

# Numerical Analysis Lecture Notes

Rebekah Dix

October 13, 2018

## Contents

<b>1</b>	<b>Results from Real Analysis</b>	<b>3</b>
<b>2</b>	<b>Solution of equations by iteration</b>	<b>3</b>
2.1	Simple Iteration . . . . .	3
2.2	Newton's Method . . . . .	5
2.3	Secant Method . . . . .	5
<b>3</b>	<b>Solution of systems of linear equations</b>	<b>6</b>
3.1	LU Decomposition . . . . .	6
3.2	Least Squares . . . . .	6
3.3	Gram-Schmidt Orthogonalization . . . . .	7
3.4	QR Factorization . . . . .	7
3.4.1	Application to Least Squares . . . . .	8
3.5	Norms and Condition Numbers . . . . .	8
3.5.1	Conditioning . . . . .	12
<b>4</b>	<b>Special Matrices</b>	<b>13</b>
4.1	Symmetric Positive Definite Matrices . . . . .	13
4.2	Banded Matrices and Differential Equations . . . . .	15
4.3	Cholesky Factorization . . . . .	15
<b>5</b>	<b>Simultaneous nonlinear equations</b>	<b>16</b>
5.1	Analysis Preliminaries . . . . .	16
5.2	Simultaneous iteration . . . . .	16
<b>6</b>	<b>Eigenvalues of Eigenvectors of a symmetric matrix</b>	<b>19</b>
6.1	Why we use iteration to calculate eigenvalues/eigenvectors . . . . .	19
6.2	Power Iteration . . . . .	19
6.3	Inverse Iteration . . . . .	20
6.4	Simultaneous Iteration . . . . .	20

6.5	Shifted Power Iteration . . . . .	21
6.6	QR Algorithm . . . . .	21
6.7	Simultaneous Iteration equivalent to QR Algorithm . . . . .	21

# 1 Results from Real Analysis

**Theorem 1.** (*The Mean Value Theorem*) Suppose  $f$  is a real-valued function, defined and continuous on the closed interval  $[a, b] \in \mathbb{R}$  and  $f$  differentiable on the open interval  $(a, b)$ . Then there exists a number  $\xi \in (a, b)$  such that

$$f(b) - f(a) = f'(\xi)(b - a) \quad (1)$$

## 2 Solution of equations by iteration

### 2.1 Simple Iteration

**Theorem 2.** (*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  $\xi$  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.  $\square$

- The converse of Theorem 2 is clearly false.

**Theorem 3.** (*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  $\xi = g(\xi)$ .  $\xi$  is called a fixed point of the function  $g$ .

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

$$f(a)f(b) = (a - g(a))(b - g(b)) \leq 0 \quad (2)$$

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 2, there exists a  $\xi \in [a, b]$  such that  $f(\xi) = 0$ . Then, for this  $\xi$ ,  $g(\xi) = \xi$ .  $\square$

**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

$$x_{k+1} = g(x_k) \quad (3)$$

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

$$\xi = \lim_{k \rightarrow \infty} x_{k+1} = \lim_{k \rightarrow \infty} g(x_k) = g\left(\lim_{k \rightarrow \infty} x_k\right) = g(\xi) \quad (4)$$

**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

$$|g(x) - g(y)| \leq L|x - y| \quad \forall x, y \in [a, b] \quad (5)$$

**Theorem 4.** (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  $\xi$  in the interval  $[a, b]$ . Moreover, the sequence  $(x_k)$  defined by simple iteration converges to  $\xi$  as  $k \rightarrow \infty$  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  $\epsilon$  away from the fixed point  $\xi$  (i.e.  $|x_k - \xi| \leq \epsilon$ ) for all  $k \geq k_0(\epsilon)$ . Then,

$$k_0(\epsilon) \leq \left\lfloor \frac{\ln|x_1 - x_0| - \ln(\epsilon(1 - L))}{\ln(1/L)} \right\rfloor + 1 \quad (6)$$

*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

$$E_k \leq L^k E_0 \quad (7)$$

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

**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  $\xi$  denote a fixed point of  $g$ .  $\xi$  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  $\xi$  whenever the starting value  $x_0$  is sufficiently close to  $\xi$ . Conversely, if no sequence  $(x_k)$  defined by this iteration converges to  $\xi$  for any starting value  $x_0$  close to  $\xi$ , except for  $x_0 = \xi$ , then we say that  $\xi$  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  $\xi = \lim_{k \rightarrow \infty} x_k$ . Define  $E_k = |x_k - \xi|$ .

- The sequence  $(x_k)$  converges to  $\xi$  linearly if there exists a number  $\mu \in (0, 1)$  such that

$$\lim_{k \rightarrow \infty} \frac{E_{k+1}}{E_k} = \mu \quad (8)$$

- The sequence  $(x_k)$  converges to  $\xi$  superlinearly if  $\mu = 0$ . That is, the sequence of  $\mu_k$  generated at each step  $\rightarrow 0$  as  $k \rightarrow \infty$ .
- The sequence  $(x_k)$  converges to  $\xi$  with order  $q$  if there exists a  $\mu > 0$  such that

$$\lim_{k \rightarrow \infty} \frac{E_{k+1}}{E_k^q} = \mu \quad (9)$$

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

## 2.2 Newton's Method

**Definition 5.** (Newton's Method) Newton's method for the solution of  $f(x) = 0$  is defined by

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \quad (10)$$

Geometrically,  $(x_{n+1}, 0)$  is the intersection of the  $x$ -axis and the tangent of the graph of  $f$  at  $(x_n, f(x_n))$ .

