LECTURE NOTES
on
LINEAR ALGEBRA, LINEAR DIFFERENTIAL EQUATIONS,
and
LINEAR SYSTEMS

A. S. Morse

August 29, 2020
Contents

I Linear Algebra and Linear Differential Equations 1

1 Fields and Polynomial Rings 3

2 Matrix Algebra 7

2.1 Matrices: Notation 7
2.2 Types of Matrices 8
2.3 Determinants: Review 9
2.4 Matrix Rank 9
2.5 Basic Matrix Operations 10
2.5.1 Transposition 10
2.5.2 Scalar Multiplication 11
2.5.3 Addition 11
2.5.4 Multiplication 12
2.5.5 Linear Algebraic Equations 14
2.5.6 Matrix Inversion 15

2.6 Linear Recursion Equations 17

2.7 Invariants under Premultiplication 19
2.7.1 Elementary Row Operations 20
2.7.2 Echelon Form 21
2.7.3 Gauss Elimination 22
2.7.4 Elementary Column Operations 27
2.7.5 Equivalent Matrices 28

3 Linear Algebra 31
3.1 Linear Vector Spaces ................................................................. 31
  3.1.1 Subspaces ................................................................. 32
  3.1.2 Subspace Operations ....................................................... 33
  3.1.3 Distributive Rule .......................................................... 34
  3.1.4 Independent Subspaces and Direct Sum ................................ 36
  3.1.5 Linear Combination and Span .......................................... 36
  3.1.6 Linear Independence ...................................................... 36
  3.1.7 Basis ................................................................. 37
  3.1.8 Dimension ................................................................. 40

3.2 Functions .............................................................................. 40
  3.2.1 Linear Transformations ..................................................... 41
  3.2.2 Operations With Linear Transformations ............................. 42
  3.2.3 Representations of Linear Transformations .......................... 42
  3.2.4 Coordinate Transformations ............................................. 43
  3.2.5 Defining a Linear Transformation on a Basis ....................... 44
  3.2.6 Linear Equations: Existence, Image, Epimorphism ............... 45
  3.2.7 Linear Equations: Uniqueness, Kernel, Monomorphism .......... 47
  3.2.8 Isomorphisms and Isomorphic Vector Spaces ....................... 48
  3.2.9 Endomorphisms and Similar Matrices ................................ 48

4 Basic Concepts from Analysis .................................................... 51
  4.1 Normed Vector Spaces ....................................................... 51
    4.1.1 Open Sets ............................................................... 53
  4.2 Continuous and Differentiable Functions ................................ 53
  4.3 Convergence ........................................................................ 55

5 Ordinary Differential Equations - First concepts ......................... 57
  5.1 Types of Equation ............................................................... 57
  5.2 Modeling ............................................................................. 59
  5.3 State Space Systems ............................................................ 62
    5.3.1 Conversion to State-Space Form ..................................... 63
  5.4 Initial Value Problem ......................................................... 65
5.5 Picard Iterations ................................................................. 68
  5.5.1 Three Relationships ...................................................... 69
  5.5.2 Convergence ............................................................... 70
  5.5.3 Uniqueness ................................................................. 72
  5.5.4 Summary of Findings .................................................... 74
5.6 The Concept of State ......................................................... 77
  5.6.1 Trajectories, Equilibrium Points and Periodic Orbits ................. 78
  5.6.2 Time-Invariant Dynamical Systems .................................... 80
5.7 Simulation ............................................................................ 81
  5.7.1 Numerical Techniques ...................................................... 81
  5.7.2 Analog Simulation .......................................................... 83
  5.7.3 Programming Diagrams .................................................... 84

6 Linear Differential Equations ................................................. 87
  6.1 Linearization ..................................................................... 87
  6.2 Linear Homogenous Differential Equations ............................... 90
    6.2.1 State Transition Matrix ................................................ 91
    6.2.2 The Matrix Exponential ............................................... 92
    6.2.3 State Transition Matrix Properties ................................... 93
  6.3 Forced Linear Differential Equations ....................................... 96
    6.3.1 Variation of Constants Formula ...................................... 96
  6.4 Periodic Homegeneous Differential Equations ........................... 97
    6.4.1 Coordinate Transformations .......................................... 98
    6.4.2 Lyapunov Transformations ............................................ 99
    6.4.3 Floquet's Theorem .................................................... 100

7 Matrix Similarity ................................................................. 103
  7.1 Motivation .......................................................................... 103
    7.1.1 The Matrix Exponential of a Diagonal Matrix ..................... 104
    7.1.2 The State Transition Matrix of a Diagonalizable Matrix ......... 104
  7.2 Eigenvalues ....................................................................... 105
    7.2.1 Characteristic Polynomial ............................................ 105
8.3 Gram's Criterion .......................................................... 134
8.4 Cauchy-Schwartz Inequality ........................................... 135
8.5 Orthogonalization of a Set of Vectors ............................... 136
8.6 Orthogonal Projections .................................................. 137
8.7 Bessel's Inequality ....................................................... 138
8.8 Least Squares ............................................................ 139

9 Normal Matrices .................................................................. 143

9.1 Algebraic Groups ........................................................... 143
  9.1.1 Orthogonal Group .................................................... 144
  9.1.2 Unitary Group ........................................................ 145

9.2 Adjoint Matrix ............................................................... 145
  9.2.1 Orthogonal Complement ........................................... 145
  9.2.2 Properties of Adjoint Matrices ................................. 146

9.3 Normal Matrices ............................................................ 146
  9.3.1 Diagonalization of a Normal Matrix ......................... 147

9.4 Hermitian, Unitary, Orthogonal and Symmetric Matrices .... 149
  9.4.1 Hermitian Matrices .................................................. 149
  9.4.2 Unitary Matrices ...................................................... 150
  9.4.3 Orthogonal Matrices ................................................. 150
  9.4.4 Symmetric Matrices ................................................ 150

9.5 Real Quadratic Forms ..................................................... 152
  9.5.1 symmetric Quadratic Forms ..................................... 153
  9.5.2 Change of Variables: Congruent Matrices ................. 153
  9.5.3 Reduction to Principal Axes .................................... 153
  9.5.4 Sum of Squares ...................................................... 154

9.6 Positive Definite Quadratic Forms .................................. 156
  9.6.1 Conditions for Positive Definiteness ....................... 156
  9.6.2 Matrix Square Roots .............................................. 157

9.7 Simultaneous Diagonalization ......................................... 158
  9.7.1 Constrained Optimization ...................................... 159
II Linear Systems

10 Introductory Concepts

10.1 Linear Systems

10.2 Continuous-Time Linear Systems

10.2.1 Time-Invariant, Continuous-Time, Linear Systems

10.3 Discrete-Time Linear Systems

10.3.1 Time-Invariant Discrete Time Systems

10.3.2 Sampled Data Systems

10.4 The Concept of a Realization

10.4.1 Existence of Continuous-Time Linear Realizations

11 Controllability

11.1 Reachable States

11.2 Controllability

11.2.1 Controllability Reduction

11.3 Time-Invariant Continuous-Time Linear Systems

11.3.1 Properties of the Controllable Space of \((A, B)\)

11.3.2 Control Reduction for Time-Invariant Systems

11.3.3 Controllable Decomposition

12 Observability

12.1 Unobservable States

12.2 Observability

12.2.1 Observability Reduction

12.3 Minimal Systems

12.4 Time-Invariant Continuous-Time Linear Systems

12.4.1 Properties of the Unobservable Space of \((C, A)\)

12.4.2 Observability Reduction for Time-Invariant Systems

12.4.3 Observable Decomposition
Part I

Linear Algebra and Linear Differential Equations
An algebraic system \{e.g., field, ring, vector space\} is a set together with certain operations under which the set is closed. Each algebraic system is defined by a list of postulates. We denote by $\mathbb{R}$ the algebraic system consisting of the set of all real numbers together with the operations of ordinary addition $\{+\}$ and multiplication $\{\cdot\}$. Let us note that addition and multiplication in $\mathbb{R}$ are each associative

\[
r_1 + (r_2 + r_3) = (r_1 + r_2) + r_3
\]

and commutative

\[
r_1 + r_2 = r_2 + r_1
\]

\[
r_1 \cdot r_2 = r_2 \cdot r_1
\]

operations and that multiplication distributes over addition:

\[
r \cdot (r_1 + r_2) = r \cdot r_1 + r \cdot r_2.
\]

Furthermore, $\mathbb{R}$ contains distinct additive and multiplicative units, namely 0 and 1 respectively, which satisfy $0 + r = r$, $1 \cdot r = r$, $\forall r \in \mathbb{R}$. Finally, for each nonzero $r \in \mathbb{R}$, there exist additive and multiplicative inverses, namely $-r$ and $r^{-1}$ respectively, such that $r + (-r) = 0$ and $r \cdot r^{-1} = 1$. These properties are enough to guarantee that $\mathbb{R}$ is an algebraic field.

The set of all complex numbers together with ordinary complex addition and complex multiplication is also an algebraic field since it has all of the properties outlined above. The complex field is denoted by $\mathbb{C}$. Although much of what follows holds for arbitrary fields, we shall limit the scope of these notes to just $\mathbb{R}$ and $\mathbb{C}$. The symbol $\mathbb{K}$ is sometimes used to denote either $\mathbb{R}$ or $\mathbb{C}$ when discussing a concept valid for both fields.

**Example 1** The set consisting of just the integers 0 and 1 together with addition $\{+\}$ and multiplication $\{\cdot\}$ operations defined by

\[
\begin{align*}
0 + 1 & \triangleq 1 \\
1 + 0 & \triangleq 1 \\
0 + 0 & \triangleq 0 \\
1 + 1 & \triangleq 0
\end{align*}
\]

and

\[
\begin{align*}
0 \cdot 1 & \triangleq 0 \\
1 \cdot 0 & \triangleq 0 \\
0 \cdot 0 & \triangleq 0 \\
1 \cdot 1 & \triangleq 1
\end{align*}
\]
is a field. This field, called GF(2) (Galois Field 2) or the field of integers modulo 2, is used in the coding of messages in communication systems. Note that all of these definitions are the same as the standard ones from integer arithmetic except for the definition of 1 + 1. To ensure closure, this sum must be defined to be either 0 or 1. If 1 + 1 \( \equiv 1 \) rather than as above, the resulting algebraic system would not be a field because there would be no additive inverse for 1; i.e., there would be no number in the set \( \{0, 1\} \) which, when added to 1 would give the additive unit 0. In fact if one were to define 1 + 1 \( \equiv 1 \) rather than 1 + 1 \( \equiv 0 \), what one would have would be the “Boolean Algebra” upon which digital logic is based; in this case the symbols 1, 0, + and \( \cdot \) would denote true, false, or and and or, respectively.

Consider next the set of all polynomials \( \rho(s) \) with indeterminate (i.e., ‘variable’) \( s \) and coefficients in \( \mathbb{R} \); i.e.,

\[
\rho(s) = r_{n+1}s^n + r_n s^{n-1} + \ldots + r_2s + r_1,
\]

where \( r_i \in \mathbb{R}, n = \deg \rho(s) \) and \( r_{n+1} \neq 0 \). Include in this set all polynomials of degree zero and the zero polynomial. Let \( \mathbb{R}[s] \) denote the algebraic system consisting of all such polynomials together with ordinary polynomial addition and multiplication. Note that \( \mathbb{R}[s] \) possesses all of the algebraic properties which defined the field \( \mathbb{R} \) except one: there are nonzero elements in \( \mathbb{R}[s] \) which don’t have multiplicative inverses \( \{\text{in } \mathbb{R}[s]\} \). In particular, if \( \rho(s) \in \mathbb{R}[s] \) is any polynomial of positive degree, there is no polynomial \( \gamma(s) \in \mathbb{R}[s] \) such that \( \rho(s) \gamma(s) = 1 \). It is true that \( \rho(s)(\rho(s)^{-1}) = 1 \), but \( \rho(s)^{-1} \) is not a polynomial since \( \deg \rho(s) > 0 \). Algebraic systems such as \( \mathbb{R}[s] \) which have all field properties except the one just described are called rings. The set of all polynomials with coefficients in \( \mathbb{C} \) together with polynomial addition and multiplication is the ring \( \mathbb{C}[s] \). We often denote either \( \mathbb{R}[s] \) or \( \mathbb{C}[s] \) by \( \mathbb{K}[s] \).

The following elementary property of \( \mathbb{K}[s] \) is important.

**Proposition 1** For each pair of polynomials \( \alpha, \beta \in \mathbb{K}[s] \) with \( \beta \neq 0 \), there exist unique polynomials \( \gamma, \delta \in \mathbb{K}[s] \) such that

\[
\alpha = \beta \gamma + \delta \quad \text{and} \quad \deg \delta < \deg \beta
\]

{The degree of the zero polynomial is defined as \( -\infty \).} The quotient \( \gamma \) and remainder \( \delta \) of the pair \( (\alpha, \beta) \) are sometimes denoted by \( q(\alpha, \beta) \) and \( r(\alpha, \beta) \) respectively. These polynomials can be computed by ordinary polynomial division. In case \( r(\alpha, \beta) = 0 \), i.e., \( \alpha = \beta \gamma \), we say \( \beta \) divides \( \alpha \); if \( \beta \) divides 1, \( \beta \) is invertible. Note that the set of all invertible elements in \( \mathbb{K}[s] \) together with 0 is precisely the set \( \mathbb{K} \) which, in turn, is a subset of \( \mathbb{K}[s] \).

**Example 2** If \( a \equiv s^3 + 3s^2 + 3s + 1 \) and \( b = s^2 + 2s + 3 \), then \( q = s + 1 \) and \( r = -2s - 2 \). To compute these polynomials in Matlab one inputs \( A = [1,3,3,1] \) and \( B = [1,2,3] \) and then uses the command \( [Q,R] = \text{deconv}(A,B) \) to get \( Q = [1,1] \) and \( R = [-2,-2] \). Try it with your favorite polynomials.

A polynomial \( \alpha \in \mathbb{K}[s] \) is prime or irreducible if 1) \( \alpha \not\in \mathbb{K} \) and 2) for each \( \beta \in \mathbb{K}[s] \) such that \( \beta \) divides \( \alpha \), either \( \alpha \) divides \( \beta \) or \( \beta \in \mathbb{K} \). Irreducibles in \( \mathbb{K}[s] \) turn out to be all polynomials of degree one and those of degree two with roots not in \( \mathbb{K} \); irreducibles in \( \mathbb{C}[s] \) are all polynomials of degree one. An \( n \)th degree polynomial \( \alpha(s) = k_{n+1}s^n + k_ns^{n-1} + \ldots + k_1 \) is monic if \( k_{n+1} = 1 \).

