1 February 16th, 2021

1.1 Basic Iterative Method

In this chapter, we will introduce iterative methods. There will be a lot of overlap with MATH5311. For iterative methods, we make use of the fact that matrix vector products are fast for sparse matrices.

Remark 1.1 — If the matrix is sparse, then the matrix vector product is on the order of non-zero entries.

Example 1.2
For the Discrete Laplacian in 2D, the matrix vector product is $O(N)$.

We will solve $Ax = b$ by stationary iterative methods. Given $x_k \in \mathbb{R}^n$, we want to improve the quality of x_k using:

$$x_{k+1} = Gx_k + f, \quad k \in 0, 1, 2 \ldots$$

where $G \in \mathbb{R}^{n\times n}$ and $f \in \mathbb{R}^n$ are stationary matrices and vectors.

Definition 1.3. G is a **stationary matrix**, as it does not depend on k.

1.2 Jacobi Iteration

- $(y)_i$ denotes the i-th component of a vector y
- $\xi_i^{(k)}$ denotes the i-th component of x_k
- ξ_i denotes the i-th component of x (true solution)
- ξ_i denotes the i-th component of b

The idea of the Jacobi iteration is, given x_k, we obtain x_{k+1} by solving the i-th unknown from the i-th equation. More precisely, we are solving:

$$(Ax - b)_i = 0,$$

with ξ_j, $j \neq i$, fixed to be $\xi_j^{(k)}$, for $i = 1, \ldots, n$. As such, we have:

$$\begin{aligned}
&(Ax - b)_i = 0 \\
\iff\ & a_{ii}\xi_i^{(k+1)} + \sum_{j\neq i} a_{ij}\xi_j^{(k)} = \beta_i \\
\iff\ & \xi_i^{(k+1)} = (\beta_i - \sum_{j\neq i} a_{ij}\xi_j^{(k)})/a_{ii}.
\end{aligned}$$

This can be expressed as Algorithm 1. In order to perform this effeciently, we reformulate this in matrix notation to make use of BLAS. Let $A = D - E - F$, where:

$$A = \begin{bmatrix} d_1 & & * \\ & \ddots & \\ * & & d_n \end{bmatrix} = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_n \end{bmatrix} - \begin{bmatrix} 0 & & 0 \\ & \ddots & \\ -* & & 0 \end{bmatrix} - \begin{bmatrix} 0 & & -* \\ & \ddots & \\ 0 & & 0 \end{bmatrix} = D - E - F$$

Thus, we have Algorithm 2, which is in stationary form.

Algorithm 1 Element Wise Jacobi Iteration

for $k = 0, 1, 2, \dots$ **do**
  for $i = 1, \dots, n$ **do**
    $\xi_i^{(k+1)} = (\beta_i - \sum_{j \neq i} a_{ij} \xi_j^{(k)}) / a_{ii}$
  end for
end for

Algorithm 2 Jacobi Iteration in Matrix Form

for $k = 0, 1, 2, \dots$ **do**
  $x_{k+1} = D^{-1}(b + (E + F)x_k)$
end for

Remark 1.4 — Some other equivalent forms of the Jacobi Iteration are:

$$x_{k+1} = D^{-1}(E + F)x_k + D^{-1}b$$

$$x_{k+1} = D^{-1}(D - A)x_k + D^{-1}b$$
$$x_{k+1} = (I - D^{-1}A)x_k + D^{-1}b$$

1.3 Review on Norms

1.3.1 Vector Norms

Definition 1.5. A (vector) **norm** on R^n is a function $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}$ that satisfies:

1. $\|x\| \geq 0 \quad \forall x \in \mathbb{R}^n$ and $\|x\| = 0 \iff x = 0$.
2. $\|\alpha x\| = |\alpha| \|x\| \quad \forall \alpha \in \mathbb{R}$ and $x \in \mathbb{R}^n$.
3. $\|x + y\| \leq \|x\| + \|y\| \quad \forall x, y \in \mathbb{R}^n$ (**triangle inequality**).

This defines a **metric** on $\mathbb{R}^n$.

Definition 1.6. A p-norm on $\mathbb{R}^n$ is defined as:

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_1|^p \right)^{1/p}$$

Example 1.7 (Spedial p Norm)
Here are a few common norms on $\mathbb{R}^n$.

- **p-norm** $(p \geq 1)$:

- **Euclidean norm** $(p = 2)$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_1|^2 \right)^{1/2}$$

- **1-norm** $(p = 1)$

$$\|x\|_1 = \sum_{i=1}^{n} |x_1|$$

- **∞-norm** $(p = \infty)$

$$\|x\|_\infty = \max_{i=1}^{n} |x_1|$$

Theorem 1.8 (**Holder's Inequality**)

$$|x^T y| \leq \|x\|_p \|y\|_q$$

if $\frac{1}{p} + \frac{1}{q} = 1$, with $p, q \geq 1$.