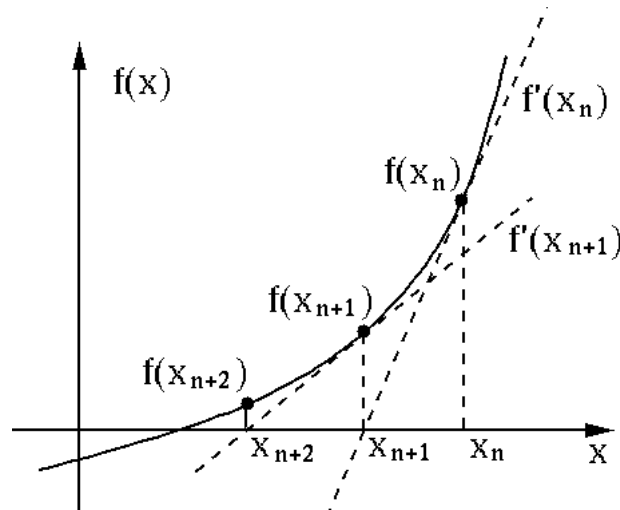


Figure 1: Geometric Interpretation of Newton's Method in  $\mathbb{R}$

## 2.3 Secant Method

Observe that Newton's method requires us to know the first derivative  $f'$  of  $f$ . In applications, we might not know  $f'$  or it could be expensive to calculate. This motivates

approximating the  $f'(x_k)$  in Newton's method with

$$f'(x_k) \approx \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \quad (11)$$

**Definition 6.** (*Secant Method*) The secant method is defined by

$$x_{k+1} = x_k - f(x_k) \frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})} \quad (12)$$

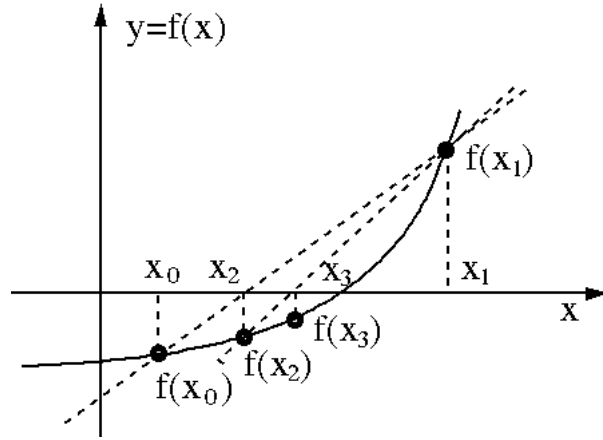


Figure 2: Geometric Interpretation of Secant Method in  $\mathbb{R}$

**Theorem 5.** (*Convergence of Secant Method*) Suppose that  $f$  is a real-valued function, defined and continuously differentiable on an interval  $I = [\xi - h, \xi + h]$ ,  $h > 0$ , with center point  $\xi$ . Suppose further that  $f(\xi) = 0$ ,  $f'(\xi) \neq 0$ . Then, the sequence  $(x_k)$  defined by the secant method converges at least linearly to  $\xi$  provided that  $x_0$  and  $x_1$  are sufficiently close to  $\xi$ .

*Proof.*

□

### 3 Solution of systems of linear equations

#### 3.1 LU Decomposition

#### 3.2 Least Squares

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

$$\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 \quad (13)$$

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

$$\nabla_x F = 0 = 2A^T A x - 2A^T b \quad (14)$$

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  $A^{-1}$ .

### 3.3 Gram-Schmidt 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  $k$ ,  $q_k = a_k - \langle a_k, q_1 \rangle q_1 - \langle a_k, q_2 \rangle q_2 - \dots - \langle a_k, q_{k-1} \rangle q_{k-1}$ . 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.  $q_k \in \text{span}(a_1, \dots, a_k)$  and  $a_k \in \text{span}(q_1, \dots, q_k)$  so that  $\text{span}(a_1, \dots, a_k) = \text{span}(q_1, \dots, q_k)$ .

[[Write proof for 2]].

### 3.4 QR Factorization

**Definition 7.** (Unitary Matrix) A matrix  $Q = [q_1 \dots q_n] \in \mathbb{R}^{m \times n}$  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  $Q^T = Q^{-1}$ .

### 3.4.1 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

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

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

## 3.5 Norms and Condition Numbers

**Definition 8.** (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:

$$\|v\|_1 = \sum_{i=1}^n |v_i| = |v_1| + \cdots + |v_n| \quad (15)$$

2. 2-norm:

$$\|v\|_2 = \left( \sum_{i=1}^n v_i^2 \right)^{\frac{1}{2}} = \sqrt{v_1^2 + \cdots + v_n^2} = \sqrt{v^T v} \quad (16)$$

3.  $\infty$ -norm

$$\|x\|_\infty = \max_{i=1, \dots, n} |v_i| \quad (17)$$

4.  $p$ -norm

$$\|v\|_p = \left( \sum_{i=1}^n |v_i|^p \right)^{\frac{1}{p}} \quad (18)$$

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



**Definition 9.** (Operator Norm) Let  $A$  be an  $m \times n$  matrix. That is,  $A$  is a linear transformation from  $\mathbb{R}^n$  to  $\mathbb{R}^m$ . Then the operator norm (or subordinate matrix norm) of  $A$  is

$$\|A\|_{p,q} = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_q}{\|x\|_p}. \quad (19)$$

