Numerical Analysis Lecture Notes

Rebekah Dix

October 2, 2018

Contents

1 Solution of equations by iteration

Theorem 1. *(Existence of Root) Let f be a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line. Assume further, that $f(a)f(b) \leq 0$; then, there exists ξ in $[a, b]$ such that $f(\xi) = 0$.*

Proof. The condition $f(a)f(b) \leq 0$ implies that $f(a)$ and $f(b)$ have opposite signs, or one of them is 0. If either $f(a)$ or $f(b)$ is 0, then we've found a root. Suppose that both endpoints are non-zero (in which case they have opposite signs). In this case, 0 must belong to the open interval whose endpoints are $f(a)$ and $f(b)$. The intermediate value theorem gives the existence of a root in the open interval (a, b). Thus, in both cases, a zero is guaranteed. □

- The converse of Theorem 1 is clearly false.

Theorem 2. *(Brouwer's Fixed Point Theorem) Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Then, there exists $\xi \in [a, b]$ such that $\xi = g(\xi)$. ξ is called a fixed point of the function g.*

Proof. Define a function $f(x) = x - g(x)$. If we find a root ξ of f, then ξ is a fixed point of g. Then,

$$f(a)f(b) = (a - g(a))(b - g(b)) \leq 0 \tag{1}$$

By assumption, $a \leq g(a), g(b) \leq b$. Therefore, the first term is negative and the second term is positive. Therefore, $f(a)f(b) \leq 0$. By Theorem 1, there exists a $\xi \in [a, b]$ such that $f(\xi) = 0$. Then, for this ξ, $g(\xi) = \xi$. □

Definition 1. *(Simple Iteration) Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Given that $x_0 \in [a, b]$, the recursion defined by*

$$x_{k+1} = g(x_k) \tag{2}$$

is called simple iteration; the numbers x_k, $k \geq 0$, are referred to as iterates.

- If this sequence converges, the limit must be a fixed of g, since g is continuous on a closed interval. Note that

$$\xi = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} g(x_k) = g\left(\lim_{k\to\infty} x_k\right) = g(\xi) \tag{3}$$

Definition 2. *(Contraction) Let g be a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line. Then, g is said to be a contraction on $[a, b]$ if there exists a constant L such that $0 < L < 1$ and*

$$|g(x) - g(y)| \leq L|x - y| \quad \forall x, y \in [a, b] \tag{4}$$

Theorem 3. *(Contraction Mapping Theorem) Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$. Suppose g is a contraction on $[a, b]$. Then, g has a unique fixed point ξ in the interval $[a, b]$. Moreover, the sequence (x_k) defined by simple iteration converges to ξ as $k \to \infty$ for any starting value x_0 in $[a, b]$.*

Let $\epsilon > 0$ be a certain tolerance, and let $k_0(\epsilon)$ denote the smallest positive integer such that x_k is no more than ϵ away from the fixed point ξ (i.e. $|x_k - \xi| \leq \epsilon$) for all $k \geq k_0(\epsilon)$. Then,

$$k_0(\epsilon) \leq \left\lfloor \frac{\ln|x_1 - x_0| - \ln(\epsilon(1 - L))}{\ln(1/L)} \right\rfloor + 1 \tag{5}$$

Proof. Let $E_k = |x_k - \xi|$ be the error at k. Then

$$\begin{aligned} |x_{k+1} - \xi| &= |g(x_k) - g(\xi)| \\ &< L|x_k - \xi| \end{aligned}$$

Therefore

$$E_k \leq L^k E_0 \tag{6}$$

Since $L < 1$, $L^k \to 0$ as $k \to \infty$. □

Definition 3. *(Stable, Unstable Fixed Point) Suppose that g is a real-valued function, defined and continuous on a bounded closed interval $[a, b]$ of the real line, and let $g(x) \in [a, b]$ for all $x \in [a, b]$, and let ξ denote a fixed point of g. ξ is a stable fixed point of g if the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k \geq 0$, converges to ξ whenever the starting value x_0 is sufficiently close to ξ. Conversely, if no sequence (x_k) defined by this iteration converges to ξ for any starting value x_0 close to ξ, except for $x_0 = \xi$, then we say that ξ is an unstable fixed point of g.*

- With this definition, a fixed point may be neither stable nor unstable.