Theorem 1.9 (**Cauchy-Schwartz Inequality**)

$$|\langle u, v \rangle| \leq \|u\| \|v\|, \quad \forall u, v \in \mathbb{R}^n$$

Example 1.10 (Weighted Norm)
Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite matrix. Then:

$$\|x\|_A = (x^T A x)^{1/2}$$

is a norm, called the **weighted norm**.

From functional analysis, because $\mathbb{R}^n$ is finite dimensional, any two norms are equivalent. More formally.

Theorem 1.11 (Norm equivalence of $\mathbb{R}^n$)
Given $\|\cdot\|_a$ and $\|\cdot\|_b$, $\exists C_1, C_2 > 0$ independent of x, s.t.

$$C_1\|x\|_b \leq \|x\|_a \leq C_2\|b\| \forall x \in \mathbb{R}^n$$

Consequently, from Theorem 1.11, the convergence of vectors in $\mathbb{R}^n$ under any norm is the same. Thus, we can analyze the convergence under any norm.

Remark 1.12 — Theorem 1.11 does not hold for infinte dimensional space. However, for numerical analysis, we always work with finite dimensional space.

Example 1.13 (Equivalence of 1-norm and other p-norms)

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2 \quad \forall x \in \mathbb{R}^n$$
$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty \quad \forall x \in \mathbb{R}^n$$

1.3.2 Matrix Norm

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Let $A \in \mathbb{R}^{n\times n}$ be a matrix.

Definition 1.14. The **norm of A induced by the vector norm $\|\cdot\|$** is:

$$\|A\| = \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|}{\|x\|} = \max_{\|x\|=1} \|Ax\|$$

Remark 1.15 — The second equality in 1.14 is due to the scaling property of A and because the norm is continuous. However, this might not be the case in infinite-dimensional spaces.

We can check that $\|A\|$ is a matrix, i.e.:

- $\|A\| \geq 0 \quad \forall A \in \mathbb{R}^{n\times n}$ and $\|A\| = 0 \iff A = 0$.
- $\|\alpha A\| = |\alpha|\|A\| \quad \forall \alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{n\times n}$.
- $\|A + B\| \leq \|A\| + \|B\| \quad \forall A, B \in \mathbb{R}^{n\times n}$ (**triangle inequality**).

In addition, since it is an **operator norm** that is induced, it has some consistency properties, namely

- $\|AB\| \leq \|A\|\|B\| \quad \forall A, B \in \mathbb{R}^{n\times n}$
- $\|Ax\| \leq \|A\|\|x\| \quad \forall A \in \mathbb{R}^{n\times n}, x \in \mathbb{R}^n$

Example 1.16 (matrix 2-norm)

$$\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2 = \left(\max_{\|x\|_x=1} \|Ax\|_2^2\right)^{\frac{1}{2}} = \left(\max_{x^Tx=1} x^TA^TAx\right)^{\frac{1}{2}}$$

$$= (\text{maximum eivenvale of } A^TA)^{\frac{1}{2}}$$

which is the maximum **singular value** of A.

Remark 1.17 — The last equality in Example 1.16 can be shown by taking the eigenvalue decomposition of A.

Theorem 1.18

$$\|A\|_1 = \max_{1\le j\le n} \|a_j\|_1, \quad \text{where } A = \begin{bmatrix} a_1 & a_2 & \dots & a_n \end{bmatrix}, a_j \in \mathbb{R}^n,$$

i.e. the maximum column 1-norm (column sum).

Proof.

- $\forall x \in \mathbb{R}^n$ with $\|\|_{1=1}$, we have:

$$\|Ax\|_1 = \|\sum_{j=1}^n x_ja_j\|_1 \le \sum_{j=1}^n |x_j|\|a_j\|_1 \le \max_{1\le j\le n} \|a_j\|_1 \sum_{j=1}^n |x_j| = \max_{1\le j\le n} \|a_j\|_1$$

Taking the max over all $x : \|x\|_1 = 1$, we obtain:

$$\|A\|_1 \le \max_{1\le j\le n} \|a_j\|_1$$

- Let $j_0 = \arg\max_{1\le j\le n} \|a_j\|_1$. Consider $x = e_{j_0}$. Then $\|x\|_1 = 1$ and $Ax = Ae_{j_0} = a_{j_0}$. Thus:

$$\|Ax\|_1 = \|a_{j_0}\|_1 = \max_{1\le j\le n} \|a_j\|_1$$

Therefore:

$$\|A\|_1 \ge \|Ax\|_1 = \max_{1\le j\le n} \|a_j\|_1$$

□

Remark 1.19 — This means that for the matrix 1-norm, the maximum is attained at the image of one of the standard unit vector. This is true, since the 1-ball is a convex polytope.

Theorem 1.20

$$\|A\|_\infty = \max_{1\le i\le n} \|a^{(i)}\|_\infty, \quad \text{where } A = \begin{bmatrix} (a^{(1)})^T \\ \vdots (a^{(n)})^T \end{bmatrix}, a^{(i)} \in \mathbb{R}^n,$$

i.e. the maximum row 1-norm (maximum row sum).