A subset \( I \) of \( \mathbb{K}[s] \) is called an ideal if for each pair of polynomials \( \alpha, \beta \in I \), there follows \( \alpha \rho + \beta \delta \in I \), \( \forall \rho, \delta \in \mathbb{K}[s] \). Thus an ideal is a subset of \( \mathbb{K}[s] \) closed under ring addition and multiplication by elements from \( \mathbb{K}[s] \). The set consisting of just the zero polynomial is the zero ideal. Suppose \( I \) is a nonzero ideal and let \( \beta \in I \) be any nonzero polynomial of least \{finite\} degree. Let \( \alpha \in I \) be arbitrary and use Proposition 1 to write \( \alpha = \beta q(\alpha, \beta) + r(\alpha, \beta) \). Clearly \( \beta q(\alpha, \beta) \in I \). Since \( r(\alpha, \beta) = \alpha - \beta q(\alpha, \beta) \), it follows that \( r(\alpha, \beta) \in I \).
But by Proposition 1, \( \deg r(\alpha, \beta) < \deg \beta \). Since \( \beta \) has least degree, it must be that \( r(\alpha, \beta) = 0 \); this means that \( \beta \) divides \( \alpha \). Clearly \( \beta \) is uniquely determined up to an invertible factor in \( \mathbb{K}[s] \). So we may as well make things unique by taking \( \beta \) to be monic. We summarize:

**Proposition 2** Let \( I \) be a nonzero ideal of \( \mathbb{K}[s] \). There exists a unique, monic polynomial \( \beta \in I \) such that for each \( \alpha \in I \),

\[
\alpha = \gamma \beta
\]

for some \( \gamma \in \mathbb{K}[s] \).

Let \( \{\alpha_1, \alpha_2, \ldots, \alpha_m\} \) be a fixed set of polynomials in \( \mathbb{K}[s] \). The greatest common divisor \( \{\gcd\} \) of \( \{\alpha_1, \alpha_2, \ldots, \alpha_m\} \) is the unique monic polynomial of greatest degree which divides each \( \alpha_i \). The least common multiple \( \{\lcm\} \) of \( \{\alpha_1, \alpha_2, \ldots, \alpha_m\} \) is the unique monic polynomial of least degree which each \( \alpha_i \) divides. If \( \alpha \) and \( \beta \) are monic polynomials, then

\[
\gcd(\alpha, \beta) \lcm(\alpha, \beta) = \alpha \beta
\]  

(1.1)

Thus the lcm of two polynomials can easily be computed once the corresponding gcd is known. There are many procedures for computing \( \gcd \{\alpha, \beta\} \), the Euclidean Algorithm [1] being the classical method. The Matlab command for computing the gcd \( G \) of two polynomials \( A \) and \( B \) is \( G = \gcd(A, B) \). Try computing \( \gcd\{s^2 + (2 + \epsilon)s + 2 + \epsilon, s^2 + 5s + 6\} \) for \( \epsilon = 0 \) and also for very small nonzero values of \( \epsilon \).

With \( \alpha, \beta \) as above, consider the set \( I(\alpha, \beta) \equiv \{\alpha \pi + \beta \mu : \pi, \mu \in \mathbb{K}[s]\} \); clearly \( I(\alpha, \beta) \) is an ideal. Thus by Proposition 2, there exists a monic polynomial \( \delta \in I(\alpha, \beta) \) which divides both \( \alpha \) and \( \beta \); thus if \( \gamma \) is the greatest common divisor of \( \alpha \) and \( \beta \) then \( \delta \) divides \( \gamma \). In addition, since \( \delta \in I(\alpha, \beta) \), we can write \( \delta = \alpha \sigma + \beta \rho \) for some \( \sigma, \rho \in \mathbb{K}[s] \); from this and the fact that \( \gamma \) divides both \( \alpha \) and \( \beta \) we conclude that \( \gamma \) divides \( \delta \). It is apparent that if \( \delta \) is monic, then \( \delta = \gamma \) and from this there follows a useful fact.

**Proposition 3** For each pair of polynomials \( \alpha, \beta \in \mathbb{K}[s] \), there exist polynomials \( \sigma, \rho \in \mathbb{K}[s] \) such that

\[
\gcd(\alpha, \beta) = \alpha \sigma + \beta \rho
\]

(1.2)

Note that if \( \alpha \) and \( \beta \) are coprime \( \{\text{i.e., if } \gcd(\alpha, \beta) = 1\} \), then \( \sigma \) and \( \rho \) can be chosen so that \( \alpha \sigma + \beta \rho = 1 \). For any two polynomials \( \alpha \) and \( \beta \), represented in Matlab format by matrices \( A \) and \( B \), the Matlab command for computing \( \gcd(\alpha, \beta) \) together with polynomials \( \sigma \) and \( \rho \) satisfying (1.2), is \( [G, S, R] = \gcd(A, B) \). Here \( G, S \) and \( R \) are Matlab representations of \( \gcd(\alpha, \beta) \), \( \sigma \), and \( \rho \) respectively.
CHAPTER 1. FIELDS AND POLYNOMIAL RINGS
In the sequel are discussed the basic concepts of matrix algebra.

### 2.1 Matrices: Notation

A \( p \times q \) matrix \( M \), over a field \( \mathbb{K} \), is an array of numbers of the form

\[
M = \begin{bmatrix}
  k_{11} & k_{12} & \cdots & k_{1q} \\
  k_{21} & k_{22} & \cdots & k_{2q} \\
  \vdots & \vdots & \ddots & \vdots \\
  k_{p1} & k_{p2} & \cdots & k_{pq}
\end{bmatrix}
\]

The elements or entries \( k_{ij} \) in \( M \), are elements of \( \mathbb{K} \); \( p \times q \) is the size of \( M \); i.e., \( M \) has \( p \) rows and \( q \) columns \{e.g., the 3rd row of \( M \) is \( [k_{31}, k_{32}, \ldots, k_{3q}] \}\). We often use the abbreviated notations

\[
M = [k_{ij}] \quad \text{or} \quad M = [k_{ij}]_{p \times q}
\]

\( M \) is a **square** matrix if \( p=q \); i.e., if \( M \) has the same number of rows as columns. Otherwise \( M \) is **rectangular**.

It is possible to partition a matrix into ‘submatrices’, by drawing lines between some of its rows and columns. For examples, the lines within the matrix

\[
A = \begin{bmatrix}
  1 & 2 & 7 & 7 & 5 \\
  1 & 2 & 1 & 7 & 1 \\
  4 & 3 & 2 & 8 & 1 \\
  6 & 4 & 3 & 5 & 1
\end{bmatrix}
\]

**partitions** \( A \) into the six submatrices, namely

\[
\begin{align*}
A_{11} &= \begin{bmatrix} 1 & 2 \end{bmatrix} & A_{12} &= \begin{bmatrix} 7 & 7 & 5 \end{bmatrix} \\
A_{21} &= \begin{bmatrix} 1 & 2 \end{bmatrix} & A_{22} &= \begin{bmatrix} 1 & 7 & 1 \end{bmatrix} \\
A_{31} &= \begin{bmatrix} 4 & 3 \\ 6 & 4 \end{bmatrix} & A_{23} &= \begin{bmatrix} 2 & 8 & 1 \\ 3 & 5 & 1 \end{bmatrix}
\end{align*}
\]
If we wish, we can then write $A$ as
\[
A = \begin{bmatrix}
A_{11} & A_{12} \\
A_{21} & A_{22} \\
A_{31} & A_{32}
\end{bmatrix}
\]

There are of course many different ways to partition a matrix.

### 2.2 Types of Matrices

There are a number of special types of matrices. Here are a few of the most basic.

**Zero matrices:** Matrices of any size with zeros in every location. Zero matrices are usually denoted by 0.

In Matlab, the command `zeros(n, m)` produces the $n \times m$ zero matrix.

**$n$-vectors:** Matrices with $n$ rows and one column. In these notes such matrices are usually denoted by lower case Roman letters $x, y, z, \ldots$; e.g.,
\[
x = \begin{bmatrix} 1 \\ 1 \\ 5 \\ 0 \end{bmatrix}
\]

**Unit $n$-vectors:** $n$-vectors with exactly one nonzero entry which in turn is a 1.

**Zero vectors:** Zero matrices with one column.

**Square matrices:** Matrices with the same number of rows as columns; e.g.,
\[
A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}
\]

**Diagonal matrices:** Square matrices with zeros everywhere except along the main diagonal; e.g.,
\[
D = \begin{bmatrix} d_1 & 0 & \cdots & 0 \\
0 & d_2 & \cdots & 0 \\
0 & 0 & \ddots & 0 \\
0 & 0 & \cdots & d_p \end{bmatrix}_{p \times p}
\]

If $d \triangleq [d_1 \ d_2 \ \cdots \ d_p]$, the Matlab statement $D = \text{diag}(d)$ inputs the diagonal matrix $D$ above.

**Symmetric matrices:** Any square matrix $A = [a_{ij}]_{n \times n}$ which is symmetric with respect to its main diagonal; i.e., $a_{ij} = a_{ji}$, $\forall \ i, j \in \{1, 2, \ldots, n\}$

**Upper triangular matrices:** Square matrices with zeros everywhere below the main diagonal; e.g.,
\[
U = \begin{bmatrix}
u_{11} & u_{12} & \cdots & u_{1p} \\
0 & u_{22} & \cdots & u_{2p} \\
0 & 0 & \ddots & \vdots \\
0 & 0 & \cdots & u_{pp}
\end{bmatrix}_{p \times p}
\]

A *lower triangular matrix* is similarly defined.
Identity matrices: Diagonal matrices with 1’s along the main diagonal. Identity matrices are usually denoted by \( I \); e.g.,

\[
I_{3 \times 3} = \begin{bmatrix}
1 & 0 & 0 \\
0 & 1 & 0 \\
0 & 0 & 1
\end{bmatrix}
\]

The Matlab statement \( I = \text{eye}(n) \) defines \( I \) to be the \( n \times n \) identity. A list of Matlab commands defining other elementary matrix can be seen by typing \texttt{help elmat}.

### 2.3 Determinants: Review

Let \( A = [a_{ij}]_{n \times n} \) be a fixed matrix. The \textit{cofactor} of the \( ij \)th element \( a_{ij} \) is a number \( \bar{a}_{ij} \) defined by the formula

\[
\bar{a}_{ij} \overset{\Delta}{=} (-1)^{(i+j)} \det \bar{A}_{ij}
\]

where \( \bar{A}_{ij} \) is the \( (n-1) \times (n-1) \) matrix obtained by deleting row \( i \) and column \( j \) from \( A \). The \textit{Laplace Expansion} \{along row \( i \)} provides a means for computing \( \det A \):

\[
\det A = \sum_{j=1}^{n} a_{ij} \bar{a}_{ij}
\]

This is useful for hand computation and theoretical development, but not for machine computation, especially when \( n \) is large. The above formula holds for any fixed \( i \).

If \( A = [a]_{1 \times 1} \), then \( \det A \overset{\Delta}{=} a \). From this and the Laplace expansion, one can derive the following general formula for \( 2 \times 2 \) matrices:

\[
\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc
\]

**Example 3** If

\[
A = \begin{bmatrix} 1 & 4 & 9 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}
\]

then

\[
\bar{4} = (-1)^{(1+2)} \det \begin{bmatrix} 2 & 8 \\ 3 & 9 \end{bmatrix} = 6
\]

and \( \det A = -6 \). The Matlab commands for computing \( \bar{a}_{ij} \), \( [\bar{a}_{ij}] \) and \( \det A \) are \texttt{cofactor(A, i, j)}, \texttt{cofactor(A)} and \texttt{det(A)} respectively.

It is worth noting that the determinant of an upper \{or lower\} triangular matrix is the product of its diagonal elements. The reader should try to prove this.

### 2.4 Matrix Rank

Any matrix obtained by deleting any number of rows and columns from a matrix \( M \) is called a \textit{subarray} of \( M \). For example,

\[
\begin{bmatrix}
m_{21} & m_{23} \\
m_{31} & m_{33} \\
m_{51} & m_{53}
\end{bmatrix}
\]
is the subarray of $M = [m_{ij}]_{3\times 5}$ obtained by deleting rows 1 and 4 as well as column 2.

By a minor of a matrix $M$ is meant the determinant of a square subarray of $M$; if the subarray has size $k \times k$, then the minor is of order $k$. The largest among the orders of the non-zero minors of $M$ is the rank of $M$. Clearly

$$\text{rank } M_{p\times q} \leq \min\{p, q\} \quad (2.1)$$

In other words, the rank of a matrix doesn’t exceed its number of rows or number of columns, whichever is smaller. $M$ has full rank if (2.1) holds with equality.

**Example 4** The determinants of the three $2 \times 2$ subarrays

$$
\begin{bmatrix}
1 & 2 \\
2 & 4
\end{bmatrix} \quad 
\begin{bmatrix}
1 & 3 \\
2 & 6
\end{bmatrix} \quad 
\begin{bmatrix}
2 & 3 \\
4 & 6
\end{bmatrix}
$$

of the matrix

$$M = \begin{bmatrix}
1 & 2 & 3 \\
2 & 4 & 6
\end{bmatrix}$$

are all zero. Since $[4]$ is a $1 \times 1$ subarray with non-zero determinant, it must be true that rank $M = 1$. The Matlab command for computing rank $M$ is \{you guessed it!\} rank($M$).

### 2.5 Basic Matrix Operations

Here we review four basic matrix operations, namely transposition, scalar multiplication, addition, and multiplication.

#### 2.5.1 Transposition

The transpose of a matrix

$$M = [m_{ij}]_{p\times q}$$

is denoted by $M'$ and is defined by the formula

$$M' \triangleq [m_{ji}]_{q\times p}$$

$M$ is obtained by simply interchanging the rows and columns of $M$. For example, if

$$M = \begin{bmatrix}
1 & 3 \\
2 & 7 \\
4 & 5
\end{bmatrix} \quad \text{then} \quad M' = \begin{bmatrix}
1 & 2 & 4 \\
3 & 7 & 5
\end{bmatrix}$$

Note that a matrix is symmetric just in case it equals its transpose.

**Facts:** The determinant of a square matrix is invariant under transposition. From this and the definition of rank, it is not hard to see why the rank of a matrix \{square or not\} is invariant under transposition.

The Matlab statement $B = A'$ defines $B$ as the transpose of $A$. 

2.5.2 Scalar Multiplication

The scalar multiplication of a matrix

\[ M = [m_{ij}]_{p \times q} \]

by a scalar \( k \in \mathbb{K} \) is denoted by either \( kM \) or \( Mk \) and is defined by

\[ kM \triangleq [km_{ij}]_{p \times q} \]

For example, with \( M \) as above

\[ kM = \begin{bmatrix} 1k & 3k \\ 2k & 7k \\ 4k & 5k \end{bmatrix} \]

Matlab treats scalars and \( 1 \times 1 \) matrices as one and the same. In Matlab, the product \( kM \) is written as \( k*M \).