Definition 4. *(Rate of Convergence) Suppose $\xi = \lim_{k\to\infty} x_k$. Define $E_k = |x_k - \xi|$.*

- The sequence (x_k) converges to ξ linearly if there exists a number $\mu \in (0, 1)$ such that
$$\lim_{k\to\infty} \frac{E_{k+1}}{E_k} = \mu \tag{7}$$

- The sequence (x_k) converges to ξ superlineraly if $\mu = 0$. That is, the sequence of μ_k generated at each step $\to 0$ as $k \to \infty$.

- The sequence (x_k) converges to ξ with order q if there exists a $\mu > 0$ such that
$$\lim_{k\to\infty} \frac{E_{k+1}}{E_k^q} = \mu \tag{8}$$

  In particular, if $q = 2$, then the sequence converges quadratically.

2 Solution of systems of linear equations

2.1 Least Squares

Given a system of equations $Ax = b$, the least squares problem is

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \tag{9}$$

We can expand the objective function out as

$$\begin{aligned} \|Ax - b\|_2^2 &= (Ax - b)^T (Ax - b) \\ &= x^T A^T A x - 2b^T A x + b^T b \end{aligned}$$

To find the x that minimizes this expression we find the x that satisfies $\nabla_x F = 0$. That is

$$\nabla_x F = 0 = 2A^T A x - 2A^T b \tag{10}$$

Therefore the minimizer is $x = (A^T A)^{-1} A^T b$. $(A^T A)^{-1} A^T$ is called the pseudo-inverse of A. If A is square and invertible, then the pseudo-inverse equals A^{-1}.

2.2 Gram-Schimdt Orthogonalization

Algorithm: Denote the columns of A by a_i.

1. $q_1 = a_1$. Then normalized by $q_1 = \frac{q_1}{\|q_1\|}$.
2. $q_2 = a_2 - \langle q_1, a_2 \rangle q_1$. Then normalize by $q_2 = \frac{q_2}{\|q_2\|}$. It's simple to verify that $q_2 \perp q_1$.
3. For an arbitrary k, $q_k = a_k - \langle a_k, q_1 \rangle q_1 - \langle a_k, q_2 \rangle q_2 - \ldots - \langle a_k, q_{k-1} \rangle q_{k-1}$. Then normalize by $q_k = \frac{q_k}{\|q_k\|}$.

We can observe the following properties:

1. $\|q_i\| = 1$ (this follows directly)
2. $q_i \perp q_j$ for all $i \neq j$
3. $q_k \in span(a_1, \ldots, a_k)$ and $a_k \in span(q_1, \ldots, q_k)$ so that $span(a_1, \ldots, a_k) = span(q_1, \ldots, q_k)$.

[[Write proof for 2]].

2.3 QR Factorization

Definition 5. *(Unitary Matrix) A matrix $Q = [q_1 \ldots q_n] \in \mathbb{R}^{m \times n}$ is unitary if and only if $\langle q_i, q_j \rangle = \delta_{ij}$.*

Observations about this definition:

1. $Q^T Q = I$

2. If Q is square, then $Q^T = Q^{-1}$.

2.3.1 Application to Least Squares

Suppose that we can write $A = QR$, where $A \in \mathbb{R}^{m\times n}$, $Q \in \mathbb{R}^{m\times n}$ and unitary, and $R \in \mathbb{R}^{n\times n}$ and upper triangular. Then the least squares solution to $Ax = b$ is given by

$$\begin{aligned} x &= (A^T A)^{-1} A^T b \\ &= (R^T Q^T Q R)^{-1} R^T Q^T b \\ &= (R^T R)^{-1} R^T Q^T b \\ \implies (R^T R)x &= R^T Q^T b \\ Rx &= Q^T b \qquad \text{(assume } R \text{ is invertible (i.e. no zeros on the diagonal))} \end{aligned}$$

We can then solve for x using back substitution, which is $\mathcal{O}(n^2)$.