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 8. 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  $x \in \mathbb{R}^n, x \neq 0$ ,

$$\|Ax\|_q \leq \|A\|_{p,q} \|x\|_p \quad (20)$$

We can generalize this inequality to claim that

$$\|AB\| \leq \|A\| \|B\| \quad (21)$$

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

$$\|ABx\| \leq \|A\| \|Bx\| \leq \|A\| \|B\| \|x\| \quad (22)$$

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

$$\frac{\|ABx\|}{\|x\|} \leq \|A\| \|B\| \quad (23)$$

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

**Theorem 6.** (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  $a_j, j = 1, \dots, n$ . Then  $\|A\|_1 = \max_{j=1, \dots, n} \sum_{i=1}^m |a_{ij}| = \max_{j=1, \dots, n} \|a_j\|$ .

*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  $i$ th

element is  $\sum_{j=1}^n a_{ij}x_j$ . Then

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

Therefore  $\frac{\|Ax\|_1}{\|x\|_1} \leq C$  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  $J$ th entry. Clearly  $\|e_J\|_1 = 1$ . But then

$$\|Ae_J\|_1 = \|a_J\|_1 = C \quad (24)$$

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

$$\frac{\|Ax\|_1}{\|x\|_1} \leq C \quad (25)$$

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

$$\|A\|_1 = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_1}{\|x\|_1} = C = \max_{j=1,\dots,n} \sum_{i=1}^m |a_{ij}| = \max_{j=1,\dots,n} \|a_j\|_1 \quad (26)$$

□

**Theorem 7.** (The  $\infty$ -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  $\|A\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^n |a_{ij}| = \max_{i=1,\dots,m} \|b_i\|_1$ .

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

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

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

$$x_j = \begin{cases} 1 & a_{Ij} > 0 \\ -1 & a_{Ij} < 0 \end{cases} \quad (27)$$

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

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

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

$$\|A\|_\infty = \sup_{x \in \mathbb{R}^n, x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = C = \max_{i=1, \dots, m} \sum_{j=1}^n |a_{ij}| = \max_{i=1, \dots, m} \|b_i\| \quad (28)$$

□

**Theorem 8.** (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

$$\|A\|_2 = \max_{i=1, \dots, n} |\lambda_i| \quad (29)$$

*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  $x_1, \dots, x_n$ . Then

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

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

$$\frac{\|Ax\|_2}{\|x\|_2} = \sqrt{\frac{\sum_{i=1}^n c_i^2 \lambda_i^2}{\sum_{i=1}^n c_i^2}} \leq \max_i |\lambda_i| = |\lambda_I| \quad (30)$$

Now we'll find an  $x$  such that we actually achieve equality. Call  $I$  the index of the max-

imum absolute value of an eigenvalue. Then, consider the eigenvector associated with this eigenvalue, called  $x_i$ . Then

$$\frac{\|Ax_I\|_2}{\|x_I\|_2} = \frac{|\lambda_I| \|x_I\|}{\|x_I\|} = |\lambda_I| \quad (31)$$

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

**Theorem 9.** (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  $\lambda_i$ ,  $i = 1, \dots, n$ . Then

$$\|A\|_2 = \max_i \sqrt{\lambda_i} \quad (32)$$

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

### 3.5.1 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 10.** (Absolute Condition Number)

$$\text{Cond}(f) = \sup_{x, y \in D, x \neq y} \frac{\|f(x) - f(y)\|}{\|x - y\|} \quad (33)$$

**Definition 11.** (Absolute Local Condition Number)

$$\text{Cond}_x(f) = \sup_{x + \delta x \in D, \delta x \neq 0} \frac{\|f(x + \delta x) - f(x)\|}{\|\delta x\|} \quad (34)$$

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

**Definition 12.** (Relative Local Condition Number)

$$\text{cond}_x(f) = \sup_{x + \delta x \in D, \delta x \neq 0} \frac{\|f(x + \delta x) - f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|} \quad (35)$$

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  $\text{Cond}(f) = \frac{1}{2}$ .

2. If  $D = [0, 1]$ , then  $\text{Cond}(f) = \infty$ .
3. If  $D = (0, \infty)$ , then the absolute local condition number of  $f$  at  $x \in D$  is

$$\text{Cond}_x(f) = \frac{1}{2\sqrt{x}} \quad (36)$$

Thus as  $x \rightarrow 0$ ,  $\text{Cond}_x(f) \rightarrow \infty$ , and as  $x \rightarrow \infty$ ,  $\text{Cond}_x(f) \rightarrow 0$ .

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

**Definition 13.** (*Condition Number of a Nonsingular Matrix*) The condition number of a nonsingular matrix  $A$  is defined by

$$\kappa(A) = \|A\| \|A^{-1}\| \quad (37)$$

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

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \|A^{-1}\| \quad (38)$$

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.

## 4 Special Matrices

### 4.1 Symmetric Positive Definite Matrices

**Definition 14.** (*Symmetric, Positive Definite, spd*) The real matrix  $A$  is said to be symmetric if  $A = A^T$ . A square  $n \times n$  matrix is called positive definite if

$$\mathbf{x}^T A \mathbf{x} > 0 \quad (39)$$

for all  $\mathbf{x} \in \mathbb{R}^n$ ,  $\mathbf{x} \neq 0$ .

