $\mathbf{x}^{(k)}$ the matrix $\mathbf{Q}$ needs to be invertible and solving systems with $\mathbf{Q}$ needs to be a cheap operation (for example $\mathbf{Q}$ could be diagonal or triangular). If the iteration converges, the limit $\mathbf{x}$ must satisfy

$$\mathbf{Qx} = (\mathbf{Q}-\mathbf{A})\mathbf{x} + \mathbf{b}. \tag{23}$$

In other words: if the iteration (22) converges, the limit solves the linear system $\mathbf{Ax} = \mathbf{b}$. The next theorem gives a sufficient condition for convergence.

THEOREM 4. *If* $\|\mathbf{I}-\mathbf{Q}^{-1}\mathbf{A}\| < 1$ *for some matrix induced norm, then the iterates* (22) *converge to the solution to* $\mathbf{Ax} = \mathbf{b}$ *regardless of the initial iterate* $\mathbf{x}^{(0)}$.

PROOF. Subtracting the equations

$$\begin{aligned}
\mathbf{x}^{(k)} &= (\mathbf{I}-\mathbf{Q}^{-1}\mathbf{A})\mathbf{x}^{(k-1)} + \mathbf{Q}^{-1}\mathbf{b} \\
\mathbf{x} &= (\mathbf{I}-\mathbf{QA})\mathbf{x} + \mathbf{Q}^{-1}\mathbf{b},
\end{aligned} \tag{24}$$

we obtain a relation between the error at step k and the error at step $k-1$:

$$\mathbf{x}^{(k)} - \mathbf{x} = (\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A})(\mathbf{x}^{(k-1)} - \mathbf{x}). \tag{25}$$

Taking norms we get:

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \|\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A}\| \|\mathbf{x}^{(k-1)} - \mathbf{x}\| \le \ldots \le \|\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A}\|^k \|\mathbf{x}^{(0)} - \mathbf{x}\|. \tag{26}$$

Thus if $\|\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A}\| < 1$ we have $\mathbf{x}^{(k)} \to \mathbf{x}$ as $k \to \infty$. □

As a stopping criterion we can look at the difference between two consecutive iterates $\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\|$. To see this, let $\delta = \|\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A}\| < 1$. Then by the proof of the previous theorem we must have

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \delta \|\mathbf{x}^{(k-1)} - \mathbf{x}\| \le \delta(\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\| + \|\mathbf{x}^{(k)} - \mathbf{x}\|). \tag{27}$$

Hence by isolating $\|\mathbf{x}^{(k)} - \mathbf{x}\|$ we can bound the error by the difference between two consecutive iterates:

$$\|\mathbf{x}^{(k)} - \mathbf{x}\| \le \frac{\delta}{1-\delta} \|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k)}\|. \tag{28}$$

Of course there can be issues if δ is very close to 1.

We now look at examples of matrix splitting methods. Let us first introduce a standard notation for partitioning the matrix $\mathbf{A}$ into its diagonal elements $\mathbf{D}$, strictly lower triangular part $-\mathbf{E}$ and strictly upper triangular part $-\mathbf{F}$ so that

$$\mathbf{A} = \mathbf{D} - \mathbf{E} - \mathbf{F}. \tag{29}$$

2.3.1. *Richardson method.* Here the splitting matrix is $\mathbf{Q} = \mathbf{I}$, so the iteration is

$$\mathbf{x}^{(k)} = (\mathbf{I} - \mathbf{A})\mathbf{x}^{(k-1)} + \mathbf{b} = \mathbf{x}^{(k-1)} + \mathbf{r}^{(k-1)}, \tag{30}$$

where the residual vector is $\mathbf{r} = \mathbf{b} - \mathbf{A}\mathbf{x}$. Using the theorem on convergence of splitting methods, we can expect convergence when $\|\mathbf{I} - \mathbf{A}\| < 1$ in some matrix induced norm, or in other words if the matrix $\mathbf{A}$ is sufficiently close to the identity.

2.3.2. *Jacobi method.* Here the splitting matrix is $\mathbf{Q} = \mathbf{D}$, so the iteration is

$$\mathbf{D}\mathbf{x}^{(k)} = (\mathbf{E} + \mathbf{F})\mathbf{x}^{(k-1)} + \mathbf{B}. \tag{31}$$

We can expect convergence when $\|\mathbf{I} - \mathbf{D}^{-1}\mathbf{A}\| < 1$ for some matrix induced norm. If we choose the $\|\cdot\|_\infty$ norm, we can get an easy to check sufficient condition for convergence of the Jacobi method. Indeed:

$$\mathbf{D}^{-1}\mathbf{A} = \begin{bmatrix} 1 & a_{12}/a_{11} & a_{13}/a_{11} & \ldots & a_{1n}/a_{11} \\ a_{21}/a_{22} & 1 & a_{23}/a_{22} & \ldots & a_{2n}/a_{22} \\ \vdots & & & & \vdots \\ a_{n1}/a_{nn} & a_{n2}/a_{nn} & \ldots & a_{nn-1}/a_{nn} & 1 \end{bmatrix}. \tag{32}$$