2.4 Norms and Condition Numbers

Definition 6. *(Norm) Suppose that $\mathcal{V}$ is a linear space over the field $\mathbb{R}$. The nonnegative real-valued function $\|\cdot\|$ is a norm on $\mathcal{V}$ if the following axioms are satisfied: Fix $v \in \mathcal{V}$*

1. *Positivity: $\|v\| = 0$ if and only if $v = 0$*

2. *Scale Preservation: $\|\alpha v\| = |\alpha|\,\|v\|$ for all $\alpha \in \mathbb{R}$*

3. *Triangle Inequality: $\|v + w\| \leq \|v\| + \|w\|$.*

Example 1. (Examples of Norms)

1. 1-norm:
$$\|v\|_1 = \sum_{i=1}^{n} |v_i| = |v_1| + \cdots + |v_n| \tag{11}$$

2. 2-norm:
$$\|v\|_2 = \left(\sum_{i=1}^{n} v_i^2\right)^{\frac{1}{2}} = \sqrt{v_1^2 + \cdots + v_n^2} = \sqrt{v^T v} \tag{12}$$

3. ∞-norm
$$\|x\|_\infty = \max_{i=1,\dots,n} |v_i| \tag{13}$$

4. p-norm

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}} \tag{14}$$

For the p-norm, proving the triangle inequality follows from the Minkowski's inequality.

Definition 7. *(Operator Norm) Let A be an $m \times n$ matrix. That is, A is a linear transformation form $\mathbb{R}^n$ to $\mathbb{R}^m$. Then the operator norm (or subordinate matrix norm) of A is*

$$\|A\|_{p,q} = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_q}{\|x\|_p}. \tag{15}$$

Observations about this definition:

1. It's easy to check that this definition of the operator norm satisfies the properties of a norm given in Definition 6. For the triangle inequality, observe that

$$\begin{aligned} \|(A+B)x\|_p &\leq \|Ax\|_p + \|Bx\|_p && \text{(from Minkowski's inequality)} \\ \implies \frac{\|(A+B)x\|_p}{\|x\|_p} &\leq \frac{\|Ax\|_p}{\|x\|_p} + \frac{\|Bx\|_p}{\|x\|_p} \end{aligned}$$

Taking the supremum of both sides over x shows that $\|A+B\|_p \leq \|A\|_p + \|B\|_p$.

2. The definition immediately implies that for an arbitrary $x \in \mathbb{R}^n$, $x \neq 0$,

$$\|Ax\|_q \leq \|A\|_{p,q} \|x\|_p \tag{16}$$

We can generalize this inequality to claim that

$$\|AB\| \leq \|A\| \|B\| \tag{17}$$

for conformable matrices A, B. Indeed, fix $0 \neq x \in R^n$. Then

$$\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\| \tag{18}$$

We can divide all inequalities by $\|x\|$ to see that for all $x \neq 0$,

$$\frac{\|ABx\|}{\|x\|} \leq \|A\| \|B\| \tag{19}$$

Taking the supremum over x on the left hand side shows that $\|AB\| \leq \|A\| \|B\|$.

Theorem 4. *(The 1-norm of a matrix is the largest absolute-value column sum) Let $A \in \mathbb{R}^{m \times n}$ and denote the columns of A by a_j, $j = 1, \ldots, n$. Then $\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}| = \max_{j=1,\ldots,n} \|a_j\|$.*

Proof. Fix $x \in \mathbb{R}^n$. Let $C = \max_{j=1,\dots,n} \sum_{i=1}^m |a_{ij}|$. First consider the product $A \cdot x$. The ith element is $\sum_{j=1}^n a_{ij} x_j$. Then

$$\begin{aligned}
\|Ax\|_1 &= \sum_{i=1}^m |(Ax)_i| = \sum_{i=1}^m |\sum_{j=1}^n a_{ij} x_j| \\
&\le \sum_{i=1}^m \sum_{j=1}^n |a_{ij}||x_j| && \text{(triangle inequality)} \\
&= \sum_{j=1}^n |x_j| \left(\sum_{i=1}^m |a_{ij}| \right) && \text{(interchange order of summation, assumed finite)} \\
&\le C \|x\|_1
\end{aligned}$$