**Theorem 10.** (*Properties of spd matrices*) Let  $A$  be an  $n \times n$  real, spd matrix. Then

1.  $a_{ii} > 0$  for all  $i = 1, \dots, n$  (the diagonal elements of  $A$  are positive).

2.  $Ax_i = \lambda_i x_i \implies \lambda_i \in \mathbb{R}_{>0}, x_i \in \mathbb{R}^n \setminus \{0\}$  (the eigenvalues of  $A$  are real and positive, and the eigenvectors of  $A$  belong to  $\mathbb{R}^n \setminus \{0\}$ ).
3.  $x_i \perp x_j$  if  $\lambda_i \neq \lambda_j$  (the eigenvectors of distinct eigenvalues of  $A$  are orthogonal)
4.  $\det(A) > 0$  (the determinant of  $A$  is positive)
5. Every submatrix  $B$  of  $A$  obtained by deleting any set of rows and the corresponding set of columns from  $A$  is symmetric and positive definite (in particular, every principal submatrix is positive definite).

*Proof.* We prove each claim in the theorem as follows

1. Let  $e_i$  be the  $i$ th canonical basis vector in  $\mathbb{R}^n$ . Then

$$a_{ii} = e_i^T A e_i > 0 \quad (40)$$

since  $A$  is pd. A few observations: this only relies on  $A$  being pd.  $e_i^T A$  picks out the  $i$ th row of  $A$ .  $A e_i$  picks out the  $i$ th column of  $A$ .

2. We'll first show that the eigenvalues of  $A$  are real. Suppose  $\lambda, x$  are an eigenvalue/vector pair of  $A$ . Thus  $Ax = \lambda x$ . We can conjugate this equation to find that  $\bar{A}\bar{x} = A\bar{x} = \bar{\lambda}\bar{x}$  (thus complex eigenvalues of real valued matrices come in conjugate pairs). Then

$$\begin{aligned} x^T A \bar{x} &= \bar{\lambda} x^T \bar{x} \\ x^T A^T \bar{x} &= (Ax)^T \bar{x} = \lambda x^T \bar{x} \end{aligned}$$

Since  $A = A^T$ , we know that  $\lambda x^T \bar{x} = \bar{\lambda} x^T \bar{x}$ . As long as  $x \neq 0$ , then  $x^T \bar{x} \neq 0$ . Therefore  $\bar{\lambda} = \lambda$ , which shows  $\lambda \in \mathbb{R}$ .

The fact that the eigenvector associated with  $\lambda$  has real elements follows by noting that all elements of the singular matrix  $A - \lambda I$  are real numbers. Therefore, the columns of  $A - \lambda I$  are linearly dependent in  $\mathbb{R}^n$ . Hence there exists an  $x \in \mathbb{R}^n$  such that  $(A - \lambda I)x = 0$ .

This proof only requires that  $A$  is symmetric – therefore any real, symmetric matrix has real eigenvalues/vectors.

Next we'll show the eigenvalues of  $A$  are positive. Suppose  $\lambda, x$  are an eigenvalue/vector pair of  $A$ . Then

$$0 < x^T A x = \lambda x^T x \quad (41)$$

Since  $x \neq 0$  and  $x^T x$  is positive (it's actually the squared 2-norm of  $x$ ), then  $\lambda > 0$ . Note that this part of the proof requires  $A$  be pd.

3. Let  $\lambda_i, \lambda_j$  be distinct eigenvalues of  $A$ , and  $\mathbf{x}_i, \mathbf{x}_j$  the corresponding eigenvectors. Then

$$\begin{aligned}\mathbf{x}_i^T A \mathbf{x}_j &= \lambda_j \mathbf{x}_i^T \mathbf{x}_j \\ \mathbf{x}_i^T A^T \mathbf{x}_j &= (A \mathbf{x}_i)^T \mathbf{x}_j = \lambda_i \mathbf{x}_i^T \mathbf{x}_j\end{aligned}$$

Since  $A$  is symmetric, these two string of equalities must be equal. We can subtract them to find that

$$(\lambda_i - \lambda_j) \mathbf{x}_i^T \mathbf{x}_j = 0 \quad (42)$$

Since we assumed  $\lambda_i \neq \lambda_j$ , then it must be that  $\mathbf{x}_i^T \mathbf{x}_j = 0$ . Therefore  $\mathbf{x}_i \perp \mathbf{x}_j$ . This proof again only relies on the symmetry of  $A$ .

4. This follows from the fact that the determinant of  $A$  is equal to the product of its eigenvalues.
- 5.

□

## 4.2 Banded Matrices and Differential Equations

## 4.3 Cholesky Factorization

Notation:

- $L_{k-1}$ : the first  $k-1 \times k-1$  upper left corner of  $L$
- $a_k$ : the first  $k-1$  entries in column  $k$  of  $A$
- $l_k$ : the first  $k-1$  entries in column  $k$  of  $L^T$  [[?]]
- $a_{kk}, l_{kk}$ : the  $kk$  entries of  $A$  and  $L$ , respectively

---

### Algorithm 1 Cholesky Factorization

---

**Require:**  $A \in \mathbb{R}^{n \times n}$ , SPD

- 1:  $L_1 \leftarrow \sqrt{a_{11}}$
  - 2: **for**  $k \leftarrow 2, 3, \dots, n$  **do**
  - 3:   Solve  $L_{k-1} l_k = a_k$  for  $l_k$
  - 4:    $l_{kk} \leftarrow \sqrt{a_{kk} - l_k^T l_k}$
  - 5:    $L_k \leftarrow \begin{pmatrix} L_{k-1} & 0 \\ l_k^T & l_{kk} \end{pmatrix}$
  - 6: **end for**
-