2.5.3 Addition

The sum of two matrices is defined only if both matrices are of the same size. If \( A = [a_{ij}]_{m \times n} \) and \( B = [b_{ij}]_{m \times n} \), then their sum is denoted by \( A + B \) and is defined by

\[ A + B \triangleq [a_{ij} + b_{ij}]_{m \times n} \]

For example

\[
\begin{bmatrix} 1 & 1 & 3 \\ 1 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 1 & 7 & 9 \\ 0 & 4 & 4 \end{bmatrix} = \begin{bmatrix} 2 & 9 & 12 \\ 1 & 5 & 5 \end{bmatrix}
\]

The Matlab command for a matrix sum is just what you’d expect, namely \( A + B \).

Observe that matrix addition is

associative: \( A + (B + C) = (A + B) + C \)

commutative: \( A + B = B + A \)

and for each matrix \( A \), there is an ‘additive inverse,’ namely \( -A \), such that \( A + (-A) = 0 \).

The idea of matrix addition extends in a natural way to the sum of two partitioned matrices, provided both matrices are partitioned in the same way. For example, if

\[
A_{p \times q} = \begin{bmatrix} B_{m \times n} & C_{m \times (q-n)} \\ D_{(p-m) \times n} & E_{(p-m) \times (q-n)} \end{bmatrix}
\]

and

\[
\bar{A}_{p \times q} = \begin{bmatrix} \bar{B}_{m \times n} & \bar{C}_{m \times (q-n)} \\ \bar{D}_{(p-m) \times n} & \bar{E}_{(p-m) \times (q-n)} \end{bmatrix}
\]

then

\[
A + \bar{A} = \begin{bmatrix} B + \bar{B} & C + \bar{C} \\ D + \bar{D} & E + \bar{E} \end{bmatrix}
\]
2.5.4 Multiplication

Just like matrix addition, matrix multiplication is defined only for pairs of matrices whose sizes are consistent. In particular, if \( A_{p \times m} \) and \( B_{n \times q} \) are matrices, their product \( AB \) is defined only if the number of columns of \( A \) equals the number of rows of \( B \). That is if \( n = m \). If \( A = [a_{ij}]_{p \times m} \) and \( B = [b_{jk}]_{m \times q} \), then product \( AB \) is a \( p \times q \) matrix \( C = AB \) whose elements are defined by the formulas

\[
c_{ik} = \sum_{j=1}^{m} a_{ij}b_{jk} \quad i \in \{1, 2, \ldots, p\}, \quad k \in \{1, 2, \ldots, q\}
\]  

(2.2)

When thinking about the product of \( A_{p \times n} \) and \( B_{n \times q} \), it is useful to keep in mind 1) that the ‘inner’ integer \( n \) must be the same for multiplication to be defined and 2) that the size of the product \( C_{p \times q} \) is determined by the outer integers \( p \) and \( q \).

The Matlab command for computing the product \( AB \) is simply \( A*B \). Although Matlab uses (2.2) as it stands to carry out this computation, it turns out that this formula is not the most useful way to think about what matrix multiplication means. In the sequel we will elaborate on this point.

Multiplication of a matrix times a vector

Suppose \( A \) is an \( m \times n \) matrix and \( x \) is an \( n \)-vector. We can write \( A \) and \( x \) in partitioned form as

\[
A = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}
\]

where \( a_i \) is the \( i \)th column of \( A \) and \( x_i \) is the \( i \)th element of \( x \). The product of \( A \) times \( x \) turns out to be an \( m \)-vector of the form

\[
Ax = a_1x_1 + a_2x_2 + \cdots + a_nx_n
\]

In other words, the product \( Ax \) can be thought of as the sum of \( n, m \)-vectors, the \( i \)th \( m \)-vector being the product of the \( i \)th element of \( x \) times the \( i \)th column of \( A \). For example, if

\[
A = \begin{bmatrix} 1 & 0 & 1 & 2 \\ 2 & 1 & 1 & 2 \\ 0 & 0 & 1 & 2 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} 1 \\ 2 \\ 4 \\ 3 \end{bmatrix}
\]

then

\[
Ax = 1\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 2\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 4\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + 3\begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 11 \\ 14 \\ 10 \end{bmatrix}
\]

The reader may wish to verify that this is consistent with the definition of matrix multiplication given by formula (2.2).

Note that if \( x \) were the \( i \)th unit \( n \)-vector (i.e., that \( n \)-vector whose \( i \)th row element is 1 and whose remaining \( n-1 \) elements are all 0), then multiplication of \( A \) times \( x \) would have the effect of ‘extracting’ from \( A \) its \( i \)th column. In other words, \( Ax = a_i \).
2.5. BASIC MATRIX OPERATIONS

Multiplication of a matrix times a matrix

Suppose $A$ is an $m \times n$ matrix and that $B$ is a $n \times q$ matrix. Since the inner integer $n$ is the same for both matrices, the product $AB$ is a well-defined $m \times q$ matrix whose $ik$th entry is given by (2.2). An especially useful way to think about what this product is, is as follows:

1. First think of $B$ as an ordered array of $q$ columns, each column being an $n$-vector; i.e.,

   $$B = \begin{bmatrix} b_1 & b_2 & \cdots & b_q \end{bmatrix}$$

   where $b_i$ is the $i$th column of $B$.

2. For each $b_i$ compute the product $Ab_i$ - remember $b_i$ is an $n$-vector and we’ve already explained how to do multiplication in this case.

3. The $i$th column of the product $AB$ is then just $Ab_i$. In other words

   $$AB = \begin{bmatrix} Ab_1 & Ab_2 & \cdots & Ab_q \end{bmatrix} \quad (2.3)$$

For example, if

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 4 & 0 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 0 & 0 \end{bmatrix}$$

then

$$AB = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 4 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 3 \\ 4 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 & 1 & 2 \\ 5 & 2 & 4 \\ 4 & 4 & 8 \end{bmatrix}$$

Multiplication of a matrix times a diagonal matrix

Note that if $B$ were a square diagonal matrix with $i$th diagonal element $b_{ii}$, then in view of (2.3), the $i$th column of the product $AB$ would be the $i$th column of $A$ times $b_{ii}$. Thus if $B$ were the $n \times n$ identity $I_{n \times n}$, then $AI_{n \times n} = A$. Similar reasoning can be used to prove that if $A$ were a square diagonal matrix with $i$th diagonal element $a_{ii}$, then the $i$th row of the product $AB$ would be the $i$th row of $B$ times $a_{ii}$. Also if $A$ were the $m \times m$ identity $I_{m \times m}$, then $I_{m \times m}B = B$. Therefore, for any $m \times n$ matrix $M$,

$$I_{m \times m}M = M = MI_{n \times n}$$

Properties of matrix multiplication

- Multiplication is an associative operation:

  $$A(BC) = (AB)C$$

- Multiplication is not a commutative operation. In other words, except for special cases $AB \neq BA$, even if $A$ and $B$ are both square. For example

  $$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \neq \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

  Two square matrices of the same size are said to commute if $AB = BA$. 

• Multiplication distributes over addition:

\[ A(B + C) = AB + AC \]

• From the equation \( AD = BD \), one cannot in general conclude that \( A = B \). For example, if

\[
A = \begin{bmatrix} 1 & -1 \\ 0 & 6 \end{bmatrix}, \quad B = \begin{bmatrix} -2 & 2 \\ 9 & -3 \end{bmatrix}, \quad \text{and} \quad D = \begin{bmatrix} 1 \\ 1 \end{bmatrix}
\]

then \( AD = BD \) even though \( A \neq B \).

• In general \( (AB)' = B'A' \). That is, the transpose of the product of two matrices equals the reverse product of the transposes of the matrices.

• The following formula is useful in many applications:

\[
\text{rank } (AB) \leq \min \{ \text{rank } A, \text{rank } B \}
\]

In other words, the rank of the product of two matrices \( A \) and \( B \) does not exceed the rank of either \( A \) or \( B \). A proof of this claim can be found in [1] on pages 10-12.

• Another important fact: If \( A \) and \( B \) are square matrices of the same size, then the determinant of the product equals the product of the constituent determinants. In other words,

\[
\det(AB) = \det(A) \det(B)
\]

For a proof of this see [1], pages 10-11.

2.5.5 Linear Algebraic Equations

For sure the most common place one encounters matrices is in dealing with linear algebraic equations. Consider the system of \( n \) simultaneous linear algebraic equations in \( m \) unknowns:

\[
y_i = a_{i1}u_1 + a_{i2}u_2 + \cdots + a_{im}u_m, \quad i \in \{1, 2, \ldots, n\}
\]  \hspace{1cm} (2.4)

By introducing the matrix \( A \triangleq [a_{ij}]_{m \times n} \) and the vectors

\[
y \triangleq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \text{and} \quad u \triangleq \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}
\]

it is possible to rewrite (2.4) compactly as

\[
y = Au
\]  \hspace{1cm} (2.5)

This is why multiplication of a matrix times a vector is defined as it is: the definition is crafted as it is to make the conversion from (2.4) to (2.5) come out right.

Suppose that \( B \) is another \( m \times n \) matrix, and that for each \( n \)-vector \( u \) satisfying (2.5)

\[
z \triangleq Bu
\]  \hspace{1cm} (2.6)

Then the definition of matrix addition implies that

\[
y + z = (A + B)u
\]
Also if \[ w = Dy \] (2.7)
is a matrix equation relating \( y \) to \( w \), then (2.5), (2.7) and the definition of matrix multiplication imply that
\[ w = (DA)u \]

In other words, matrix addition and multiplication are defined so as to be consistent with the manipulations of systems of simultaneous linear algebraic equations.

At this point some readers may be thinking that matrix algebra is little more than a convenient shorthand notation for dealing with simultaneous algebraic equations. This would be something like characterizing natural language as nothing more that a convenient way for humans to communicate. Natural language provides us not only with a convenient way to communicate, but also with a high-level conceptual framework for problem solving and working out new ideas in our own minds. The same is true of matrix algebra.

### 2.5.6 Matrix Inversion

The notions of matrix addition and matrix multiplication are natural generalizations of the notions of addition and multiplication of real numbers. In the sequel we discuss various generalizations of the arithmetic operation of inverting \( \{ \text{forming the reciprocal of} \} \) a nonzero number.

#### Nonsingular Matrices

In extending matrix inversion to square matrices, the natural generalization of a nonzero number is not a nonzero matrix but rather a ‘nonsingular’ one. A square matrix \( A \) is nonsingular if its determinant is nonzero. If \( \det A = 0 \) then \( A \) is singular. Note that the rank of an \( n \times n \) matrix is \( n \) if and only if the matrix is nonsingular.

**Remark 1** The logical statement “\( X \) is true if and only if \( Y \) is true” means that \( X \) and \( Y \) are equivalent assertions. The statement also means that “\( X \) implies \( Y \) and \( Y \) implies \( X \).” The symbol iff abbreviates if and only if. Sometimes we say that “for \( X \) to be true it is necessary and sufficient that \( Y \) be true.” This is another way of saying that “\( X \) is true if and only if \( Y \) is true.”

#### Matrix Inverses

The concept of a matrix inverse is not limited to just square matrices.

**Right and left inverses**

An \( n \times m \) matrix \( A \) is said to have a right inverse \( A_R \) if \( AA_R = I_{n \times n} \) where \( I_{n \times n} \) is the \( n \times n \) identity matrix. Similarly \( A \) has a left inverse \( A_L \) if \( A_L A = I_{m \times m} \) where \( I_{m \times m} \) is the \( m \times m \) identity.

Suppose \( A \) has a right inverse \( A_R \). Then

\[ n = \text{rank } I_{n \times n} = \text{rank } (AA_R) \leq \min \{ \text{rank } A, \text{ rank } A_R \} \leq \text{rank } A \leq \min \{ m, n \} \leq n \]

Since \( n \) appears at both ends of this chain of inequalities, the only way this can be true is if all of the above inequalities hold with equality. Therefore a necessary condition for \( A \) to have a right inverse is that the rank
of $A$ equal the number of rows of $A$. Later it will be shown that this rank condition is also sufficient for $A$ to have a right inverse. Similar reasoning can be used to show that a matrix $B$ has a left inverse if and only if its rank equals the number of its columns.

Note that the transpose of any right inverse of a matrix $M$ is a left inverse of $M'$ and visa versa. Because of this the problems of computing right and left inverses are essentially the same. The left inverse of a matrix $A_{m \times n}$ whose rank equals $m$ can be computed in Matlab using the command `pinv(A)`. Try this on your favorite $3 \times 2$ full rank matrix.

**The inverse of a square matrix**

The preceding imply that a given matrix $A$ will have both a left inverse $A_L$ and a right inverse $A_R$ just in case the number of rows of $A = \text{rank } A = \text{the number of columns of } A$; in other words, just in case $A$ is a square and nonsingular matrix. Suppose

$$AA_R = I_{n \times n}$$
$$A_L A = I_{m \times m}$$

Clearly

$$A_R = I_{m \times m} A_R = (A_L A) A_R = A_L (A A_R) = A_L I_{n \times n} = A_L$$

so $A_L = A_R$. When this is the case we write $A^{-1}$ for $A_R$ and call $A^{-1}$ the \{unique\} inverse of $A$. It follows from the preceding discussion that $A^{-1}$ exists if and only if $A$ is a square nonsingular matrix. Note in addition that $A^{-1} A = AA^{-1} = I_{n \times n}$.

**Inversion formulas**

In general left and right inverses are not unique and because of this there are no specific formulas for calculating them. Nevertheless left and right inverses can be computed when the exist, using Gauss Elimination – an important algorithm which we will discuss a little later. The Matlab commands for the left inverse of a $n \times m$ matrix $A$ of rank $m$ is is `pinv(A)`. If $A = [a_{ij}]_{n \times n}$ is a nonsingular matrix, then there is an explicit formula for its inverse. In particular

$$A^{-1} = \left( \frac{1}{\det A} \right) [\bar{a}_{ij}]'_{n \times n}$$

where $\bar{a}_{ij}$ is the cofactor of $a_{ij}$. \{See pages 13-17 of [1] for a proof of this.\} While this formula is convenient for hand calculation and for theoretical development, it is a very poor way to compute an inverse when $n$ is larger than (say) 3 or 4. The Matlab command for computing the inverse of a square matrix $A$ is either `inv(A)` or $A^\cdot(-1)$.

**Remark 2** There is an important point here. There are mathematical objects \{e.g., the right inverse of a matrix $A$\} which may or may not exist and which may or may not be unique even when they exist. Moreover there may be no explicit formula for calculating such an object even though there may well be a computer algorithm for doing just that. As we shall see there are many mathematical objects including solutions to some types of linear algebraic equations which share these properties.