Hence

$$\|\mathbf{I}-\mathbf{Q}^{-1}\mathbf{A}\|_\infty = \max_{i=1,\dots,n} \sum_{j=1, j\neq i}^{n} \frac{|a_{ij}|}{|a_{ii}|}. \tag{33}$$

A matrix satisfying the condition $\|\mathbf{I}-\mathbf{Q}^{-1}\mathbf{A}\|_\infty < 1$ is said to be **diagonally dominant** since it is equivalent to saying that in every row the diagonal element is larger than the sum of all the other ones (in magnitude)

$$|a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}|, \text{ for } i = 1, \dots, n. \tag{34}$$

This result can be summarized as a theorem:

THEOREM 5. *If* **A** *is diagonally dominant, then the Jacobi method converges regardless of the initial iterate to the solution of* $\mathbf{Ax} = \mathbf{b}$.

We emphasize that this is only a sufficient condition for convergence. The Jacobi method may converge for matrices that are not diagonally dominant.

The pseudocode for the Jacobi algorithm is

```
for k = 1,2,...
  x = x + (b-A*x)./diag(diag(A))
end
```

Each iteration involves a multiplication by **A** and division by the diagonal elements of **A**.

2.3.3. *Gauss-Seidel method.* Here $\mathbf{Q} = \mathbf{D} - \mathbf{E}$, i.e. the lower triangular part of **A**. The iterates are:

$$(\mathbf{D}-\mathbf{E})\mathbf{x}^{(k)} = \mathbf{F}\mathbf{x}^{(k-1)} + \mathbf{b}. \tag{35}$$

Each iteration involves multiplication by the strictly upper triangular part of **A** and solving a lower triangular system (forward substitution). Here is an easy to check sufficient condition for convergence.

THEOREM 6. *If* **A** *is diagonally dominant, then the Gauss-Seidel method converges regardless of the initial iterate to the solution of* $\mathbf{Ax} = \mathbf{b}$.

The proof of this theorem is deferred to later, when we will find a necessary and sufficient condition for convergence of matrix splitting methods. Gauss-Seidel usually outperforms the Jacobi method.

2.3.4. *Successive Over Relaxation (SOR) method.* Here $\mathbf{Q} = \omega^{-1}(\mathbf{D} - \omega\mathbf{E})$ and ω is a parameter that needs to be chosen ahead of time. For symmetric positive definite matrices choosing $\omega \in (0,2)$ gives convergence. The iterates are:

$$(\mathbf{D}-\omega\mathbf{E})\mathbf{x}^{(k)} = \omega(\mathbf{F}\mathbf{x}^{(k-1)} + \mathbf{b}) + (1-\omega)\mathbf{D}\mathbf{x}^{(k-1)}. \tag{36}$$

The cost of an iteration is similar to that of Gauss-Seidel and with a good choice of the relaxation parameter ω, SOR can outperform Gauss-Seidel.

2.4. Convergence of iterative methods. The goal of this section is to give a sufficient and necessary condition for the convergence of the iteration

$$\mathbf{x}^{(k)} = \mathbf{G}\mathbf{x}^{(k-1)} + \mathbf{c}. \tag{37}$$

The matrix splitting methods with iteration $\mathbf{Q}\mathbf{x}^{(k)} = (\mathbf{Q}-\mathbf{A})\mathbf{x}^{(k-1)} + \mathbf{b}$ from previous section can be written in the form (37) with $\mathbf{G} = (\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A})$ and $\mathbf{c} = \mathbf{Q}^{-1}\mathbf{b}$.

If the iteration (37) converges its limit satisfies

$$\mathbf{x} = \mathbf{G}\mathbf{x} + \mathbf{c}, \tag{38}$$

that is $\mathbf{x} = (\mathbf{I}-\mathbf{G})^{-1}\mathbf{c}$, assuming the matrix $\mathbf{I}-\mathbf{G}$ is invertible. We will show the following theorem.

THEOREM 7. *The iteration* $\mathbf{x}^{(k)} = \mathbf{G}\mathbf{x}^{(k-1)} + \mathbf{c}$ *converges to* $(\mathbf{I}-\mathbf{G})^{-1}\mathbf{c}$ *if and only if* $\rho(\mathbf{G}) < 1$.

To prove theorem 7 we need the following result.

