# APMA E4300: Introduction to Numerical Methods

Michael Cai

February 12, 2025

**Abstract**

This course introduces and analyzes fundamental numerical methods in scientific computation. The main topics include elementary numerical linear algebra, least-square methods, nonlinear equations and optimization, polynomial interpolation, numerical differentiation and integration, initial-value and boundary problems for ordinary differential equations and systems, eigenvalue problems. This course prepares students for more advanced studies in the field of numerical analysis and scientific computation (for instance, APMA courses E4301, E4302, and E6302).

# Contents

**1 Linear Algebra Overview** **2**
1.1 Linear Iterative Schemes for Linear Systems . . . . . . . . . . . . . . . . . . . . . . . . . . 6

# Chapter 1

# Linear Algebra Overview

## Lecture 1

Jan 23 10:10

**Remark.**

$$Av = \lambda v$$

When $\{v_i\}_{i=1}^{n}$ is linearly independent, then we wil have that

$$Av_i = \lambda_i v_i$$

We can then write $A$ as a product of two matrices such that we have

$$AQ = Q\alpha A = Q\alpha Q^{-1}$$

**Remark.** Suppose that some of the values do not contribute much in

$$A = Q\alpha Q^{-1}$$

Then we can throw away some nonzero elements given that they are small. Thus, we can create a approximation of the matrix which was a (256 x 256) matrix. We have converted it to a $n \times 2, 2 \times 2, 2 \times n$.

**Example.** We can have an image that is 256 x 256, we can compress it without losing too much information. However, there is never any free lunch. Any two dimensional object, 3d, etc. can be done. Thus, we can generalize it further in the course...

With $\lambda$ being very small, it is negligible and can be omitted.

We would like to know how much error can be omitted.

**Definition 1.0.1.** $l^2$ norm is going to be denoted as $||\mathbf{x}||$ which is a negative number given by

$$\sqrt{\sum_{i=1}^{n} x_i^2}$$

**Theorem 1.0.1.** The $l_2$ norm satisfies the following properties:

- $||x||_{l_2} = 0 \Leftrightarrow x = 0$
- $||\alpha x|| = |\alpha| ||x||_{l_2}$

# Lecture 2

Jan 28 10:10

**Remark.** Gaussian Elimination:
Suppose we had an $n \times n$ system with n unknowns and n equations denoted as

$$Ax = y, A \in \mathbb{R}^{\ltimes \times \ltimes}$$

**Example.** Given an example of the equation

$$y = a_0 + a_1 x + \cdots + a_{m-1} x^{m-1}$$

How do we figure out the coefficients. Now suppose that we are given a set of points, it is quite easy to solve a matrix for the given system of equations

$$a_0 + a_1 x + \cdots + a_{n-1} x^{n-1} = y_1$$

$$\ldots$$

$$a_0 + a_1 x_n + \cdots = y_m$$

This results in a matrix where we can solve for the values of the coefficients now.

$$\begin{pmatrix} 1 & x_1 & x_1^{m-1} \\ \ldots & & \\ 1 & x_m & x_m^{m-1} \end{pmatrix} \begin{pmatrix} a_0 \\ \ldots \\ a_m \end{pmatrix} = \begin{pmatrix} y_0 \\ \ldots \\ y_m \end{pmatrix}$$

In practice, we could measure data points which leads to a situation where there are more equations or $x$ values than coefficients which is defined as a "overdetermined matrix". On the other hand, we can also have an undertermined problem.

**Example.** Suppose we have another example with a system of equations with $y = a_0 + a_1 \sin x + a_2 \sin 2x \ldots a_n \sin x$. If we continue this this is also a linear system of equations where the matrix of the equations will be

$$\begin{pmatrix} 1 & \sin x_1 & \sin(m-1)x_1 \\ \ldots & & \\ 1 & \sin x_m & \sin(m-1)x_m \end{pmatrix} \begin{pmatrix} a_0 \\ \ldots \\ a_m \end{pmatrix} = \begin{pmatrix} y_0 \\ \ldots \\ y_m \end{pmatrix}$$

One useful example is a gaussian mixture model where you can suerimpose a number of gaussian models to get the value of y. Thus, solving $Ax = y$ gives many possible applications.

Elementary Row Operations:

**Definition 1.0.2.** Elementary row operations can be done with multiplication. Scaling is done with the following matrix

$$S_i(p) = \begin{pmatrix} 1 & & \\ & p & \\ & & 1 \end{pmatrix}$$

which gives us a scaling factor on a row of the matrix. We simply see that the inverse of the matrix would be simply be $\frac{1}{p}$ for $S_i(p)^{-1}$

**Definition 1.0.3.** Interchange: interchanging row $i$ and row $j$ of $\vec{A}$ gives us

$$E_{ij}A = \begin{pmatrix} 1 & & & & \\ & 0 & 1 & & \\ & 1 & 0 & & \\ & & & 1 & \\ & & & & 1 \end{pmatrix} A$$

This showcases a exchange of rows $i$ and $j$. Note that the inverse of such a matrix is just itself.

**Definition 1.0.4.** Replacement is also a method too. (finish later)

**Remark.** Gaussian Elimination:

We can start with an easy example

$$A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ \dots & \dots & \dots & \dots \end{pmatrix}$$

We can then apply a matrix $L_1$ which is our linear operator we use to transform our matrix using gaussian elimination.

We can apply these operations which results in

$$L_3L_2L_1Ax = L_3L_2L_1y$$

which will give us a upper triangular matrix. We can then solve and use backwards substitution toget our answer from the upper triangular matrix.