The following formulas are especially useful:

- If $A$ and $B$ are both $n \times n$ nonsingular matrices then
  $$ (AB)^{-1} = B^{-1} A^{-1} $$
2.6 Linear Recursion Equations

Later in this course we will be studying various properties of ‘ordinary differential equations.’ Ordinary differential equations are of major importance to science and engineering because they provide good models of many different types of natural phenomena. ‘Recursion equations’ are first cousins to ordinary differential equations. In fact the way a computer calculates a solution to a differential equation is by first converting it to a recursion equation and then calculating a solution to the latter. Recursion equations are useful for many other purposes as well. For example, Newton’s method for computing the \( n \)th root of a positive number \( \mu \) using only addition, multiplication and division is described by the recursion equation

\[
y(i + 1) = \left(1 - \frac{1}{n}\right) y(i) + \frac{\mu}{ny(i)^{n-1}}
\]

This is an example of a 1-dimensional, nonlinear recursion equation. Nonlinear, because for each fixed \( i \), the right hand side of the equation is a nonlinear function of \( y(i) \), and 1-dimensional because each \( y(i) \) is a scalar. For a given initial value for \( y(1) \), the equation is ‘solved’ by first setting \( i = 1 \) and evaluating \( y(2) \), then setting \( i = 2 \) and evaluating \( y(3) \) and so on.

Let \( A \) be a given \( n \times n \) matrix. Then

\[
x(i + 1) = Ax(i)
\]

is a linear, \( n \)-dimensional recursion equation. Note that \( x(2) = Ax(1) \), \( x(3) = Ax(2) = AAx(1) \), \( x(4) = Ax(3) = AAAx(1) \) and so on. Let us define

\[
A^0 \triangleq I, \quad A^1 \triangleq A, \quad A^2 \triangleq AA, \quad A^3 \triangleq AAA
\]

and so on. Then it is easy to see that

\[
x(i) = A^{(i-j)} x(j), \quad \forall \ i \geq j \geq 0
\]

Moreover \( A^p A^q = A^{(p+q)} \) for all integers \( p \geq 0 \) and \( q \geq 0 \).

The notation generalizes in a natural way. For example suppose

\[
\alpha(s) \triangleq 5s^3 + 6s^2 + 7s + 1
\]

The notation \( \alpha(A) \) is defined to mean

\[
\alpha(A) \triangleq 5A^3 + 6A^2 + 7A + I_{n \times n}
\]

The \( n \times n \) matrix \( \alpha(A) \) is called a matrix polynomial. This formalism enables us to say (for example) that if \( \alpha(s) \) and \( \beta(s) \) are given polynomials, then

\[
\alpha(A) \beta(A) = \beta(A) \alpha(A)
\]

The matlab command for computing \( A^m \) is \( A^{m} \).

**Example 5** Here is a very simple example of where a linear recursion equation arises. Recall that a fixed-rate mortgage is an agreement to pay back to a lender over a specified period \( \{ \text{usually } N \text{ years} \} \), a certain amount of borrowed money called the principal \( P \) together with interest computed at a fixed annual rate \( R\% \). Payments are
We've used (2.11) here. The indicated computations can easily be carried out using Matlab. Here are plots of \( \pi(i) \) versus payment period for \( \mu = 0.08 \) and \( \mu = 1.0008 \) assuming \( P = 1000 \), \( R = 7\% \) and \( N = 30 \) years.

Let \( y(i) \) denote the remaining unpaid principal after payment \( i - 1 \) has been made. We thus have at once the boundary conditions

\[
\begin{align*}
y(1) &= P \\
y(n+1) &= 0
\end{align*}
\]  

(2.9)  

(2.10)

The \( i \)th monthly payment \( \pi(i), i \in \{1, 2, \ldots, n\} \) to the lender is made up of two components, the \( i \)th interest payment \( r y(i) \), and the agreed upon \( i \)th principal payment \( x(i) \). Thus

\[
\pi(i) = x(i) + r y(i), \quad i \in \{1, 2, \ldots, n\}
\]

(2.11)

Clearly

\[
y(i+1) = y(i) - x(i), \quad i \in \{1, 2, \ldots, n\}
\]

(2.12)

For historical reasons {lenders didn’t have computers when mortgages were invented}, the \( x(i) \) are usually chosen so that all \( n \) monthly payments are the same. In other words so that \( \pi(i+1) = \pi(i), \quad \forall \ i \in \{1, 2, \ldots, n - 1\} \). We’re going to consider a more progressive policy in which each monthly payment is \( \mu \) times the previous monthly payment, \( \mu \) being to agreed upon positive scale factor. That is

\[
\pi(i+1) = \mu \pi(i), \quad i \in \{1, 2, \ldots, n - 1\}
\]

(2.13)

Setting \( \mu = 1 \) gives the standard mortgage policy. Clearly there are constraints on what other values \( \mu \) can take, but let’s not worry about them now. Using (2.11)-(2.13) we get

\[
x(i+1) = r(\mu - 1)y(i) + (\mu + r)x(i), \quad i \in \{1, 2, \ldots, n - 1\}
\]

(2.14)

It is now possible to write (2.12) and (2.14) as a 2-dimensional linear recursion equation of the form

\[
\begin{bmatrix}
y(i+1) \\
x(i+1)
\end{bmatrix}
= \begin{bmatrix}
1 & -1 \\
r(\mu - 1) & \mu + r
\end{bmatrix}
\begin{bmatrix}
y(i) \\
x(i)
\end{bmatrix}, \quad i \in \{1, 2, \ldots, n - 1\}
\]

(2.15)

Recursing this equation \( n - 1 \) times and making use of (2.9) we get

\[
\begin{bmatrix}
y(n) \\
x(n)
\end{bmatrix}
= \begin{bmatrix}
1 & -1 \\
r(\mu - 1) & \mu + r
\end{bmatrix}^{(n-1)}
\begin{bmatrix}
P \\
x(1)
\end{bmatrix}
\]

We also know from (2.12) and (2.10) that

\[
\begin{bmatrix}
y(n) \\
x(n)
\end{bmatrix}
= 0
\]

Combining these two equations we get the linear equation

\[
\begin{bmatrix}
1 & -1 \\
r(\mu - 1) & \mu + r
\end{bmatrix}^{(n-1)}
\begin{bmatrix}
P \\
x(1)
\end{bmatrix}
= 0
\]

which can then be solved for \( x(1) \). Once this is accomplished, the computation of \( y(i) \) and \( x(i) \) can be carried out by recursively solving (2.15). The end result is a formula for \( \pi(i) \) of the form

\[
\pi(i) = \begin{bmatrix}
r & 1 \\
r(\mu - 1) & \mu + r
\end{bmatrix}^{(i-1)}
\begin{bmatrix}
P \\
x(1)
\end{bmatrix}, \quad i \in \{1, 2, \ldots, n\}
\]

(2.11)

We’ve used (2.11) here. The indicated computations can easily be carried out using Matlab. Here are plots of \( \pi(\cdot) \) versus payment period for \( \mu = 1 \) and \( \mu = 1.0008 \) assuming \( P = 1000 \), \( R = 7\% \) and \( N = 30 \) years.
It seems that $\mu \triangleq 1.0008$ has little to offer. How about trying $\mu \triangleq .99$?

## 2.7 Invariants under Premultiplication

To motivate what comes next, let us consider the following problems:

**P1:** Given an $n \times m$ matrix $A$, develop an algorithm for computing the rank of $A$.

**P2:** Given an $n \times m$ matrix $A$ and an $n$-vector $b$ develop an algorithm for

(a) deciding when the equation $Ax = b$ has a solution $x$, and

(b) finding a solution if one exists.

Key to the solutions of both of these problems are the following facts.

**F1:** For any nonsingular $n \times n$ matrix $T$,

$$\text{rank } TA = \text{rank } A$$

**F2:** For any nonsingular $n \times n$ matrix $T$, the set of all solutions to the equation

$$TAx = Tb$$

is the same as the set of all solutions to the equation

$$Ax = b$$

The validity of Fact 1 follows from the chain of inequalities

$$\text{rank } A = \text{rank } (T^{-1}TA) \leq \text{min}\{\text{rank } T^{-1}, \text{rank } (TA)\} \leq \text{rank } (TA) \leq \text{min}\{\text{rank } T, \text{rank } A\} \leq \text{rank } A$$
As for Fact 2, note first that if \( y \) is any solution to (2.17), then \( Ay = b \); premultiplication of both sides of this equation by \( T \) shows that \( y \) must also satisfy (2.16). Conversely, if \( z \) is any solution to (2.16), then \( TAz = Tb \); premultiplication of both sides of this equation by \( T^{-1} \) shows that \( z \) must also satisfy (2.17). Thus Fact 2 is true.

The reason Fact 1 is useful can be explained as follows. As we shall see, for any matrix \( M_{n \times p} \) it is possible to compute a finite sequence of \( n \times n \) nonsingular matrices \( E_1, E_2, \ldots, E_q \) which transform \( M \) into a matrix

\[
M^* \triangleq E_q \cdots E_2 E_1 M
\]

of very special form. So special if fact that the rank of \( M^* \) can be read off by inspection. Notice that \( E_q \cdots E_2 E_1 \) is a nonsingular matrix, since each \( E_i \) is. Thus if we had chosen \( M \) to be \( A \) and defined \( T \) to be \( T \triangleq E_q \cdots E_2 E_1 \), then this process would have transformed \( A \) into a new matrix \( TA \) whose rank could have been read off by inspection. Therefore because of Fact 1, the overall approach would have provided a means for computing the rank of \( A \).

The utility of Fact 2 admits a similar explanation. In this case one would define \( M \triangleq [A \ b] \). Transforming \( M \) into \( M^* \triangleq E_q \cdots E_2 E_1 M \) would in this case yield the partitioned matrix \( M^* = [TA \ Tb] \) where \( T \triangleq E_q \cdots E_2 E_1 \) as before. Because of \( M^* \)'s special form \{which we have yet to describe\}, \( TA \) and \( Tb \) also have special forms which make it possible to decide by inspection if there is a solution to \( TAx = Tb \) and, if there is, to read off such a solution in a direct manner. Because of Fact 2, this overall process provides an especially useful algorithm for solving Problem 2.

### 2.7.1 Elementary Row Operations

There are three very simple kinds of operations which can be performed on the rows of an arbitrary \( n \times m \) matrix \( A \). The operations are as follows.

1. Interchange two rows of \( A \)
2. Multiply a row of \( A \) by a nonzero number.
3. Add to a row of \( A \), a different row of \( A \) multiplied by a number.

Corresponding to each elementary row operation, there is a nonsingular matrix \( E \), which when multiplied times \( A \) on the left, produces in a new matrix \( EA \) which is the same as that matrix which would have resulted had the same elementary row operation been performed directly on \( A \). Moreover \( E \) is that matrix which results when the elementary row operation under consideration is applied to the \( n \times n \) identity matrix \( I_{n \times n} \). To illustrate, consider the \( 3 \times 2 \) matrix

\[
A = \begin{bmatrix}
a_{11} & a_{12} \\
a_{21} & a_{22} \\
a_{31} & a_{32}
\end{bmatrix}
\]

1. The elementary row operation of interchanging rows 1 and 3 yields

\[
B = \begin{bmatrix}
a_{31} & a_{32} \\
a_{21} & a_{22} \\
a_{11} & a_{12}
\end{bmatrix}
\]
The same operation on $I_{3\times 3}$ produces the elementary matrix

$$E_{(r_1 \leftrightarrow r_3)} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Moreover $E_{(r_1 \leftrightarrow r_3)}A = B$. Note that $E_{(r_1 \leftrightarrow r_3)}$ is nonsingular.

2. The elementary row operation of multiplying row 2 of $A$ by some number $\mu \neq 0$ yields

$$C = \begin{bmatrix} a_{11} & a_{12} \\ \mu a_{21} & \mu a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

The same operation on $I_{3\times 3}$ produces the elementary matrix

$$E_{(r_2 \rightarrow \mu r_2)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mu & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

and $E_{(r_2 \rightarrow \mu r_2)}A = C$. Note that $E_{(r_2 \rightarrow \mu r_2)}$ is nonsingular because $\mu \neq 0$.

3. The elementary row operation of adding to row 2 of $A$, row 3 multiplied by some number $\mu$, yields the matrix

$$D = \begin{bmatrix} a_{11} & a_{12} \\ \mu a_{31} + a_{21} & \mu a_{32} + a_{22} \\ a_{31} & a_{32} \end{bmatrix}$$

The same operation on $I_{3\times 3}$ produces the elementary matrix

$$E_{(r_2 \rightarrow (r_2 + \mu r_3))} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \mu \\ 0 & 0 & 1 \end{bmatrix}$$

and $E_{(r_2 \rightarrow (r_2 + \mu r_3))}A = D$. Note again that $E_{(r_2 \rightarrow (r_2 + \mu r_3))}$ is nonsingular.

### 2.7.2 Echelon Form

By performing a sequence of elementary row operations on a given $n \times n$ matrix $A$, it is possible to transform $A$ into a new, specially structured matrix $A^*$, called an *echelon form* of $A$. The structure of $A^*$ is as follows.

1. Each of the first $k \geq 0$ rows of $A^*$ are nonzero and each of the remaining $n - k$ rows has only zeros in it.

2. If $k > 0$, the first nonzero element in each row $i \in \{1, 2, \ldots, k\}$ is a 1.

3. If $j_i$ is the index of the column within which the $ith$ such 1 appears, then

$$j_1 < j_2 < \cdots j_k$$

In other words, the nonzero rows of an echelon matrix all begin with 1’s and to the left and below each such 1 are only 0’s. For example

$$A^* = \begin{bmatrix} 0 & 1 & a_{13} & a_{14} & a_{15} & a_{16} \\ 0 & 0 & 0 & 1 & a_{25} & a_{26} \\ 0 & 0 & 0 & 0 & 1 & a_{36} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$
is an echelon matrix with \( k \geq 3 \) nonzero rows.

Sometimes it is useful to perform additional elementary row operations on a matrix to bring it to a reduced echelon form. A reduced echelon form differs from an echelon form in that all of the entries above the 1’s are also zero. For example, the echelon form \( A^* \) just described can be transformed to a reduced echelon form which looks like

\[
B^* = \begin{bmatrix}
0 & 1 & 0 & 0 & a_{13} & 0 & a_{16} \\
0 & 0 & 0 & 1 & a_{26} & 0 & 0 \\
0 & 0 & 0 & 0 & a_{36} & 0 & 0
\end{bmatrix}
\]

One nice feature of an echelon form (or a reduced echelon form) is that its rank can be determined by inspection. For example, consider the echelon form \( A^* \) above. Because the last row of this matrix is zero, any \( 4 \times 4 \) subarray will have only zeros in its bottom row; thus all minors of order 4 are zero so the rank of \( A^* \) must be less than 4. On the other hand, the \( 3 \times 3 \) subarray

\[
\begin{bmatrix}
1 & a_{14} & a_{15} \\
0 & 1 & a_{25} \\
0 & 0 & 1
\end{bmatrix}
\]

has determinant 1, so \( A^* \) has a third order nonzero minor. Therefore the rank of \( A^* \) is three, which just so happens to be the number of nonzero rows of \( A^* \). This is true in general for both echelon and reduced echelon forms. The rank of any echelon or reduced echelon form \( A^* \) is equal to the number of nonzero rows of \( A^* \).