## 5 Simultaneous nonlinear equations

### 5.1 Analysis Preliminaries

**Definition 15.** (Cauchy Sequence) A sequence  $(\mathbf{x}^{(k)}) \subset \mathbb{R}^n$  is called a Cauchy sequence in  $\mathbb{R}^n$  if for any  $\epsilon > 0$  there exists a positive integer  $k_0 = k_0(\epsilon)$  such that

$$\|\mathbf{x}^{(k)} - \mathbf{x}^{(m)}\|_\infty < \epsilon \quad \forall k, m \geq k_0(\epsilon) \quad (43)$$

**Remark 1.**  $\mathbb{R}^n$  is **complete** in the sense that every Cauchy sequence  $(\mathbf{x}^{(k)})$  converges to some  $\xi \in \mathbb{R}^n$ .

**Definition 16.** (Continuous function) Let  $D \subset \mathbb{R}^n$  be nonempty and  $f : D \rightarrow \mathbb{R}^n$ . Given  $\xi \in D$ ,  $f$  is continuous at  $\xi$  if for every  $\epsilon > 0$ , there exists a  $\delta = \delta(\epsilon) > 0$  such that for every  $x \in B(\xi; \delta) \cap D$

$$\|f(x) - f(\xi)\|_\infty < \epsilon \quad (44)$$

**Lemma 1.** Let  $D \subset \mathbb{R}^n$  be nonempty and  $f : D \rightarrow \mathbb{R}^n$  be defined and continuous on  $D$ . If  $(\mathbf{x}^{(k)}) \subset D$  converges in  $\mathbb{R}^n$  to  $\xi \in D$ , then  $f(\mathbf{x}^{(k)})$  also converges to  $f(\xi)$ .

### 5.2 Simultaneous iteration

**Definition 17.** (Lipschitz condition, constant, and contraction) Let  $D$  be a closed subset of  $\mathbb{R}^n$  and  $g : D \rightarrow D$ . If there exists a positive constant  $L$  such that

$$\|g(x) - g(y)\|_\infty \leq L\|x - y\|_\infty \quad (45)$$

for all  $x, y \in D$ , then  $g$  satisfies the Lipschitz condition on  $D$  in the  $\infty$ -norm.  $L$  is called the Lipschitz constant. If  $L \in (0, 1)$ , then  $g$  is called a contraction on  $D$  in the  $\infty$ -norm.

Observations about this definition:

- Any function  $g$  that satisfies the Lipschitz condition on  $D$  is continuous on  $D$  (to see this, set  $\delta = \frac{\epsilon}{L}$ ).
- If  $g$  satisfies the Lipschitz condition on  $D$  in the  $\infty$ -norm, then it also does in the  $p$ -norm for  $p \in [1, \infty)$  and vice-versa. However the size of  $L$  depends on the choice of norm.

**Theorem 11.** (Contraction Mapping Theorem in  $\mathbb{R}^n$ ) Suppose  $D$  is a closed subset of  $\mathbb{R}^n$  and  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  is defined on  $D$ , and  $g(D) \subset D$ . Suppose further that  $g$  is a contraction on  $D$  in the  $\infty$ -norm. Then,

1.  $g$  has a unique fixed point  $\xi \in D$
2. The sequence  $(\mathbf{x}^{(k)})$  defined by  $\mathbf{x}^{(k+1)} = g(\mathbf{x}^{(k)})$  converges to  $\xi$  for any starting value  $\mathbf{x}^{(0)} \in D$ .



*Proof.* The proof has three parts:

1. First prove uniqueness, assuming existence of a fixed point.
2. Prove the iteration generates a Cauchy sequence (then convergence to some  $\xi$  follows from the completeness of the space).
3. Show  $\xi$  is indeed the fixed point.

Uniqueness: Suppose  $\xi, \eta$  are both fixed points of  $g$  in  $D$ . Then,

$$\begin{aligned}\|\xi - \eta\|_\infty &= \|g(\xi) - g(\eta)\| && (\xi, \eta \text{ are fixed points}) \\ &\leq L\|\xi - \eta\|_\infty && (g \text{ is a contraction on } D)\end{aligned}$$

We can rearrange this to see that  $(1 - L)\|\xi - \eta\|_\infty \leq 0$ . By assumption,  $L \in (0, 1)$ , and the norm of a quantity is always weakly positive. Therefore,  $\|\xi - \eta\|_\infty = 0$  which implies  $\xi = \eta$ .

Convergence: Assuming  $g$  has a fixed point  $\xi \in D$ , the sequence  $\mathbf{x}^{(k+1)} = g(\mathbf{x}^k)$  will converge to  $\xi$  for any  $\mathbf{x}^{(0)} \in D$ . This follows because

$$\|\mathbf{x}^{(k)} - \xi\|_\infty \leq L^k \frac{1}{1 - L} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty \quad (46)$$