Let us denote this as

$$U = L_3L_2L_1A$$

Thus, we can write everything as

$$Ux = L^{-1}y \Rightarrow LUx = y$$

where U is always an upper triangular matrix and the L is always a lower triangular matrix. This only works with elimination methods and not row swaps or replacements. This can be easily solvable by having

$$LUx = y \Rightarrow Lz = y$$

solve for z, do a backward substitution for $U$ and we can easily get the coefficients for $x$

# Lecture 4

Feb 6 10:10

**Remark.** We can start with the form

$$x = Px + y$$

where we wish to solve for $x$. We can construct an iterative way of calculating where $x$ is a vector. THe iteration is based on the equation

$$(I + (A - I))\,x = y$$

$$\to x = (I - A)x + y$$

We can calculate iteratively for the $k+1$ term where

$$x^{k+1} = Px^k + y, \text{given } x^0$$

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{k+1} = \begin{pmatrix} \frac{1}{2} & \\ & \frac{1}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{k} + \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \text{given } \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}^{0} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}$$

We can thus solve to get the next term in the results. If $x^k$ converges to some vector we can write this as

$$\overline{x} = P\overline{x} + y$$

This tells us that if we continue the calculations, we find the solution to the equation. Thus we wish to find the stable solutions. The trivial case is for when $P = 0$. Another case is where

$$Px^k \to 0$$

which gives a stable solution. Thus, we see that all values of $P < 1$ lead to convergence. Thus when the "size" is less than 1, we will have a convergence. We can do this to find a solution to the form

$$x = Px + y$$

Thus we only need that the size of $P$ which is $\rho(P) < 1$ and a guess $x^0$ .

**Remark.** Iterative Schemes: They are used in the solution of linear and nonlinear problems with the form above. For such schemes we need the following

1. Starting point
2. Convergence
3. Rate of Convergence
4. Stopping Criteria: We stop when $\Delta x < \epsilon$

Note that in the limiting case we will have the following equations

$$x^{k+1} = Bx^k + y$$

$$x = Bx + y$$

we will measure how far we are from the solution and we can denote $\epsilon$ as the error.

$$\epsilon^{k+1} = B\epsilon^k$$

If $\|B\| < 1$ then we get a geometric series decreasing where

$$\lim_{k\to\infty} \epsilon^k \to 0, \text{and } x^k \to x$$

**Remark.** Schemes for Linear Systems:

To solve for $Ax = y$ we can rewrite this as $x = (I - A)x + y$ where $B = I - A$. However this is not a good way to construct a $B$ that works. There is a less naive strategy where we can construct a $B$ that works. We can split up $A$ $A = P - N$

$$(P - N)x = y$$

$$x = P^{-1}Nx + P^{-1}y$$

where we must have that

$$\rho(B) = \rho(P^{-1}N) < 1$$

This can be rewritten as

$$x = Bx + \widetilde{y}$$

By choosing $P$ to be invertible, we can rewrite the iteration to be the form we had.

**Definition 1.0.5.** Jacobi Iteration:

1. $D_A$: the diagonal matrix of $A$
2. $L_A$: the negative of $A$ lower triangular
3. $U_A$: the negative of $A$ upper triangular

**Definition 1.0.6.** Richardson Iteration:

$$P = I, \text{therefore } N = I - A$$

# Lecture 4: 10:10

6 Feb 2025

## 1.1 Linear Iterative Schemes for Linear Systems

**Definition 1.1.1.** Define the error at step $k$ to be

$$e^k = x^k - x \Rightarrow e^{k+1} = Be^k$$

$$\Rightarrow e^k = B^k e^0$$

We see if $B$ is greater than 1 , we are amplifying the noise of the answer. Thus, we want to use some method to get the size of the matrix $I - A$ to be less than 1

### 1.1.1 Jacobi Iteration

For the Jacobi Iteration we acn write that

$$A = P - N$$

$$Ax = y \Rightarrow Px = Nx + y$$

**Remark.** We have rewritten the equation in this form but is not in the form $x = Bx + y$. We can write instead

$$X = P^{-1}Nx + P^{-1}y$$

We can now perform the iteration here which makes the calculations much easier.

Recall that the Jacobi is going to consist of

$$P = D_A, N = L_A + U_A$$

**Remark.** Selecting $P = D_A$ is useful because the inverse of $P$ is simply

$$\begin{pmatrix} \frac{1}{p_1} & & \\ & \frac{1}{p_2} & \\ & & \frac{1}{p_n} \end{pmatrix}$$

Thus given $x_0$,

$$x^{k+1} = D_A^{-1}(L_A + U_A)x^k + D_A^{-1}y, k \geq 0$$

Note that if we have the diagonal to be 0 for one part, the method is not possible. However, it is nice to see when the diagonal is very large.

In component form, we can write this as

$$x_1^{k+1} = \frac{y_1 - \sum_{j\neq 1} a_j x_j^K}{a_{11}}$$

$$\vdots$$

$$x_i^{k+1} = \frac{y_i - \sum_{i\neq j} a_{ij} x_j^k}{a_{ii}}$$

$$\vdots$$

$$x_n^{k+1} = \frac{y_n - \sum_{j\neq n} a_{nj} x_j^k}{a_{nn}}$$

**Stopping Criteria** We have to designate a stopping criteria for linear iterations. ONe of the methods is using a maximum amount of iterations and the other method is using the relative size of the residual. This is defined as the size of $y - Ax^k$ over the initial as

$$\frac{\|y - Ax^k\|_{l2}}{\|y - Ax^0\|_{l2}}$$