### 2.7.3 Gauss Elimination

The algorithm which transforms a given matrix \( A \) into echelon form by applying a sequence of elementary row operations, is called Gauss Elimination. Gauss Elimination is in essence nothing more than the formalization of the method commonly used to solve simultaneous linear algebraic equations by hand. It is without question the most basic of all algorithms in real numerical analysis.

The easiest way to understand Gauss Elimination is by watching someone work through an example. The algorithm can be described in words as follows.

1. Start with a given \{nonzero\} \( n \times m \) matrix \( A = [a_{ij}] \).
2. Let \( j \) denote the index of the first nonzero column of \( A \).
3. Let \( i \) be an integer for which \( a_{ij} \neq 0 \).
4. **Elementary Row Interchange:** Interchange rows 1 and \( i \) obtaining a matrix \( B \) whose 1\( j \) entry \( b_{1j} = a_{ij} \neq 0 \).
5. **Pivot on Element 1\( j \):**
   
   (a) **Elementary Row Multiplication:** Multiply row 1 of \( B \) by \( \frac{1}{b_{1j}} \) thereby obtaining a matrix \( C \) whose first \( j - 1 \) columns are zero and \( c_{1j} = 1 \).
   
   (b) **Elementary Row Additions:** For the first nonzero entry \( c_{kj} \) below \( c_{1j} \) {i.e., \( k > 1 \)}, add \(-c_{kj}\) times row 1 of \( C \) to row \( k \). Repeat this process successively on each resulting matrix thereby obtaining in at most \( n - 1 \) steps, a matrix \( D \) whose first \( j - 1 \) columns are zero and whose \( j \)th column has a 1 in row 1 and zeros below the 1.
6. **Elementary Row Interchange:** Let \( k \) be the first column of \( D \) beyond the \( j \)th which has a nonzero entry not in its first row. If \( k \) does not exist, stop because \( D \) is in echelon form. Let \( h \) be any integer greater than 1 for which \( d_{hk} \neq 0 \). Interchange rows 2 and \( h \) obtaining a matrix \( F \) whose 2\( k \) entry \( f_{2k} = d_{hk} \neq 0 \).

7. **Pivot on Element 2\( k \):**

   (a) **Elementary Row Multiplication:** Multiply row 2 of \( F \) by \( \frac{1}{f_{2k}} \) thereby obtaining a matrix \( G \) whose first \( j - 1 \) columns are zero, whose \( j \)th column has a 1 in the first row and zeros below, and whose second row has as its first nonzero element a 1 in column \( k \).

   (b) **Elementary Row Additions:** Successively eliminate the nonzero entries below \( g_{2k} \), as in step 5b above\(^1\).

8. Continue these steps until all columns have been processed. The result is an echelon matrix.

Here is an example.

\[
A \triangleq \begin{bmatrix}
0 & 0 & 0 & 3 & 1 \\
0 & 2 & 4 & 2 & 2 \\
0 & 1 & 2 & 1 & 0 \\
0 & 3 & 0 & 1 & 0 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
0 & 1 & 2 & 1 & 1 \\
0 & 0 & 0 & 3 & 1 \\
0 & 1 & 2 & 1 & 0 \\
0 & 3 & 0 & 1 & 0 \\
\end{bmatrix}
= A^* =
\begin{bmatrix}
0 & 1 & 2 & 1 & 1 \\
0 & 0 & 0 & 3 & 1 \\
0 & 1 & 2 & 1 & 0 \\
0 & 3 & 0 & 1 & 0 \\
\end{bmatrix}
\]

As noted in footnote 1, a reduced echelon form \( B^* \) of \( A \) can be obtained by performing additional elementary row additions as part of each pivot. Alternatively \( B^* \) can be obtained by first computing \( A^* \) and then performing additional elementary row additions \( A^* \) to eliminate the nonzero entries in above the 1’s. The following illustrates the latter approach.

\[
A^* = \begin{bmatrix}
0 & 1 & 2 & 1 & 1 \\
0 & 0 & 1 & \frac{1}{3} & \frac{1}{2} \\
0 & 0 & 0 & 1 & \frac{1}{3} \\
0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
0 & 1 & 0 & -\frac{1}{3} & 0 \\
0 & 0 & 1 & \frac{1}{3} & \frac{1}{2} \\
0 & 0 & 0 & 1 & \frac{1}{3} \\
0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\rightarrow
\begin{bmatrix}
0 & 1 & 0 & 0 & \frac{1}{3} \\
0 & 0 & 1 & \frac{1}{3} & \frac{1}{2} \\
0 & 0 & 0 & 1 & \frac{1}{3} \\
0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\]

\(^1\)If a reduced echelon form is sought, at this point one would also eliminate the nonzero entries above the 1 in position 2\( k \) by additional elementary row additions.
CHAPTER 2. MATRIX ALGEBRA

The entries in a reduced echelon form are not always just 1's and 0's even though this turned out to be the case here. The Matlab command for computing a reduced echelon form of a given matrix $A$ is `rref(A)`. Matlab uses essentially the method just described with the elimination of nonzero elements above 1's being carried out as part of each pivot as noted mentioned in footnote 1. A nice Matlab demo illustrating this can be seen by typing the command `rrefmovie`.

Compute of Matrix Rank

Since an echelon form or reduced echelon form $A^*$ of $A$ is obtained by applying a sequence of elementary row operations to $A$, it follows that there is a corresponding sequence of elementary matrices $E_1, E_2, \ldots, E_i$, one for each elementary row operation, such that

$$A^* = E_i E_{i-1} \cdots E_2 E_1 A$$

Since each $E_j$ is nonsingular, the product $T \triangleq E_i E_{i-1} \cdots E_2 E_1$ must be nonsingular as well. Therefore

$$A^* = TA$$

This proves that any matrix can be transformed into an echelon form or a reduced echelon form by multiplying it on the left by a suitably defined nonsingular matrix.

As we've already noted, the rank of a matrix doesn't change when multiplied on the left by a nonsingular matrix. Therefore it must be true that rank $A = \text{rank } A^*$. But we've already noted that the rank of an echelon form or a reduced echelon form is equal to its number of nonzero rows. We are therefore led to the following conclusion. The rank of any matrix $A$ is the same as the same as the number of nonzero rows in any echelon form or reduced echelon form of $A$. Together with Gauss Elimination, this provides a numerical method for obtaining the rank of a matrix. Note that this method has the desirable feature of not requiring the computation of determinants.

Matrix Inversion

Suppose that $A_{n \times n}$ is a square, nonsingular matrix. Then $A$ must have rank $n$. This means that any reduced echelon form $B^*$ of $A$ must have $n$ nonzero rows. This implies that $B^*$ must be the $n \times n$ identity $I_{n \times n}$, since reduced echelon forms are required to have 0s above and below each 1. This means that $A$ can be transformed into the $n \times n$ identity matrix by a applying to it, a sequence of elementary row operations. Thus if $E_1, E_2, \ldots, E_q$ is the set of elementary matrices obtained by applying the same sequence of elementary row operations to identity matrices, then

$$E_q E_{q-1} \cdots E_2 E_1 A = I$$
Clearly
\[ A^{-1} = E_q E_{q-1} \cdots E_2 E_1 \]
so
\[ [I \ A^{-1}] = E_q E_{q-1} \cdots E_2 E_1 [A \ I] \]

In other words, it is possible to compute the inverse of \( A \) by applying a sequence of elementary operations to the partitioned matrix \([A \ I]\) which transform it into its reduced echelon form. This form automatically turns out to be \([I \ A^{-1}]\). More symbolically
\[
[A \ I] \xrightarrow{op_q, \ op_{q-1}, \ldots, \ op_2, \ op_1} [I \ A^{-1}]
\]

where \( op_i \) denotes the \( i \)th elementary row operation applied to \([A \ I]\) in the process of transforming it to reduced echelon form. Computing \( A^{-1} \) in this way is called the Gauss Jordan method. It is preferred method for computing the inverse of an arbitrary nonsingular matrix. To illustrate, suppose
\[
A = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}
\]

Then the Gauss Jordan computation would proceed as follows.

\[
\begin{bmatrix} 1 & 3 & 1 & 0 \\ 2 & 4 & 0 & 1 \end{bmatrix} \xrightarrow{r_2 \rightarrow (r_2 - 2r_1)} \begin{bmatrix} 1 & 3 & 1 & 0 \\ 0 & -2 & -2 & 1 \end{bmatrix} \xrightarrow{r_2 \rightarrow -\frac{1}{2}r_2} \begin{bmatrix} 1 & 3 & 1 & 0 \\ 0 & 1 & 1 - \frac{1}{2} \end{bmatrix}
\]

\[
\xrightarrow{r_1 \rightarrow (r_1 - 3r_2)} \begin{bmatrix} 1 & 0 & -2 & \frac{3}{2} \\ 0 & 1 & 1 - \frac{1}{2} \end{bmatrix}
\]

Thus
\[
A^{-1} = \begin{bmatrix} -2 & \frac{3}{2} \\ 1 & -\frac{1}{2} \end{bmatrix}
\]

There are two Matlab commands for computing the inverse of a nonsingular matrix \( A \), namely \( A^{-1} \) and \( \text{inv}(A) \).

### Solutions to Linear Algebraic Equations

As we explained at the beginning of Section 2.7, the set of solutions \( x \) to a linear algebraic equation of the form
\[
A_{n \times m} x_{m \times 1} = b_{n \times 1} \tag{2.18}
\]
is the same as the set of solutions to
\[
T A x = T b
\]
provided \( T \) is a \( n \times n \) nonsingular matrix. Suppose \( op_1, \ op_2, \ldots, \ op_q \) is a sequence of elementary row operations transforming the partitioned matrix \([A \ b]\) into a reduced echelon form \( M^* \) written in partitioned form \([\bar{A} \ \bar{b}]\). In other words
\[
[A \ b] \xrightarrow{op_q, \ op_{q-1}, \ldots, \ op_2, \ op_1} [\bar{A} \ \bar{b}]
\]

Then the set of solutions to (2.18) must be the same as the set of solutions to
\[
\bar{A} x = \bar{b} \tag{2.19}
\]
because \( \bar{A} = TA \) and \( \bar{b} = Tb \) where \( T \triangleq E_q E_{q-1} \cdots E_2 E_1 \) and \( E_i \) is the elementary row matrix obtained by applying operation \( op_i \) to the \( n \times n \) identity matrix. The key point here is that because \([\bar{A} \ \bar{b}]\) is a reduced
echelon form, the problem of finding solutions to (2.19) \{ and consequently (2.18) \} is especially easy. The following example serves to illustrate this point.

Suppose that $A$ is $4 \times 5$, that $b$ is $4 \times 1$ and consequently that $x$ is a $5$ - vector of unknowns; i.e.

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix}$$

Suppose that $C^*$ is a reduced echelon form of $[A \ b]$ which looks like

$$C^* = \begin{bmatrix} 0 & 1 & c_{13} & 0 & 0 & c_{16} \\ 0 & 0 & 0 & 1 & 0 & c_{26} \\ 0 & 0 & 0 & 0 & 1 & c_{36} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{2.20}$$

Then the set of solutions to (2.18) is the same as the set of solutions to

$$\begin{bmatrix} 0 & 1 & c_{13} & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} c_{16} \\ c_{26} \\ c_{36} \\ 0 \end{bmatrix}$$

By carrying out the indicated multiplications, one obtains the simultaneous equations

$$
x_2 + c_{13}x_3 = c_{16} \\
x_4 = c_{26} \\
x_5 = c_{36}
$$

From these equations it is clear that the set of all solutions to (2.18) consists of all vectors of the form

$$x = \begin{bmatrix} \mu_1 \\ c_{16} - c_{13}\mu_3 \\ \mu_3 \\ c_{26} \\ c_{36} \end{bmatrix}$$

where $\mu_1$ and $\mu_3$ are real numbers. Because $\mu_1$ and $\mu_3$ are arbitrary elements of $\mathbb{R}$, the system of equations under consideration does not have a unique solution; in fact there are infinitely many solutions, since there are infinitely many possible values for $\mu_1$ and $\mu_3$.

To illustrate the possibility of nonexistence, suppose that in place of

$$\begin{bmatrix} c_{16} \\ c_{26} \\ c_{36} \\ 0 \end{bmatrix}$$

the last column of the reduced echelon form of $[A \ b]$ in (2.20) had turned out to be

$$\begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$
Then the resulting matrix equation
\[
\begin{bmatrix}
0 & 1 & c_{13} & 0 & 0 \\
0 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
x_5
\end{bmatrix}
= 
\begin{bmatrix}
0 \\
0 \\
0 \\
1
\end{bmatrix}
\]
would not have had a solution at all because any such solution would have had to satisfy
\[
\begin{bmatrix}
0 & 0 & 0 & 0 & 0
\end{bmatrix}
\begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4 \\
x_5
\end{bmatrix}
= 1
\]
which is impossible.

To summarize, existence, uniqueness, and explicit solutions to \(Ax = b\) can be obtained directly by inspection of a reduced echelon form of \([A \ b]\). Any such form can be obtained in a straightforward manner via elementary row operations.

**Solutions to Linear Matrix Equations**

Now suppose that \(A_{n \times m}\) and \(B_{n \times q}\) are given matrices and consider the matrix equation
\[
AX = B \tag{2.21}
\]
Let \(b_i\) and \(x_i\) denote the \(i\)th columns of \(B\) and \(X\) respectively. Thus each \(b_i\) is an \(n\) - vector, each \(x_i\) is an \(m\) - vector, and
\[
B = [b_1 \ b_2 \ \cdots \ b_q] \quad \text{and} \quad X = [x_1 \ x_2 \ \cdots \ x_q]
\]
The point here is that (2.21) is equivalent to the \(q\) ‘decoupled’ equations
\[
Ax_i = b_i, \quad i \in \{1, 2, \ldots, q\} \tag{2.22}
\]
Because they are decoupled, each such equation can thus be solved independently of the others using the method discussed in the last section. Defining these \(q\) vector solutions as the columns of a matrix \(\bar{X}\) thus provides a solution \(X = \bar{X}\) to the original matrix equation (2.21). Note that the \(q\) independent equations (2.22) can also be solved simultaneously by computing a reduced echelon form of \([A \ B]\). The Matlab command for computing a matrix \(X\) which solves \(AX = B\) is \(X = A\backslash B\).

**2.7.4 Elementary Column Operations**

Just as there are three elementary row operations which can be performed on the rows of an arbitrary \(n \times m\) matrix \(A\) there are three analogous operations which can be performed on the columns of \(A\). The operations are

1. Interchange two columns of \(A\).
2. Multiply a column of \(A\) by a nonzero number.
3. Add to a column of \(A\), a different column of \(A\) multiplied by a number.
Corresponding to each elementary column operation, there is a nonsingular matrix $E$, which when multiplied times $A$ on the right, produces in a new matrix $AE$ which is the same as that matrix which would have resulted had the elementary column operation under consideration been performed directly on $A$. Moreover $E$ is that matrix which results when the elementary column operation under consideration is applied to the $m \times m$ identity matrix $I_{m \times m}$.