THEOREM 8. *The spectral radius satisfies:*

$$\rho(\mathbf{A}) = \inf_{\|\cdot\|} \|\mathbf{A}\|, \tag{39}$$

where the inf *is taken over all induced matrix norms.*

This theorem means that the smallest possible induced matrix norm is the 2−norm, if the matrix $\mathbf{A}$ is symmetric. The proof of this theorem is deferred to the end of this section. Let us first prove theorem 7.

PROOF OF THEOREM 7. We first show that $\rho(\mathbf{G}) < 1$ is sufficient for convergence. Indeed if $\rho(\mathbf{G}) < 1$, then there is an induced matrix norm $\|\cdot\|$ for which $\|\mathbf{G}\| < 1$. The iterates (37) are:

$$\begin{aligned}
\mathbf{x}^{(1)} &= \mathbf{G}\mathbf{x}^{(0)} + \mathbf{c}\\
\mathbf{x}^{(2)} &= \mathbf{G}^2\mathbf{x}^{(0)} + \mathbf{G}\mathbf{c} + \mathbf{c}\\
\mathbf{x}^{(3)} &= \mathbf{G}^3\mathbf{x}^{(0)} + \mathbf{G}^2\mathbf{c} + \mathbf{G}\mathbf{c} + \mathbf{c}\\
&\vdots\\
\mathbf{x}^{(k)} &= \mathbf{G}^k\mathbf{x}^{(0)} + \sum_{j=0}^{k-1}\mathbf{G}^j\mathbf{c}.
\end{aligned} \tag{40}$$

The term involving the initial guess goes to zero as $k \to \infty$ because

$$\|\mathbf{G}^k\mathbf{x}^{(0)}\| \le \|\mathbf{G}\|^k\|\mathbf{x}^{(0)}\|. \tag{41}$$

Thus by the Neumann series theorem:

$$\sum_{j=0}^{k-1} \mathbf{G}^j \mathbf{c} \to (\mathbf{I}-\mathbf{G})^{-1}\mathbf{c} \text{ as } k \to \infty. \tag{42}$$

Now we need to show that $\rho(\mathbf{G}) < 1$ is necessary for convergence. Assume $\rho(\mathbf{G}) \geq 1$ and let $\lambda, \mathbf{u}$ be an eigenpair of $\mathbf{G}$ for which $|\lambda| \geq 1$. Taking $\mathbf{x}^{(0)} = \mathbf{0}$ and $\mathbf{c} = \mathbf{u}$ we get:

$$\mathbf{x}^{(k)} = \sum_{j=0}^{k-1} \mathbf{G}^j \mathbf{u} = \sum_{j=0}^{k-1} \lambda^j \mathbf{u} = \begin{cases} k\mathbf{u} & \text{if } \lambda = 1 \\ \frac{1-\lambda^k}{1-\lambda}\mathbf{u} & \text{if } \lambda \neq 1. \end{cases} \tag{43}$$

This is an example of an iteration of the form (37) that does not converge when $\rho(\mathbf{G}) \geq 1$. □

Theorem 8 applied to splitting matrix methods gives:

COROLLARY 1. *The iteration* $\mathbf{Q}\mathbf{x}^{(k)} = (\mathbf{Q}-\mathbf{A})\mathbf{x}^{(k-1)} + \mathbf{b}$ *converges to* $\mathbf{A}\mathbf{x} = \mathbf{b}$ *for any initial guess* $\mathbf{x}^{(0)}$ *if and only if* $\rho(\mathbf{I}-\mathbf{Q}^{-1}\mathbf{A}) < 1$.

In order to show theorem 8 we need the following result:

THEOREM 9. *Let* $\mathbf{A}$ *be a* $n \times n$ *matrix. There is a similarity transformation* $\mathbf{X}$ *such that*

$$\mathbf{A}\mathbf{X} = \mathbf{X}\mathbf{B} \tag{44}$$

where $\mathbf{B}$ *is an upper triangular matrix with off-diagonal components that can be made arbitrarily small.*

PROOF. By the Schur factorization any matrix $\mathbf{A}$ is similar through an unitary transformation $\mathbf{Q}$ to an upper triangular matrix $\mathbf{T}$

$$\mathbf{A} = \mathbf{Q}\mathbf{T}\mathbf{Q}^T, \text{ with } \mathbf{Q}^T\mathbf{Q} = \mathbf{I}. \tag{45}$$

Let $\mathbf{D} = \operatorname{diag}(\epsilon, \epsilon^2, \ldots, \epsilon^n)$. Then

$$(\mathbf{D}^{-1}\mathbf{T}\mathbf{D})_{ij} = t_{ij}\epsilon^{j-i}. \tag{46}$$

The elements below the diagonal ($j < i$) are zero. Those above the diagonal ($j > i$) satisfy

$$|t_{ij}\epsilon^{j-i}| \leq \epsilon |t_{ij}|. \tag{47}$$

With $\mathbf{X} = \mathbf{Q}\mathbf{D}$, the matrix $\mathbf{A}$ is similar to $\mathbf{B} = \mathbf{D}^{-1}\mathbf{T}\mathbf{D}$, and $\mathbf{B}$ is upper triangular with off-diagonal elements that can be made arbitrarily small. □