Since  $L \in (0, 1)$ ,  $\lim_{k \rightarrow \infty} L^k = 0$ , and therefore

$$\lim_{k \rightarrow \infty} \|\mathbf{x}^{(k)} - \xi\|_\infty = 0 \quad (47)$$

Existence: First observe that if  $\mathbf{x}^{(0)}$  belongs to  $D$ , then  $\mathbf{x}^{(k+1)} = g(\mathbf{x}^k) \in D$  for all  $k \geq 1$  since  $g(D) \subset D$  (this is important since the proof relies on  $g$  being a contraction on  $D$ ). Next, consider the distance between two adjacent terms on the sequence  $\mathbf{x}^{(k+1)} = g(\mathbf{x}^k)$

$$\begin{aligned}\|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|_\infty &= \|g(\mathbf{x}^{(k-1)}) - g(\mathbf{x}^{(k-2)})\|_\infty && (\text{definition of } g) \\ &\leq L\|\mathbf{x}^{(k-1)} - \mathbf{x}^{(k-2)}\|_\infty && (g \text{ is a contraction on } D) \\ &\leq L^{k-1}\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty && (\text{induction})\end{aligned}$$

Now, fix positive integers  $m, k$  such that  $m > k$ . Then

$$\begin{aligned}\|\mathbf{x}^{(m)} - \mathbf{x}^{(k)}\|_\infty &= \|\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)} + \mathbf{x}^{(m-1)} + \dots + \mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty \\ &\leq \|\mathbf{x}^{(m)} - \mathbf{x}^{(m-1)}\|_\infty + \dots + \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|_\infty && (\text{triangle inequality}) \\ &\leq (L^{m-1} + \dots + L^k)\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty && (g \text{ a contraction}) \\ &= L^k(L^{m-k-1} + \dots + 1)\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty \\ &\leq L^k \frac{1}{1 - L} \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\|_\infty && (\text{geometric series})\end{aligned}$$

Since  $L \in (0, 1)$ ,  $\lim_{k \rightarrow \infty} L^k = 0$ . Therefore,  $\mathbf{x}^{(k)}$  is a Cauchy sequence in  $\mathbb{R}^n$ , that is for all  $\epsilon > 0$ , there exists a  $k_0$  such that

$$\|\mathbf{x}^{(m)} - \mathbf{x}^{(k)}\|_\infty < \epsilon \quad \forall m, k \geq k_0 \quad (48)$$

Any Cauchy sequence in  $\mathbb{R}^n$  is convergent in  $\mathbb{R}^n$ . Thus, there exists some  $\boldsymbol{\xi} \in \mathbb{R}^n$  such that  $\boldsymbol{\xi} = \lim_{k \rightarrow \infty} \mathbf{x}^{(k)}$ .

$\boldsymbol{\xi}$  is indeed the fixed point: Since  $g$  satisfies the Lipschitz condition on  $D$ ,  $g$  is continuous on  $D$ . Therefore,

$$\boldsymbol{\xi} = \lim_{k \rightarrow \infty} \mathbf{x}^{(k+1)} = \lim_{k \rightarrow \infty} g(\mathbf{x}^{(k)}) = g\left(\lim_{k \rightarrow \infty} \mathbf{x}^{(k)}\right) = g(\boldsymbol{\xi}) \quad (49)$$

therefore  $\boldsymbol{\xi}$  is a fixed point of  $g$ , and observe that  $\boldsymbol{\xi} \in D$  since  $D$  is closed.  $\square$

**Definition 18.** (Jacobian) Let  $g = (g_1, \dots, g_n)^T : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a function defined and continuous in an (open) neighborhood of  $\boldsymbol{\xi} \in \mathbb{R}^n$ . Suppose the first partial derivatives of each  $g_i$  exist at  $\boldsymbol{\xi}$ . The Jacobian matrix  $J_g(\boldsymbol{\xi})$  of  $g$  at  $\boldsymbol{\xi}$  is the  $n \times n$  matrix with elements

$$J_g(\boldsymbol{\xi})_{ij} = \frac{\partial g_i}{\partial x_j}(\boldsymbol{\xi}) \quad (50)$$

**Theorem 12.** Let  $g = (g_1, \dots, g_n)^T : \mathbb{R}^n \rightarrow \mathbb{R}^n$  be a function defined and continuous on a closed set  $D \subset \mathbb{R}^n$ . Let  $\boldsymbol{\xi} \in D$  be a fixed point of  $g$ . Suppose the first partial derivatives of each  $g_i$  are defined and continuous in some (open) neighborhood  $N(\boldsymbol{\xi}) \in D$  of  $\boldsymbol{\xi}$ , with

$$\|J_g(\boldsymbol{\xi})\|_\infty < 1 \quad (51)$$

Then there exists  $\epsilon > 0$  such that  $g(\bar{B}_\epsilon(\boldsymbol{\xi})) \subset \bar{B}_\epsilon(\boldsymbol{\xi})$ , and the sequence  $\mathbf{x}^{(k+1)} = g(\mathbf{x}^{(k)})$  converges to  $\boldsymbol{\xi}$  for all  $\mathbf{x}^{(0)} \in \bar{B}_\epsilon(\boldsymbol{\xi})$  (in other words, the sequence converges to  $\boldsymbol{\xi}$  as long as  $\mathbf{x}^{(0)}$  is close enough to  $\boldsymbol{\xi}$ ).

**Example 3.**

## Newton's Method

**Definition 19.** (Newton's Method) The sequence defined by

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - [J_f(\mathbf{x}^{(k)})]^{-1} f(\mathbf{x}^{(k)}) \quad (52)$$

where  $\mathbf{x}^{(0)} \in \mathbb{R}^n$ , is called Newton's method.