We shall not illustrate all this. Suffice it to say that an elementary column operation on $A$ {e.g., interchange columns $i$ and $j$ of $A$} amounts to performing the corresponding elementary row operation on $A'$ {e.g., interchange rows $i$ and $j$ of $A'$} and then transposing the resulting matrix.

2.7.5 Equivalent Matrices

Let $A_{n \times m}$ be a given matrix. As we’ve already seen, by application of a sequence of elementary row operations it is possible to transform $A$ into a reduced echelon form $A^*$. Such an $A^*$ might look like

$$A^* = \begin{bmatrix} 0 & 1 & a_{13} & 0 & 0 & a_{16} \\ 0 & 0 & 0 & 1 & 0 & a_{26} \\ 0 & 0 & 0 & 1 & 0 & a_{36} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We’ve also seen that there must be a nonsingular matrix $T$ such that

$$A^* = TA \quad (2.23)$$

Now suppose that by means of elementary row operations, the transpose of $A^*$ e.g.

$$(A^*)' = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ a_{13} & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ a_{16} & a_{26} & a_{36} & 0 \end{bmatrix}$$

is transformed into a reduced echelon form $B^*$. For the example at hand, $B^*$ would have to be

$$B^* = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (2.24)$$

The very special structure of $B^*$ is no accident. Rather it is a direct consequence of the fact that $B^*$ is a reduced echelon form of the transpose of a reduced echelon form. Note that $(B^*)'$ could also have been obtained by applying elementary column operations to $A^*$.

We know that there must be a nonsingular matrix $R'$ such that

$$B^* = R'(A^*)'$$

Combining this with (2.24) and (2.23) and we see that

$$TAR = \begin{bmatrix} I_{3 \times 3} & 0_{3 \times 3} \\ 0_{1 \times 3} & 0_{1 \times 3} \end{bmatrix}$$
The matrix on the right is an example of a “left-right equivalence canonical form.”

For given nonnegative integers \( n, m \) and \( r \) with \( r \leq \min\{n, m\} \), a left-right equivalence canonical form \( M_{n \times m} = [m_{ij}] \) is a \( n \times m \) matrix whose entries are all zero except for the \( r \) entries \( m_{11}, m_{22}, \ldots, m_{rr} \) which are all 1’s. Note that \( r \) is the rank of \( M \). What we’ve just shown by example is that for any matrix \( A \) there are nonsingular matrices \( T \) and \( R \) which transform \( A \) to a left-right canonical form \( TAR \).

**Equivalence Relations**

There are aspects of the preceding which can be usefully interpreted in a broader context. This requires the concept of an equivalence relation. An equivalence is a special type of relation.

Let \( S \) be a given set. Roughly speaking, a “relation” between objects in \( S \) is an association between certain pairs of objects in \( S \) determined by a specific set of rules. For example, if \( S \) is the set of graduate students taking this course this fall, a set of rules defining a relation called “sort of similar” might be as follows: student \( s_a \) is sort of similar to student \( s_b \) if student \( s_a \) is registered in the same department as student \( s_b \) and has been at Yale at least as long as student \( s_b \). Notice that the relation “sort of similar” enables one to form ordered pairs of sort of similar students \((s_a, s_b)\). The set of all such ordered pairs is sometimes considered to be the definition of the relation.

There is a special type of relation, called an ‘equivalence relation’ which is especially useful. We’ll first define it formally, and then give some examples.

Let \( S \) be a set. Let \( S \times S \) denote the set of all ordered pairs \((s_1, s_2)\) where \( s_1 \) and \( s_2 \) are both elements of \( S \). \( S \times S \) is sometimes called the **Cartesian Product** of \( S \) with itself.\] Let \( E \) be a given subset of \( S \times S \). Let’s use the notation \( s_1E s_2 \) to mean \((s_1, s_2)\) is an element of \( E \). The subset \( E \) is called an equivalence relation on \( S \) if the following three requirements are satisfied.

1. **Reflexivity:** \( sEs, \forall s \in S \). In other words each \( s \) in \( S \) must be equivalent to itself.

2. **Symmetry:** \( tEs, \forall s, t \in S \) such that \( sEt \). In other words for each \( s \) and \( t \) in \( S \) such that \( s \) is equivalent to \( t \), \( t \) must be equivalent to \( s \).

3. **Transitivity:** \( sEu, \forall s, t, u \in S \) such that \( sEt \) and \( t Eu \). In other words for each \( s \), \( t \), and \( u \) in \( S \) such that \( s \) is equivalent to \( t \) and \( t \) is equivalent to \( u \), it must follow that \( s \) is equivalent to \( u \).

Equality of real numbers is a simple example of an equivalence relation on \( \mathbb{R} \). On the other hand, ‘less than or equal’ \( \{\text{i.e., } \leq\} \) is not an equivalence relation on \( \mathbb{R} \) because requirement 2, namely symmetry, is not satisfied. The relation “sort of similar” defined above is also not an equivalence relation because it fails to be symmetric.

An equivalence relation \( E \) on a set \( S \) provides a simple means of partitioning \( S \) into disjoint subsets of like \( \{\text{i.e., equivalent}\} \) objects. These subsets are called ‘equivalence classes’ and are defined just as you’d expect. The equivalence class of \( s \in S \) under \( E \) is the subset \([s]_E \triangleq \{t : tEs, t \in S\} \). In other words, \([s]_E \) is simply the set of all elements in \( S \) which are equivalent to \( s \). The set of all possible equivalence classes of \( S \) under \( E \) is written as \( S/E \) and is called the quotient set induced by \( E \). Thus \( S/E = \{[s]_E : s \in S\} \). The rules defining \( E \) insure that distinct equivalence classes don’t overlap:

\([s]_E \cap [t]_E = \text{the empty set if } s \not\sim t\)

The rules also insure that the union of all such equivalence classes is \( S \):

\[ S = \bigcup_{s \in S} [s]_E \]
Example 6 Suppose $S \triangleq \{1, 2, 3, 4, 5\}$ and suppose that $E$ is defined so that $sEt$ whenever $s - t$ is an integer multiple of 3. Then $E = \{(1, 4), (4, 1), (2, 5), (5, 2), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5)\}$. Since the above requirements are satisfied, $E$ is an equivalence relation on $S$. In addition $[1]_E = \{1, 4\}$, $[2]_E = \{2, 5\}$, $[3]_E = \{3\}$, $[4]_E = [1]_E$, $[5]_E = [2]_E$ and $S/E = \{[1]_E, [2]_E, [3]_E\}$.

**Left-Right Equivalence**

Later in the course we will study two special equivalence relations on sets of matrices, namely “similarity” and “congruence.” In the sequel we will consider a third which serves to illustrate the ideas we’ve just discussed.

Let $\mathcal{M}^{n\times m}$ denote the set of all $n \times m$ matrices over the field $\mathbb{K}$. Let us agree to call two matrices $M_1$ and $M_2$ in $\mathcal{M}^{n\times m}$ left-right equivalent \(^2\) if there exist nonsingular matrices $P$ and $Q$ such that

$$PM_1Q = M_2$$

We sometimes write $M_1 \mapsto QM_1P$ to emphasize that such a $Q$ and $P$ are transforming $M_1$ into $M_2$. We often call this a *left-right equivalence transformation*. The reader is encouraged to verify that left-right equivalence is an equivalence relation on $\mathcal{M}^{n\times m}$.

Earlier in this section we demonstrated by example that for any matrix $A$ in $\mathcal{M}^{n\times m}$ there are nonsingular matrices $T$ and $R$ which transform $A$ into a left-right equivalence canonical form $E^* \triangleq TAR$. Therefore $A$ and $E^*$ are left-right equivalent. We also pointed out that the structure of $E^*$ is uniquely determined by three integers $n, m,$ and $r$, the latter being the rank of $E^*$. Since matrix rank is invariant under premultiplication and postmultiplication by nonsingular matrices \(\text{i.e., under left-right equivalence transformations}\), $r$ must also be the rank of $A$. In other words, $E^*$ itself is uniquely determined by properties of $A$ \(\text{namely A’s size and rank}\), which in turn are invariant under left-right equivalence transformations.

In the light of the preceding, it is possible to determine what’s required of two matrices $A$ and $B$ in $\mathcal{M}^{n\times m}$ in order for them to be left-right equivalent. On the one hand, if they are in fact left-right equivalent, then they must both have the same rank because rank is invariant under left-right equivalence transformations. On the other hand, if $A$ and $B$ both have the same rank, they turn out to be left-right equivalent. Here’s a proof:

- Suppose $A$ and $B$ have the same rank.
- Then $A$ and $B$ must both have the same left-right equivalence canonical form $E^*$.
- By definition, $A$ and $B$ are thus both left-right equivalent to $E^*$.
- By symmetry, $E^*$ is left-right equivalent to $B$
- Since $A$ is left-right equivalent to $E^*$ and $E^*$ is left-right equivalent to $B$ it must be that $A$ is left-right equivalent to $B$ because of transitivity.

Therefore **two matrices of the same size are left-right equivalent if and only if they both have the same rank.**

Let us note that the concept of left-right equivalence enables us to decompose $\mathcal{M}^{n\times m}$ into $1 + \min\{n, m\}$ equivalence classes $C_i$, $i \in \{0, 1, \ldots, \min\{n, m\}\}$ where $M \in C_i$ just in case rank $M = i$. The rank of a matrix $M$ serves to uniquely label $M$’s left-right equivalence class in the quotient set $\mathcal{M}^{n\times m}/\text{left-right equivalence}$. Said differently, the rank of a matrix uniquely determines the matrix “up to a left-right equivalence transformation.”

\(^2\)In many texts, ‘left-right’ equivalent matrices are called simply ‘equivalent’ matrices. Later on, when there is less chance of confusion, we’ll adopt the shorter name.
Chapter 3

Linear Algebra

It is sometimes useful to think of a matrix as a “representation” of a “linear function” which “assigns” to each vector in a “vector space,” a vector in another vector space. By doing this it is possible to single out those properties of a matrix {e.g., rank} which don’t depend on the “coordinate system” in which the linear transformation is represented. In the sequel we expand on these ideas by first introducing the concept of a vector space and discussing some of its properties. Then we make precise the concept of a linear function and explain how it relates to a matrix.

3.1 Linear Vector Spaces

Let $\mathbb{K}$ denote an algebraic field {e.g., $\mathbb{R}$ or $\mathbb{C}$} and let $Y$ be a set. Let $\circ$ and $+$ be symbols such that $k \circ y$ and $y_1 + y_2$ denote well-defined elements of $Y$ for each $k \in \mathbb{K}$ and each $y, y_1, y_2 \in Y$. The algebraic system \( \{Y, \mathbb{K}, +, \circ\} \) is called a $\mathbb{K}$-vector space with vectors $y \in Y$ and scalars $k \in \mathbb{K}$ provided the following axioms hold:

1. $y_1 + (y_2 + y_3) = (y_1 + y_2) + y_3$, \( \forall y_1, y_2, y_3 \in Y \).
2. $y_1 + y_2 = y_2 + y_1$, \( \forall y_1, y_2 \in Y \).
3. There exists a vector $\bar{0} \in Y$ \{the additive unit\} call the zero vector such that $\bar{0} + y = y$, \( \forall y \in Y \).
4. For each $y \in Y$ there is a vector $-y \in Y$ \{the additive inverse\} such that $y + (-y) = \bar{0}$.
5. $k \circ (y_1 + y_2) = k \circ y_1 + k \circ y_2$, \( \forall k \in \mathbb{K}, y_1, y_2 \in Y \).
6. $(k_1 + k_2) \circ y = k_1 \circ y + k_2 \circ y$, \( \forall k_1, k_2 \in \mathbb{K}, y \in Y \).
7. $(k_1 k_2) \circ y = k_1 \circ (k_2 \circ y)$, \( \forall k_1, k_2 \in \mathbb{K}, y \in Y \).
8. $1 \circ y = y$, \( \forall y \in Y \).

To summarize, a $\mathbb{K}$-vector space is a field of scalars $\mathbb{K}$ and a set of vectors $Y$ together with two ‘operations’, namely scalar multiplication \{\( \circ \)\} and vector addition \{\( + \)\} which satisfy the familiar rules of vector algebra just listed. To simplify notation we shall henceforth write $ky$ for $k \circ y$, $y_1 + y_2$ for $y_1 + y_2$, and $0$ for $\bar{0}$. These notational simplifications are commonly used and should cause no confusion.
Throughout these notes capital calligraphic letters $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$... usually denote $\mathbb{K}$-vector spaces. Vectors are typically denoted by lower case Roman letters $x$, $y$, $z$,... Vector spaces containing just the zero vector are denoted by $0$. Note that there is no such thing as an empty vector space because every vector space must contain at least one vector, namely the zero vector!

The set of all $n \times 1$ real-valued matrices together with ordinary matrix addition and $\{\text{real}\}$ scalar multiplication provides a concrete example of an $\mathbb{R}$-vector space; this space is denoted by $\mathbb{R}^n$. Similarly, $\mathbb{C}^n$ is the $\mathbb{C}$-vector space consisting of all $n \times 1$ complex-valued matrices with addition and scalar multiplication defined in the obvious way. While these two vector spaces arise quite often in physically-motivated problems, they are by no means the only vector spaces of practical importance.

**Example 7** The set of all infinite sequences $\{k_1$, $k_2$, $\ldots\}$ of numbers $k_i \in \mathbb{K}$ together with component-wise addition defined by
\[ \{a_1$, $a_2$, $\ldots\} + \{b_1$, $b_2$, $\ldots\} = \{a_1 + b_1$, $a_2 + b_2$, $\ldots\} \]
and component-wise scalar multiplication defined by
\[ k\{k_1$, $k_2$, $\ldots\} = \{kk_1$, $kk_2$, $\ldots\} \]
is a $\mathbb{K}$-vector space. In this case the “vectors” are infinite sequences.

**Example 8** The set of all real-valued functions $f(t)$ defined on a closed interval of the real line $[a$, $b] = \{t : a \leq t \leq b\}$, together with ordinary pointwise addition and multiplication by real numbers, is a real vector space. In this case the “vectors” are functions defined on $[a$, $b]$.

**Example 9** The set of all polynomials in one variable $s$, with complex coefficients, and degree less than or equal to $n$, together with ordinary polynomial addition and “scalar” multiplication by complex numbers is a $\mathbb{C}$-vector space. In this case the “vectors” are polynomials.