Therefore $\frac{\|Ax\|_1}{\|x\|_1} \le C$ for all x. Next, we find an x such we achieve equality with C. Call index J the index such that $\|a_J\|_1 = C = \max_{j=1,\dots,n} \sum_{i=1}^m |a_{ij}|$. Then let e_J be the n-vector of zeros with a 1 in the Jth entry. Clearly $\|e_J\|_1 = 1$. But then

$$\|Ae_J\|_1 = \|a_J\|_1 = C \tag{20}$$

In sum, we first showed that for all $x \in \mathbb{R}^n$

$$\frac{\|Ax\|_1}{\|x\|_1} \le C \tag{21}$$

We then found an $x \in \mathbb{R}^n$ such that $\frac{\|Ax\|_1}{\|x\|_1} = C$. Therefore

$$\|A\|_1 = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} = C = \max_{j=1,\dots,n} \sum_{i=1}^m |a_{ij}| = \max_{j=1,\dots,n} \|a_j\| \tag{22}$$

□

Theorem 5. *(The ∞-norm of a matrix is the largest absolute-value row sum) Let $A \in \mathbb{R}^{m \times n}$ and denote the rows of A by b_i, $i = 1, \dots, m$. Then* $\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}| = \max_{i=1,\dots,m} \|b_i\|$.

Proof. Fix $x \in \mathbb{R}^n$. Let $C = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|$.

$$\begin{aligned}
\|Ax\|_\infty &= \max_{i=1,\dots,m} |\sum_{j=1}^{n} a_{ij}x_j| \\
&\le \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}||x_j| && \text{(by the triangle inequality)} \\
&\le \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}|\, \|x\|_\infty && \text{(since } |x_j| \le \|x\|_\infty \text{ for all } j\text{)} \\
&= C\, \|x\|_\infty
\end{aligned}$$

Next, we find an x such we achieve equality with C. Call I the index for which $\|b_I\|_\infty = C$. Define

$$x_j = \begin{cases} 1 & a_{Ij} > 0 \\ -1 & a_{Ij} < 0 \end{cases} \tag{23}$$

Observe that $\|x\|_\infty = 1$. Then

$$\begin{aligned}
|A \cdot x|_I &= |b_I^T \cdot x| \\
&= |\sum_{j=1}^{m} a_{Ij}x_j| \\
&= |\sum_{j=1}^{m} a_{Ij}| \\
&= C
\end{aligned}$$

We then found an $x \in \mathbb{R}^n$ such that $\frac{\|Ax\|_\infty}{\|x\|_\infty} = C$. Therefore

$$\|A\|_\infty = \sup_{x \in \mathbb{R}^n, x \ne 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = C = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}| = \max_{i=1,\dots,m} \|b_i\| \tag{24}$$

□

Theorem 6. *(The 2-norm of a symmetric positive definite matrix is the maximum absolute value of its eigenvalues) Let A be a positive definite $n \times n$ matrix. Then*

$$\|A\|_2 = \max_{i=1,\dots,n} |\lambda_i| \tag{25}$$

Proof. Since A is positive definite, A has n distinct eigenvalues, which implies that it has n linearly independent eigenvectors. Therefore, for an arbitrary $x \in \mathbb{R}^n$, we can write x

as a linearly combination of the eigenvectors $x_1, \ldots, x_n$. Then

$$\begin{aligned} x &= c_1 x_1 + \cdots + c_n x_n \\ Ax &= c_1 A x_1 + \cdots + c_n A x_n \\ &= c_1 \lambda_1 x_2 + \cdots + c_n \lambda_n x_n \end{aligned}$$

We can normalize the eigenvectors of A so that $x_i^T x_i = 1$. Then $\|Ax\|_2 = \sqrt{\sum_{i=1}^n c_i^2 \lambda_i^2}$ and $\|x\|_2 = \sqrt{\sum_{i=1}^n c_i^2}$. Therefore

$$\frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\frac{\sum_{i=1}^n c_i^2 \lambda_i^2}{\sum_{i=1}^n c_i^2}} \leq \max_i |\lambda_i| = |\lambda_I| \tag{26}$$