PROOF OF THEOREM 8. We start by proving that $\rho(\mathbf{A}) \leq \inf_{\|\cdot\|} \|\mathbf{A}\|$. Pick a vector norm $\|\cdot\|$ and let $\lambda, \mathbf{x}$ be an eigenpair of $\mathbf{A}$ with $\|\mathbf{x}\| = 1$. Then

$$\|\mathbf{A}\| \geq \|\mathbf{A}\mathbf{x}\| = \|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|. \tag{48}$$

Since this is true for all eigenvalues λ of $\mathbf{A}$ we must have $\rho(\mathbf{A}) \le \|\mathbf{A}\|$. Since this is true for all induced matrix norms, it must also be true for their inf, i.e. $\rho(\mathbf{A}) \le \inf_{\|\cdot\|} \|\mathbf{A}\|$.

We now show the reverse inequality $\rho(\mathbf{A}) \le \inf_{\|\cdot\|} \|\mathbf{A}\|$. By theorem 9, for any $\epsilon > 0$, there is a non-singular matrix $\mathbf{X}$ such that $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathbf{D} + \mathbf{T}$, where $\mathbf{D}$ is diagonal and $\mathbf{T}$ is strictly upper triangular with $\|\mathbf{T}\|_\infty \le \epsilon$ (why does the component by component result from 9 imply this inequality with the induced matrix norm?). Therefore:

$$\|\mathbf{X}^{-1}\mathbf{A}\mathbf{X}\|_\infty = \|\mathbf{D}+\mathbf{T}\|_\infty \le \|\mathbf{D}\|_\infty + \|\mathbf{T}\|_\infty \le \rho(\mathbf{A}) + \epsilon. \tag{49}$$

It is possible to show that the norm $\|\mathbf{A}\|'_\infty \equiv \|\mathbf{X}^{-1}\mathbf{A}\mathbf{X}\|_\infty$ is an induced matrix norm. Hence

$$\inf_{\|\cdot\|} \|\mathbf{A}\| \le \|\mathbf{A}\|'_\infty \le \rho(\mathbf{A}) + \epsilon. \tag{50}$$

Since $\epsilon > 0$ is arbitrary, we have $\rho(\mathbf{A}) \le \inf_{\|\cdot\|} \|\mathbf{A}\|$. □

2.4.1. *Convergence of Gauss-Seidel method.* As an application of the general theory above, we will determine that the Gauss-Seidel method converges when the matrix is diagonally dominant. Again, this is only a sufficient condition for convergence, the Gauss-Seidel method may converge for other matrices that are not diagonally dominant.

THEOREM 10. *If $\mathbf{A}$ is diagonally dominant then the Gauss-Seidel method converges for any initial guess $\mathbf{x}^{(0)}$.*

PROOF. We need to show that when $\mathbf{A}$ is diagonally dominant we have $\rho(\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A}) < 1$. Let $\lambda, \mathbf{x}$ be an eigenpair of $\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A}$ with $\|\mathbf{x}\|_\infty = 1$. Then $(\mathbf{I} - \mathbf{Q}^{-1}\mathbf{A})\mathbf{x} = \lambda\mathbf{x}$, or equivalently $(\mathbf{Q} - \mathbf{A})\mathbf{x} = \lambda\mathbf{Q}\mathbf{x}$. Written componentwise this becomes:

$$-\sum_{j=i+1}^{n} a_{ij}x_j = \lambda \sum_{j=1}^{i} a_{ij}x_j, \; 1 \le i \le n. \tag{51}$$

Isolating the diagonal component:

$$\lambda a_{ii}x_i = -\lambda \sum_{j=1}^{i-1} - \sum_{j=i+1}^{n} a_{ij}x_j, \; 1 \le i \le n. \tag{52}$$

Now pick the index i such that $|x_i| = 1$ and write

$$|\lambda||a_{ii}| \le |\lambda| \sum_{j=1}^{i-1} |a_{ij}| + \sum_{j=i+1}^{n} |a_{ij}|. \tag{53}$$

Isolating for λ and using the diagonal dominance of $\mathbf{A}$ we obtain

$$|\lambda| \le \frac{\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}|} < 1. \tag{54}$$

We conclude by noticing that this holds for all eigenvalues λ of $\mathbf{I}-\mathbf{Q}^{-1}\mathbf{A}$, and therefore must also hold for $\rho(\mathbf{I}-\mathbf{Q}^{-1}\mathbf{A})$. $\square$

2.5. Extrapolation.

3. Conjugate gradient and other Krylov subspace methods