### 3.1.1 Subspaces

Recall that a *subset* $V$ of a set $X$ is itself a set whose elements are also elements of $X$. Let $\mathcal{X}$ be a given vector space over $\mathbb{K}$. A nonempty subset $\mathcal{V}$ whose elements are vectors in $\mathcal{X}$, is a *subspace* of $\mathcal{X}$ if $\mathcal{V}$ is a vector space under the operations of $\mathcal{X}$. In other words for $\mathcal{V}$ to be a subspace of $\mathcal{X}$, each element of $\mathcal{V}$ must be a vector in $\mathcal{X}$ and $\mathcal{V}$ must be *closed* under the addition and scalar multiplication operations of $\mathcal{X}$:
\[ k_1v_1 + k_2v_2 \in \mathcal{V}, \quad \forall k_i \in \mathbb{K}, \quad v_2 \in \mathcal{V} \]

Let us note that the zero subspace and the whole space $\mathcal{X}$ are subspaces of $\mathcal{X}$. Any other subspace of $\mathcal{X}$ \{i.e., any subspace other than the whole space $\mathcal{X}$ or the zero subspace $0$\} is said to be a *proper* subspace of $\mathcal{X}$. To indicate that $\mathcal{V}$ is a subspace of $\mathcal{X}$ one often uses the notation $\mathcal{V} \subset \mathcal{X}$ and says that $\mathcal{V}$ is *contained* in $\mathcal{X}$.

**Example 10** The set of all vectors of the form
\[
\begin{bmatrix}
  r_1 \\
  r_2 \\
  0
\end{bmatrix}, \quad r_1$, $r_2 \in \mathbb{R}
\]
is a subspace of $\mathbb{R}^3$. 
Example 11 The set of all vectors of the form

\[ \begin{bmatrix} r_1 \\ r_2 \\ 2 \end{bmatrix}, \quad r_1, r_2 \in \mathbb{R} \]

is not a subspace of \( \mathbb{R}^3 \) because this set is not closed under addition; i.e., although

\[ \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix} \]

are vectors in the set, their sum

\[ \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix} \]

is not since \( 4 \neq 2 \).

### 3.1.2 Subspace Operations

Let \( \mathcal{U} \) and \( \mathcal{V} \) be subspaces of a given vector space \( \mathcal{X} \). The sum of \( \mathcal{U} \) and \( \mathcal{V} \), written \( \mathcal{U} + \mathcal{V} \), is defined by

\[ \mathcal{U} + \mathcal{V} \coloneqq \{ u + v : u \in \mathcal{U}, \ v \in \mathcal{V} \} \]

In other words, the sum of \( \mathcal{U} \) and \( \mathcal{V} \) is the set of all possible vectors of the form \( u + v \) where \( u \) and \( v \) are vectors in \( \mathcal{U} \) and \( \mathcal{V} \) respectively. For example, if \( \mathcal{U} \) is the subspace of all vectors of the form

\[ \begin{bmatrix} u_1 \\ 0 \\ u_3 \\ 0 \end{bmatrix}, \quad u_1, u_3 \in \mathbb{R} \]

and \( \mathcal{V} \) is the subspace of all vectors of the form

\[ \begin{bmatrix} v_1 \\ 0 \\ 0 \\ v_4 \end{bmatrix}, \quad v_1, v_4 \in \mathbb{R} \]

then \( \mathcal{U} + \mathcal{V} \) is the set of all vectors of the form

\[ \begin{bmatrix} r_1 \\ 0 \\ r_3 \\ r_4 \end{bmatrix}, \quad r_1, r_3, r_4 \in \mathbb{R} \]

Note that “sum” and “union” are not the same thing. The union of the elements of \( \mathcal{U} \) and \( \mathcal{V} \) is the subset

\[ \mathcal{U} \cup \mathcal{V} \coloneqq \{ w : w \in \mathcal{U} \text{ or } w \in \mathcal{V} \} \]

Clearly

\[ (\mathcal{U} \cup \mathcal{V}) \subset (\mathcal{U} + \mathcal{V}) \]

but the reverse inclusion may not hold.
Both \( \mathcal{U} + \mathcal{V} \) and \( \mathcal{U} \cap \mathcal{V} \) are subsets of \( \mathcal{X} \). However only \( \mathcal{U} + \mathcal{V} \) and not \( \mathcal{U} \cup \mathcal{V} \) is a subspace of \( \mathcal{X} \). To prove that \( \mathcal{U} + \mathcal{V} \) is a subspace it is enough to show that \( \mathcal{U} + \mathcal{V} \) is closed under addition and scalar multiplication. To prove closure under addition, suppose that \( x_1 \) and \( x_2 \) are vectors in \( \mathcal{U} + \mathcal{V} \). Then, because of the definition of \( \mathcal{U} + \mathcal{V} \) there must be vectors \( u_1, v_1 \in \mathcal{U} \) and \( u_2, v_2 \in \mathcal{V} \) such that \( x_1 = u_1 + v_1 \) and \( x_2 = u_2 + v_2 \). Clearly \( x_1 + x_2 = (u_1 + u_2) + (v_1 + v_2) \). Since \( (u_1 + u_2) \in \mathcal{U} \) and \( (v_1 + v_2) \in \mathcal{V} \), it must be that \( (x_1 + x_2) \in (\mathcal{U} + \mathcal{V}) \).

Thus \( \mathcal{U} + \mathcal{V} \) is closed under addition. To establish closure under scalar multiplication, let us note that for any \( k \in \mathbb{K} \), \( k x_1 = ku_1 + kv_1 \); but \( ku_1 \in \mathcal{U} \) and \( kv_1 \in \mathcal{V} \) so \( k x_1 \in (\mathcal{U} + \mathcal{V}) \). Hence \( \mathcal{U} + \mathcal{V} \) is closed under scalar multiplication. Therefore \( \mathcal{U} + \mathcal{V} \) is a subspace. \( \blacksquare \)

Let \( \mathcal{X} \), \( \mathcal{U} \) and \( \mathcal{V} \) be as above. The \textit{intersection} of \( \mathcal{U} \) and \( \mathcal{V} \), written \( \mathcal{U} \cap \mathcal{V} \), is the set of all vectors which are in both \( \mathcal{U} \) and \( \mathcal{V} \). In other words

\[
\mathcal{U} \cap \mathcal{V} = \{ x : x \in \mathcal{U} \text{ and } x \in \mathcal{V} \}
\]

It turns out that \( \mathcal{U} \cap \mathcal{V} \), like \( \mathcal{U} + \mathcal{V} \), is also a subspace of \( \mathcal{X} \). The reader should try to prove that this is so.

To illustrate intersection, suppose that \( \mathcal{U} \) and \( \mathcal{V} \) are subspaces of \( \mathbb{R}^4 \) consisting of all vectors of the forms

\[
\begin{bmatrix}
  u_1 \\
  0 \\
  u_3 \\
  0
\end{bmatrix}
\]

and

\[
\begin{bmatrix}
  v_1 \\
  0 \\
  0 \\
  v_4
\end{bmatrix}
\]

respectively. Then \( \mathcal{U} \cap \mathcal{V} \) is all vectors of the form

\[
\begin{bmatrix}
  r_1 \\
  0 \\
  0 \\
  0
\end{bmatrix}
\]

The definitions of sum and intersection of subspaces extend to finite families of subspaces in a natural way. For example, if \( \mathcal{S}_1 \), \( \mathcal{S}_2 \) and \( \mathcal{S}_3 \) are subspaces of \( \mathcal{X} \) then \( \mathcal{S}_1 + \mathcal{S}_2 + \mathcal{S}_3 \triangleq (\mathcal{S}_1 + \mathcal{S}_2) + \mathcal{S}_3 \) and \( \mathcal{S}_1 \cap \mathcal{S}_2 \cap \mathcal{S}_3 \triangleq (\mathcal{S}_1 \cap \mathcal{S}_2) \cap \mathcal{S}_3 \).

### 3.1.3 Distributative Rule

Suppose that \( \mathcal{U} \), \( \mathcal{V} \) and \( \mathcal{W} \) are subspaces of a vector space \( \mathcal{X} \). We claim that

\[
(\mathcal{W} \cap \mathcal{U}) + (\mathcal{W} \cap \mathcal{V}) \subset \mathcal{W} \cap (\mathcal{U} + \mathcal{V})
\]  

(3.1)

To prove that this is so, let \( x \) be any vector in \( (\mathcal{W} \cap \mathcal{U}) + (\mathcal{W} \cap \mathcal{V}) \). Then \( x \in \mathcal{W} \) and \( x = u + v \) for some \( u \in \mathcal{U} \) and some \( v \in \mathcal{V} \). Since \( u + v \in (\mathcal{U} + \mathcal{V}) \) it must be true that \( x \in (\mathcal{U} + \mathcal{V}) \). Since \( x \) is also in \( \mathcal{W} \), it follows that \( x \in \mathcal{W} \cap (\mathcal{U} + \mathcal{V}) \). As \( x \) was selected arbitrarily in \( (\mathcal{W} \cap \mathcal{U}) + (\mathcal{W} \cap \mathcal{V}) \), the containment \( x \in \mathcal{W} \cap (\mathcal{U} + \mathcal{V}) \) must hold for all \( x \in (\mathcal{W} \cap \mathcal{U}) + (\mathcal{W} \cap \mathcal{V}) \). In other words, (3.1) is true. \( \blacksquare \)

**Example 12** Suppose \( \mathcal{U} \), \( \mathcal{V} \) and \( \mathcal{W} \) are subspaces of \( \mathbb{R}^2 \) consisting of all vectors of the forms

\[
\begin{bmatrix}
  r_1 \\
  0
\end{bmatrix}, \quad \begin{bmatrix}
  0 \\
  r_2
\end{bmatrix}, \quad \begin{bmatrix}
  r_3 \\
  0
\end{bmatrix}
\]

respectively. Then

\[
\mathcal{U} \cap \mathcal{W} = 0, \quad \mathcal{V} \cap \mathcal{W} = 0, \quad \text{and} \quad \mathcal{U} + \mathcal{V} = \mathbb{R}^2
\]

Thus

\[
(\mathcal{U} \cap \mathcal{W}) + (\mathcal{V} \cap \mathcal{W}) = 0 \quad \text{and} \quad \mathcal{W} \cap (\mathcal{U} + \mathcal{V}) = \mathcal{W}
\]
Clearly
\[ W \cap (U + V) \neq (U \cap W) + (V \cap W) \]

Geometrically, things look like this.

Here all points in the plane of the paper represent points in \( \mathbb{R}^2 \) and the point of intersection of the three lines is the zero vector. \( V \) consists of all points on the extension of the vertical line shown, from \(-\infty\) to \(\infty\). \( W \) and \( U \) admit similar interpretations.

It can be shown by example that the reverse inclusion in (3.1) does not necessarily hold so the containment symbol in (3.1) cannot in general be replaced with an equal sign. An exception to this occurs if it happens to be true that
\[ U \subset W \] (3.2)

Our aim is to prove that this is so; i.e., that (3.2) implies that
\[ (W \cap U) + (W \cap V) = W \cap (U + V) \] (3.3)

The implication
\[ U \subset W \Rightarrow (W \cap U) + (W \cap V) = W \cap (U + V) \]
is called the modular distributive rule for subspaces. In view of (3.1), to prove the rule's validity all we need to show is that (3.2) implies
\[ (W \cap U) + (W \cap V) \supset W \cap (U + V) \] (3.4)

This can be done as follows.

Let \( x \) be any vector in \( W \cap (U + V) \). Then \( x \in W \) and \( x = u + v \) for some vectors \( u \in U \) and \( v \in V \). Since \( u \in U \), we have by (3.2) that \( u \in W \) so \( u \in U \cap W \). Since \( v = x - u \) and both \( x \) and \( u \) are in \( W \), it must be true that \( v \in W \); but \( v \in V \) so \( v \in V \cap W \). Therefore \( u + v \in (W \cap U) + (W \cap V) \) which implies that \( x \in (W \cap U) + (W \cap V) \). Since \( x \) can be any vector in \( W \cap (U + V) \), all such vectors must be in \( (W \cap U) + (W \cap V) \). In other words, (3.4) must hold. ■
Two subspaces \{U, V\} of a vector space \mathcal{X} are independent if their intersection is the zero subspace in \mathcal{X}; i.e., \( U \cap V = 0 \). Note that every subspace of \mathcal{X} contains the zero vector, so the intersection of subspaces is never the empty set. The subspaces \( U \) and \( W \) in Example 12 are independent as are the subspaces \( V \) and \( W \).

A finite family of subspaces \( \{S_i : i \in \{1, 2, \ldots, n\}\} \) is independent if
\[
S_i \cap \left( \sum_{j \neq i} S_j \right) = 0, \quad i \in \{1, 2, \ldots, n\}
\]
If \( \{S_i : i \in \{1, 2, \ldots, n\}\} \) is an independent family and \( S \) is the sum \( S = S_1 + S_2 + \cdots + S_n \), then \( S \) is called the direct sum of the \( S_i \). The symbol \( \oplus \) is usually used to denote direct sum. Thus if \( \{S_i : i \in \{1, 2, \ldots, n\}\} \) is an independent family, then
\[
S = S_1 \oplus S_2 \oplus \cdots \oplus S_n
\]
One consequence of \( S \) being a direct sum is that each vector \( s \in S \) can be written as \( s = s_1 + s_2 + \cdots + s_n \) where each vector \( s_i \in S_i \) is unique.

### 3.1.5 Linear Combination and Span

Let \( \{x_i : i \in \{1, 2, \ldots, n\}\} \) be a set of vectors in a vector space \( \mathcal{X} \). A vector \( x \in \mathcal{X} \) is said to be a linear combination of the vectors \( x_1, x_2, \ldots, x_n \) if there exist scalars \( k_i \in \mathbb{K} \) such that \( x = k_1x_1 + k_2x_2 + \cdots + k_nx_n \). By the span of the set of vectors \( \{x_i : i \in \{1, 2, \ldots, n\}\} \) is meant the set of all possible linear combinations of \( x_1, x_2, \ldots, x_n \). It is easy to verify that the span of \( x_1, x_2, \ldots, x_n \) is a subspace of \( \mathcal{X} \). We say that \( x_1, x_2, \ldots, x_n \) spans a given subspace \( \mathcal{U} \subset \mathcal{X} \) just in case \( \mathcal{U} \) is the span of \( x_1, x_2, \ldots, x_n \). \( \mathcal{X} \) is said to be a finite dimensional vector space if it can be spanned by a finite set of vectors. If no such set exists, \( \mathcal{X} \) is called an infinite dimensional space. It will soon be clear, if it is not already, that subspaces of finite dimensional spaces are finite dimensional spaces. Unless otherwise stated, we shall deal only with finite dimensional subspaces in these notes.

**Example 13** Since \( \mathbb{R}^n \) is spanned by the unit n-vectors\(^1\) \( e_1, e_2, \ldots, e_n \), \( \mathbb{R}^n \) is a finite dimensional vector space. The same vectors \( e_1, e_2, \ldots, e_n \) also span \( \mathbb{C}^n \) so \( \mathbb{C}^n \) is also finite dimensional. The vector space of polynomials in one variable \( s \), with coefficients in \( \mathbb{C} \) and degrees not exceeding \( n \) \{see Example 9\} is finite dimensional. The vector space of all infinite sequences \( \{k_1, k_2, \ldots\} \) defined in Example 8 is infinite dimensional as is the vector space of all real-valued functions on \([a, b]\) defined in Example 7.