**Theorem 13.** Suppose  $f(\boldsymbol{\xi}) = 0$ , that in some (open) neighborhood  $N(\boldsymbol{\xi})$  of  $\boldsymbol{\xi}$ , where  $f$  is defined and continuous, all the second-order partial derivatives of  $f$  are defined and continuous, and that the Jacobian matrix  $J_f(\mathbf{x}^{(k)})$  of  $f$  at the point  $\boldsymbol{\xi}$  is nonsingular. Then the sequence defined by Newton's method converges to  $\boldsymbol{\xi}$  provided that  $\mathbf{x}^{(0)}$  is sufficiently close to  $\boldsymbol{\xi}$ .

## 6 Eigenvalues of Eigenvectors of a symmetric matrix

### 6.1 Why we use iteration to calculate eigenvalues/eigenvectors

We call  $\lambda$  an eigenvalue and  $x \neq 0$  an eigenvector of  $A$  if  $Ax = \lambda x$ . Thus,  $(Ax - \lambda I)x = 0$ . Therefore,  $x \in \text{Null}(A - \lambda I)$ . Since  $x \neq 0$ ,  $A - \lambda I$  has a non-trivial nullspace, so we must have  $\det(A - \lambda I) = 0$ . This suggests a way to transform an eigenvalue finding problem to a root finding problem. Define

$$\rho(\lambda) = \det(A - \lambda I) \quad (53)$$

Recall that the determinant of a matrix is the product of its eigenvalues. If  $A$  is a  $n \times n$  real, symmetric matrix, then  $\rho(\lambda)$  is an  $n$ -th order polynomial in  $\lambda$ , whose roots are the eigenvalues of  $A$ .

**Theorem 14.** (Abel(-Ruffini) Theorem, or “No-go Theorem”) *There is no algebraic solution (that is, a solution expressed in terms of radicals) to general polynomial equations of degree five or higher with arbitrary coefficients.*

Therefore, there is no finite-number operation procedure that provides an eigenvalue decomposition.

### 6.2 Power Iteration

Find the biggest eigenvalue/vector.

---

#### Algorithm 2 Power Iteration

---

**Require:**  $v^{(0)}$  = some vector with  $\|v^{(0)}\| = 1$

- 1: **for**  $k \leftarrow 1, 2, \dots$  **do**
  - 2:    $w \leftarrow Av^{(k-1)}$  ▷ Apply  $A$
  - 3:    $v^{(k)} \leftarrow w / \|w\|$  ▷ Normalize
  - 4:    $\lambda^{(k)} \leftarrow (v^{(k)})^T Av^{(k)} = \langle v^{(k)}, Av^{(k)} \rangle$  ▷ Rayleigh Quotient
  - 5: **end for**
- 

**Theorem 15.** (Convergence of Power Iteration) *Suppose  $|\lambda_1| > |\lambda_2| \geq \dots \geq |\lambda_n|$  and  $q_1^T v^{(0)} \neq 0$ . Then the iterates of power iteration satisfy*

$$\|v^{(k)} - (\pm q_1)\| = \mathcal{O} \left( \left| \frac{\lambda_2}{\lambda_1} \right|^k \right) \quad (\text{error of eigenvector})$$

$$|\lambda^{(k)} - \lambda_1| = \mathcal{O} \left( \left| \frac{\lambda_2}{\lambda_1} \right|^2 k \right) \quad (\text{error of eigenvalue})$$

*Proof.* Convergence of eigenvector: Write  $v^{(0)} = v$  as a linear combination of the orthonormal eigenvectors  $q_i$ :

$$v = c_1 q_1 + \dots + c_n q_n \quad (54)$$

$v^{(k)}$  is a scalar multiple of  $A^k v^{(0)}$ . Therefore

$$\begin{aligned} v^{(k)} &= \alpha_k A^k v^{(0)} && (\alpha_k \text{ a normalization constant}) \\ &= \alpha_k \left( \sum_{i=1}^n \lambda_i^k a_i q_i \right) \\ &= \alpha_k \lambda_1^k \left( c_1 q_1 + c_2 \left( \frac{\lambda_2}{\lambda_1} \right)^k q_2 + \dots + c_n \left( \frac{\lambda_n}{\lambda_1} \right)^k q_n \right) \end{aligned}$$

We can choose  $\alpha_k$  such that  $\alpha_k \lambda_1^k$  is 1. Therefore,  $c_1 q_1$  is dominating (as long as  $c_1 \neq 0$ ). The other terms are of order  $\mathcal{O} \left( \left| \frac{\lambda_2}{\lambda_1} \right|^k \right)$ .

Convergence of eigenvalue: [[Incomplete]] □

**Theorem 16.** (Error of Rayleigh Quotient) Let  $x_1$  be the eigenvector that corresponds to the largest (in absolute value) eigenvalue. If  $\|x - x_1\| = \mathcal{O}(\epsilon)$ , then

$$\left| \frac{\langle x, Ax \rangle}{\langle x, x \rangle} - \lambda_1 \right| = \mathcal{O}(\epsilon^2) \quad (55)$$

### 6.3 Inverse Iteration

Find the smallest eigenvalue/vector.

### 6.4 Simultaneous Iteration

Obtain the full set of eigenvalues/vectors simultaneously.

---

#### Algorithm 3 Simultaneous Iteration

---

**Require:**  $Q^{(0)} = V = I$ , a list of vectors  $V$ , which we choose to be the identity

- 1: **for**  $k \leftarrow 1, 2, \dots$  **do**
  - 2:    $Z \leftarrow A Q^{(k-1)}$  ▷ Apply  $A$
  - 3:    $Z \leftarrow \underline{Q}^{(k)} R^{(k)}$  ▷ QR factorization of  $Z$
  - 4:    $A^{(k)} \leftarrow (\underline{Q}^{(k)})^T A Q^{(k)}$
  - 5: **end for**
-

## 6.5 Shifted Power Iteration

Find the eigenvalue close to a specific number.