Proof. (omitted). □

Definition 1.21. The **spectral radius** of a matrix A is defined as:

$$\rho(A) = \max\{|\lambda_i| : \lambda_i \text{ is an eigenvalue of } A\}$$

Theorem 1.22

Let $A \in \mathbb{C}^{n\times n}$. Then:

1. $\|A\| \geq \rho(A)$ for any matrix norm induced by $\|\cdot\|$.
2. For any $\epsilon > 0$, we can find a vector norm $\|\cdot\|$, s.t. the induced matrix norm satsfies:
$$\|A\| \leq \rho(A) + \epsilon$$
3. From (1) and (2), we have:
$$\rho(A) = \inf \|A\|$$
4. If A is diagonalizable, there exists a matrix operator norm s.t.
$$\rho(A) = \|A\|$$
5. In particular, when A is symmetric, $\rho(A) = \|A\|_2$.

Proof.

1. Let λ_0, x_0 be an eigenpair of A satisfying $|\lambda_0| = \rho(A)$. Assume that $\|x_0\| = 1$. Then, for any vector norm $\|\cdot\|$, its induced operator norm satisfies:
$$\|A\| \geq \|Ax_0\| = \|\lambda_0 x\| = |\lambda_0|\|x_0\| = \rho(A)$$
2. We use a construction proof by finding such vector norm. Let
$$A = X \begin{bmatrix} \lambda_1 & \delta_1 & & & \\ & \lambda_2 & \delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \delta_{n-1} \\ & & & & \lambda_n \end{bmatrix}$$
be the Jordan decomposition, where: $\delta_i \in \{0, 1\}$, and λ_i are eigenvalues of A.
Given $\epsilon > 0$, we define:
$$\|x\|_\epsilon = \|(XD_\epsilon)^{-1}x\|_\infty, \quad \text{with } D_\epsilon = \begin{bmatrix} 1 & & & & \\ & \epsilon & & & \\ & & \epsilon^2 & & \\ & & & \ddots & \\ & & & & \epsilon^{n-1} \end{bmatrix}.$$
We can check that $\|\cdot\|_\epsilon$ is a norm on $\mathbb{C}^n$. So:
$$\|A\|_\epsilon = \max_{\|x\|_\epsilon=1} \|Ax\|_\rho = \max_{\|(XD_\epsilon)^{-1}x\|_\infty=1} \|(XD_\epsilon)^{-1}Ax\|_\infty$$

Let $y = (XD_\epsilon)^{-1}x$, we have:

$$\|A\|_\epsilon = \max_{\|y\|_\infty=1} \|(XD_\epsilon)^{-1}A(XD_\epsilon)y\|_\infty = \|(XD_\epsilon)^{-1}A(XD_\epsilon)\|_\infty$$

Note that we have:

$$(XD_\epsilon)^{-1}A(XD_\epsilon) = D_\epsilon^{-1}X^{-1}AXD_\epsilon = D_\epsilon \begin{bmatrix} \lambda_1 & \delta_1 & & & \\ & \lambda_2 & \delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \delta_{n-1} \\ & & & & \lambda_n \end{bmatrix} D_\epsilon$$

$$= \begin{bmatrix} \lambda_1 & \epsilon\delta_1 & & & \\ & \lambda_2 & \epsilon\delta_2 & & \\ & & \ddots & \ddots & \\ & & & \lambda_{n-1} & \epsilon\delta_{n-1} \\ & & & & \lambda_n \end{bmatrix}$$

Thus, since the infinity norm is the maximum row sum, we have:

$$\|A\|_\epsilon \leq \max_{1\leq i\leq n}(|\lambda_i| + \epsilon) \leq \rho(A) + \epsilon$$

3. By part 2, if A is diagonalizable, $\delta_i = 0$ for all i. Then $\|A\|_\epsilon = \max_i |\lambda_i| = \rho(A)$. If $A = A^T$, then $\delta_i = 0$ for all i, λ_i are real, and X is unitary. Thus:

$$A^TA = X^{-1} \begin{bmatrix} \lambda_1^2 & & & \\ & \lambda_2^2 & & \\ & & \ddots & \\ & & & a\lambda_n^2 \end{bmatrix}$$

$$\rho(A) = (\rho(A^TA))^{\frac{1}{2}} = \|A\|_2$$

□

Remark 1.23 — 1. and 2. imply

$$\rho(A) = \inf\{\|A\| : \|\cdot\| \text{ is an operator norm }\}$$

In particular

- If A diagonalizable, then minimum is attainable, meaning:

$$\rho(A) = \min\{\|A\| : \|\cdot\| \text{ is an operator norm }\}$$

- If A is symmetric:

$$\rho(A) = \|A\|_2 = \min\{\|A\| : \|\cdot\| \text{ is an operator norm }\}$$