Now we'll find an x such that we actually achieve equality. Call I the index of the maximum absolute value of an eigenvalue. Then, consider the eigenvector associated with this eigenvalue, called x_i. Then

$$\frac{\|Ax_I\|_2}{\|x_I\|_2} = \frac{|\lambda_I| \, \|x_I\|}{\|x_I\|} = |\lambda_I| \tag{27}$$

This shows that $\|A\|_2 = \max_i |\lambda_i|$. □

Theorem 7. *(The 2-norm of a matrix $A_{m\times n}$ equals its largest singular value) Let A be an $m \times n$ matrix and denote the eigenvalues of the matrix $B = A^T A$ by λ_i, $i = 1, \ldots, n$. Then*

$$\|A\|_2 = \max_i \sqrt{\lambda_i} \tag{28}$$

The square roots of the (nonnegative) eigenvalues of $A^T A$ are referred to as the singular values of A.

2.4.1 Conditioning

Conditioning helps us quantify the sensitivity of the output to perturbations of the input. In what follows, let f be a mapping from a subset D of a normed linear space $\mathcal{V}$ to another normed linear space $\mathcal{W}$.

Definition 8. *(Absolute Condition Number)*

$$Cond(f) = \sup_{x,y \in D, x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|} \tag{29}$$

Definition 9. *(Absolute Local Condition Number)*

$$Cond_x(f) = \sup_{x+\delta x \in D, \delta x \neq 0} \frac{\|f(x+\delta x) - f(x)\|}{\|\delta x\|} \tag{30}$$

The previous two definitions depend on the magnitudes of $f(x)$ and x. In applications, it's often better to rescale as follows

Definition 10. *(Relative Local Condition Number)*

$$cond_x(f) = \sup_{x+\delta x \in D, \delta x \neq 0} \frac{\|f(x+\delta x) - f(x)\| \;/\; \|f(x)\|}{\|\delta x\| \;/\; \|x\|} \tag{31}$$

In these definitions, if f is differentiable then we can replace the differences with the appropriate derivatives.

Example 2. (Example of conditions numbers) Let D be a subinterval of $[0, \infty)$ and $f(x) = \sqrt{x}$. Then $f'(x) = \frac{1}{2\sqrt{x}}$.

1. If $D = [1, 2]$, then $Cond(f) = \frac{1}{2}$.

2. If $D = [0, 1]$, then $Cond(f) = \infty$.

3. If $D = (0, \infty)$, then the the absolute local condition number of f at $x \in D$ is

$$Cond_x(f) = \frac{1}{2\sqrt{x}} \tag{32}$$

   Thus as $x \rightarrow$, $Cond_x(f) \rightarrow \infty$, and as $x \rightarrow \infty$, $Cond_x(f) \rightarrow 0$.

4. If $D = (0, \infty)$, then the relative local condition number of f is $cond_x(f) = 1/2$ for all $x \in D$.

Definition 11. *(Condition Number of a Nonsingular Matrix) The condition number of a nonsingular matrix A is defined by*

$$\kappa(A) = \|A\| \, \|A^{-1}\| \tag{33}$$

If $\kappa(A) \gg 1$, the matrix is said to be ill-conditioned.

Observations about this definition:

1. $\kappa(A) = \kappa(A^{-1})$

2. For all A, $\kappa(A) \geq 1$. This follows because

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \, \|A^{-1}\| \tag{34}$$

3. The condition number of a matrix is unaffected by scaling all its elements by multiplying by a nonzero constant.

4. There is a condition number for each norm, and the size of the condition number is strongly dependent on the choice of norm.

3 Special Matrices

3.1 Symmetric Positive Definite Matrices

Definition 12. *The matrix A is said to be symmetric if $A = A^T$. A square $n \times n$ matrix is called positive definite if*

$$x^T A x > 0 \tag{35}$$

for all $x \in \mathbb{R}^n$.