## 6.6 QR Algorithm

The QR can be viewed as a stable procedure for computing QR factorizations of the matrix powers  $A, A^2, A^3, \dots$

---

### Algorithm 4 QR Algorithm (without shifts)

---

**Require:**  $A^{(0)} = A$

- 1: **for**  $k \leftarrow 1, 2, \dots$  **do**
  - 2:    $Q^{(k)}R^{(k)} \leftarrow A^{(k-1)}$  ▷ QR factorization of  $A^{(k-1)}$
  - 3:    $A^{(k)} \leftarrow R^{(k)}Q^{(k)}$  ▷ Recombine factors in reverse order
  - 4: **end for**
- 

## 6.7 Simultaneous Iteration equivalent to QR Algorithm

The QR algorithm is equivalent to simultaneous iteration applied to a full set of initial vectors, namely,  $\hat{Q}^{(0)} = I$ . Summary of each algorithm:

### Simultaneous Iteration

$$\begin{aligned}\underline{Q}^{(0)} &= I && \text{(initial condition)} \\ Z &= A\underline{Q}^{(k-1)} && \text{(apply } A) \\ Z &= \underline{Q}^{(k)}R^{(k)} && \text{(resemblance of normalization, QR factorization of } Z) \\ A^{(k)} &= (\underline{Q}^{(k)})^T A \underline{Q}^{(k)} && \text{(resemblance of Rayleigh quotient)}\end{aligned}$$

### QR Algorithm

$$\begin{aligned}A^{(0)} &= A && \text{(initial condition)} \\ A^{(k-1)} &= Q^{(k)}R^{(k)} && \text{(compute QR factorization)} \\ A^{(k)} &= R^{(k)}Q^{(k)} && \text{(reverse order of factors)} \\ \underline{Q}^{(k)} &= Q^{(1)}Q^{(2)} \dots Q^{(k)} && \text{(definition of } \underline{Q}^{(k)})\end{aligned}$$

and

$$\underline{R}^{(k)} = R^{(k)}R^{(k-1)} \dots R^{(1)} \quad \text{(definition of } \underline{R}^{(k)})$$

**Theorem 17.** (*Equivalence of Simultaneous Iteration and the QR Algorithm*) Simultaneous Iteration and the QR Algorithm generate identical sequences of matrices  $\underline{R}^{(k)}, \underline{Q}^{(k)}, A^{(k)}$ . Both give

$$(a) : A^{(k)} = \underline{Q}^{(k)} \underline{R}^{(k)} \quad (\text{QR factorization of the } k\text{th power of } A)$$

$$(b) : A^{(k)} = (\underline{Q}^{(k)})^T A \underline{Q}^{(k)} \quad (\text{projection})$$

*Proof.* By induction on  $k$  (number of iterations). The base case  $k = 0$  is trivial.

1. QR gives (a): Assume  $A^{(k-1)} = \underline{Q}^{(k-1)} \underline{R}^{(k-1)}$ . The inductive hypothesis for (b) gives that  $A^{(k-1)} = (\underline{Q}^{(k-1)})^T A \underline{Q}^{(k-1)}$  or that  $\underline{Q}^{(k-1)} A^{(k-1)} = A \underline{Q}^{(k-1)}$ . Then

$$\begin{aligned} A^{(k)} &= A A^{(k-1)} && (\text{decompose to use inductive hypothesis}) \\ &= A \underline{Q}^{(k-1)} \underline{R}^{(k-1)} && (\text{inductive hypothesis}) \\ &= \underline{Q}^{(k-1)} A^{(k-1)} \underline{R}^{(k-1)} && (\text{inductive hypothesis from (b)}) \\ &= \underline{Q}^{(k-1)} \underline{R}^{(k-1)} \underline{Q}^{(k-1)} \underline{R}^{(k-1)} && (\text{from algorithm}) \\ &= \underline{Q}^{(k)} \underline{R}^{(k)} && (\text{from definitions of } \underline{Q}^{(k)}, \underline{R}^{(k)}) \end{aligned}$$

2. QR gives (b): Assume  $A^{(k-1)} = (\underline{Q}^{(k-1)})^T A \underline{Q}^{(k-1)}$ . From the relationship  $A^{(k-1)} = \underline{Q}^{(k-1)} \underline{R}^{(k-1)}$  and the fact that  $\underline{Q}^{(k)}$  is orthogonal, we can apply  $(\underline{Q}^{(k)})^T$  to both sides (on the left) to get that  $(\underline{Q}^{(k)})^T A^{(k-1)} = \underline{R}^{(k)}$ . Then

$$\begin{aligned} A^{(k)} &= \underline{R}^{(k)} \underline{Q}^{(k)} \\ &= (\underline{Q}^{(k)})^T A^{(k-1)} \underline{Q}^{(k)} \\ &= (\underline{Q}^{(k)})^T (\underline{Q}^{(k-1)})^T A \underline{Q}^{(k-1)} \underline{Q}^{(k)} && (\text{inductive hypothesis}) \\ &= (\underline{Q}^{(k)})^T A \underline{Q}^{(k)} && (\text{definition of } \underline{Q}^{(k)}) \end{aligned}$$

□