### 3.1.6 Linear Independence

A set of vectors \( \{x_i : i \in \{1, 2, \ldots, n\}\} \) in \( \mathcal{X} \) is linearly dependent if there exist scalars \( k_i \in \mathbb{K} \), not all zero, such that \( k_1x_1 + k_2x_2 + \cdots + k_nx_n = 0 \). If the set is not linearly dependent it is linearly independent. For example, the set of vectors
\[
x_1 = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \quad x_2 = \begin{bmatrix} 3 \\ -2 \\ 0 \end{bmatrix}, \quad x_3 = \begin{bmatrix} -1 \\ 4 \\ 4 \end{bmatrix}
\]
is linearly dependent since \( 2x_1 - x_2 - x_3 = 0 \). On the other hand, the vectors \( x_1 \) and \( x_2 \) are linearly independent since the equation \( k_1x_1 + k_2x_2 = 0 \) implies that \( k_1 \) and \( k_2 \) must both equal zero.

\(^1\)\( e_i \) is a \( n \times 1 \) matrix with a 1 in row \( i \) and 0’s elsewhere.
### 3.1.7 Basis

A linearly independent set \{x_i : i \in \{1, 2, \ldots, n\}\}, ordered by \(i\), is a basis for \(X\) if the set spans \(X\). The set of unit vectors \(\{e_1, e_2, \ldots, e_n\}\) is a basis for both \(\mathbb{R}^n\) and \(\mathbb{C}^n\). The ordered set consisting of the vectors \(x_1\) and \(x_2\) just defined, is a basis for their span which, in turn, is a subspace of \(\mathbb{R}^3\).

Suppose that \(\{z_1, z_2, \ldots, z_m\}\) is a finite set of vectors which spans \(X\). If this set is independent, it is a basis for \(X\). If not, it is possible to reduce \(\{z_1, z_2, \ldots, z_m\}\) to a smaller set \(\{x_1, x_2, \ldots, x_n\}\) which is a basis. Here’s how:

Suppose that \(\{z_1, z_2, \ldots, z_m\}\) is a finite set of linearly dependent vectors which spans \(X\). Then there must exist scalars \(k_i\), not all zero, such that

\[
k_1 z_1 + k_2 z_2 + \cdots + k_m z_m = 0
\]

Without loss of generality, assume \(k_m \neq 0\). Then \(z_m\) can be written as a linear combination of the remaining vectors \(z_1, z_2, \ldots, z_{m-1}\). That is

\[
z_m = \frac{-k_1}{k_m} z_1 + \frac{-k_2}{k_m} z_2 + \cdots + \frac{-k_{m-1}}{k_m} z_{m-1}
\]

From this it should be clear that the reduced set of vectors \(\{z_1, z_2, \ldots, z_{m-1}\}\) also spans \(X\). If this set is linearly independent, then it is a basis. If not, the process can be continued for at most \(m - 1\) additional steps until a linearly independent subset \(\{x_i : i \in \{1, 2, \ldots, n\}\} \subset \{z_1, z_2, \ldots, z_m\}\) is finally obtained. The result is a basis for \(X\). We have therefore shown that every finite dimensional vector space has a basis. Note: the zero subspace has no basis, but it doesn’t need one since it contains exactly one vector, namely 0.

#### Representations

Let \(\{y_i : i \in \{1, 2, \ldots, n\}\}\) be a basis for a vector space \(Y\) and let \(y\) be any given vector in \(Y\). Because \(\{y_i : i \in \{1, 2, \ldots, n\}\}\) spans \(Y\) there must be scalars \(k_i \in \mathbb{K}\) such that

\[
y = k_1 y_1 + k_2 y_2 + \cdots + k_n y_n\tag{3.5}
\]

Moreover the \(k_i\) must be unique because \(\{y_i : i \in \{1, 2, \ldots, n\}\}\) is a linearly independent set. In other words, for each basis \(\{y_i : i \in \{1, 2, \ldots, n\}\}\) and each vector \(y\) there is a unique, ordered set of numbers \(\{k_i : i \in \{1, 2, \ldots, n\}\}\) for which (3.5) holds. The \(k_i\) are called the coordinates of \(y\) in the basis \(\{y_i : i \in \{1, 2, \ldots, n\}\}\). The column vector

\[
\begin{bmatrix}
k_1 \\
k_2 \\
\vdots \\
k_n
\end{bmatrix}
\]

is called a representation of \(y\) in the basis \(\{y_i : i \in \{1, 2, \ldots, n\}\}\). For example, the vector space of polynomials of degree not exceeding \(n\) defined in Example 9, has as a basis the set of polynomials \(\{1, s, s^2, \ldots, s^n\}\). In this basis, any polynomial \(c_p s^{(p-1)} + c_{p-1} s^{(p-1)} + \cdots + c_2 s + c_1\) in the space can be represented by a \((n + 1) \times 1\) matrix of the form

\[
\begin{bmatrix}
c_1 \\
c_2 \\
\vdots \\
c_p \\
0 \\
\vdots \\
0
\end{bmatrix}
\]
Representations provide an important link between abstract vectors (which might be polynomials, sequences, etc.) and matrices which makes it possible to determine various properties of the original vectors by studying analogous properties of their matrix representations. In the sequel we shall illustrate this by developing a procedure for determining when a given set of vectors is linearly independent by evaluating the rank of a matrix whose columns are representations of the vectors under consideration.

**Linear Independence and Rank**

Let \( \mathcal{X} \) be any finite dimensional vector space and let \( \{ x_j : j \in \{1, 2, \ldots, n\} \} \) be a basis. Let \( \{ y_i : i \in \{1, 2, \ldots, m\} \} \) be any set of vectors in \( \mathcal{X} \). Let

\[
a_i \triangleq \begin{bmatrix} a_{1i} \\ a_{2i} \\ \vdots \\ a_{ni} \end{bmatrix}
\]

be the representation of \( y_i \) in the basis \( \{ x_j : j \in \{1, 2, \ldots, n\} \} \). That is

\[
y_i = \sum_{j=1}^{n} a_{ji} x_j, \quad i \in \{1, 2, \ldots, m\} \quad (3.6)
\]

Next let \( \{ b_i : i \in \{1, 2, \ldots, m\} \} \) be any set of numbers such that

\[
\sum_{i=1}^{m} b_i y_i = 0 \quad (3.7)
\]

Substituting (3.6) into (3.7) we obtain the equation

\[
\sum_{i=1}^{m} b_i \sum_{j=1}^{n} a_{ji} x_j = 0 \quad (3.8)
\]

Since the \( x_j \) are independent, their coefficients in (3.8) must each be zero. In other words,

\[
\sum_{i=1}^{m} a_{ji} b_i = 0, \quad j \in \{1, 2, \ldots, n\} \quad (3.9)
\]

If we define

\[
A \triangleq [a_{ji}]_{n \times m} \quad \text{and} \quad b \triangleq [b_i]_{m \times 1}
\]

then (3.9) can be written as

\[
Ab = 0 \quad (3.10)
\]

What we’ve just proved is that (3.7) implies (3.10). The reader should verify that the reverse implication is also true. In other words (3.7) and (3.10) are equivalent assertions. From this and the definition of linear independence it follows that \( \{ y_i : i \in \{1, 2, \ldots, m\} \} \) will be a linearly independent set just in case the only vector \( b \) for which (3.10) holds is the zero vector. We now develop a condition for testing when the latter is so.

Let \( P \) and \( Q \) be nonsingular matrices which transform \( A \) into its left-right equivalence canonical form \( B^* \); i.e. \( B^* = PAQ \). Note that if \( b \) is a solution to (3.10), then \( Q^{-1}b \) is a solution \( d \) to \( B^*d = 0 \). Conversely, if \( d \) is a solution to \( B^*d = 0 \), then \( Q^{-1}d \) is a solution to (3.10). From this and the fact that \( Q \) is nonsingular it follows that (3.10) will have \( b = 0 \) as its only solution just in case \( B^*d = 0 \) has \( d = 0 \) as its only solution. But because of \( B^* \)'s special structure, the equation \( B^*d = 0 \) will have \( d = 0 \) as its only solution if and only if the
number of 1’s in $B^*$ [i.e., the rank of $B^*$] equals $m$, the number of columns of $B^*$. But rank $B^* = \text{rank } A$. From this it follows that (3.10) will have $b = 0$ as its only solution just in case rank $A = m$. Thus \{y_i : i \in \{1, 2, \ldots, m\}\} will be a linearly independent set if and only if rank $A = m$. We’ve proved the following.

**Proposition 4** Let $\mathcal{X}$ be a finite dimensional space with basis \{x_j : j \in \{1, 2, \ldots, n\}\}, and let \{y_i : i \in \{1, 2, \ldots, m\}\} be any set of vectors in $\mathcal{X}$. Then \{y_i : i \in \{1, 2, \ldots, m\}\} is a linearly independent set if and only if

$$\text{rank } \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix}_{n \times m} = m$$

where for $i \in \{1, 2, \ldots, m\}$, $a_i$ is the representation of $y_i$ in the basis \{x_j : j \in \{1, 2, \ldots, n\}\}.

Note that (3.10) can also be written as

$$a_1 b_1 + a_2 b_2 + \cdots + a_m b_m = 0$$

Clearly the columns of $A$ will be linearly independent just in case $b = 0$; i.e., just in case rank $A = m$. In the sequel, this line of reasoning is carried further.

Suppose that $G_{n \times m}$ is an arbitrary matrix with columns $g_1$, $g_2$, ..., $g_m$. Let $G$ denote the span of \{g_i : i \in \{1, 2, \ldots, m\}\}. As before, we can remove dependent columns from \{g_i : i \in \{1, 2, \ldots, m\}\} without changing span, until we end up with a linearly independent subset which is a basis for $G$. Without loss of generality, assume that this subset consists of the first $q$ columns of $G$. Therefore $g_{q+1}$ can be written as

$$g_{q+1} = k_1 g_1 + k_2 g_2 + \cdots + k_q g_q$$

It follows that if $-k_1$ times $g_1$ plus $-k_2$ times $g_2$ plus ... plus $-k_q$ times $g_q$ is added to column $q + 1$ of $G$, the result will be a zero column. By continuing this process, all columns of $G$, from $q + 1$ to $m$, can be made equal to zero. The end result is a matrix $\tilde{G}_{n \times m}$ of the form

$$\tilde{G} = \begin{bmatrix} \hat{G}_{n \times q} & 0_{n \times (m-q)} \end{bmatrix}$$

where

$$\hat{G} = \begin{bmatrix} g_1 & g_2 & \cdots & g_q \end{bmatrix}$$

Moreover since the steps involved in going from $G$ to $\tilde{G}$ are elementary column operations, there must be a nonsingular matrix $T$ such that

$$\tilde{G} = GT$$  \hspace{1cm} (3.11)

Now since the columns of $\hat{G}$ are independent, the number of columns of $\hat{G}$, namely $q$, must equal the rank of $\hat{G}$. Thus $\hat{G}$ and therefore $\tilde{G}$ must contain a nonzero minor of order $q$. It follows that rank $\tilde{G} \geq q$. But any square subarray of $\tilde{G}$ larger than $q \times q$ must contain a zero row or column and hence must be singular. This means that $q$ is the largest integer for which there exists a nonzero minor of order $q$ in $\tilde{G}$. Therefore rank $\tilde{G} = q$. But from (3.11) and the fact that $T$ is nonsingular, it follows that rank $G = q$. We’ve proved the following very useful fact.

**Proposition 5** The rank of a matrix is equal to the number of linearly independent columns of the matrix.

Since the rank of a matrix equals the rank of its transpose, we also have the following:

**Proposition 6** The rank of a matrix is equal to the number of linearly independent rows of the matrix.
Note that the two propositions imply that the number of linearly independent columns of a matrix equals the number of linearly independent rows.

It is sometimes possible to tell by inspection just how many linearly independent columns or rows a matrix has. For example, the matrix
\[
\begin{bmatrix}
2 & 1 & 3 \\
1 & 1 & 2 \\
3 & 1 & 4 \\
0 & 1 & 1
\end{bmatrix}
\]
has two independent columns since the first two columns are clearly independent and the third is the sum of the first and second. Therefore the rank of the matrix is 2.

### 3.1.8 Dimension

The *dimension* of a vector space $X$, written $\dim(X)$, is the least number of vectors required to span $X$. The dimension of any zero space is defined to be zero.

Suppose $\{x_j : j \in \{1, 2, \ldots, n\}\}$ and $\{y_i : i \in \{1, 2, \ldots, m\}\}$ are two bases for $X$. We can express the vectors in one basis as linear combinations of vectors in the other basis. For example

\[
y_i = \sum_{j=1}^{n} a_{ji} x_j \quad i \in \{1, 2, \ldots, m\}
\]

Since $\{y_i : i \in \{1, 2, \ldots, m\}\}$ is an independent set, it follows from Proposition 4 that rank $[a_{ji}]_{n \times m} = m$. Since the rank of a matrix cannot be larger than the number of its rows, it must be true that $m \leq n$. Now the reverse inequality, namely $n \leq m$, can also be established by interchanging the roles of the $y_i$ and $x_j$ and using similar reasoning. This means that $n = m$. We’ve thus proved the following.

**Proposition 7** All bases for a vector space $X$ contain the same number of vectors.

Observe that the dimension of $X$ cannot exceed the number of vectors in a basis for $X$ because bases are spanning sets. Conversely, the number of vectors in a basis cannot exceed the dimension of $X$; for if this were false, there would exist a spanning set for $X$ containing less vectors than are in a basis. Since the spanning set could be reduced to a basis by eliminating dependent vectors, this would imply that $X$ has two bases containing different numbers of vectors – a clear contradiction of Proposition 7. We have thus proved the following.

**Proposition 8** Each basis for a vector space $X$ contains exactly $\dim(X)$ vectors.

Note that any set of $\dim(X)$ vectors which spans $X$ must be a basis for $X$.

### 3.2 Functions

Having completed our discussion of the basic properties of finite dimensional linear vector spaces, we turn to the concept of a linear transformation. What is a linear transformation? A linear transformation is a linear function. What is a function? To pin down with reasonable precision just what a function is we need first two sets $R$ and $S$. Then a *function* from *domain* $R$ to *codomain* $S$, written $f : R \rightarrow S$, is a rule which assigns to each element $r$ in its domain $R$ a corresponding element $s$ in its codomain $S$; $s$ is called the *value* of $f$.
functions at \( r \) and is often denoted by \( f(r) \). We sometimes write \( r \mapsto f(r) \) to indicate the action of \( f \) on a typical element \( r \) in its domain. As we shall soon see, it is essential in talking about a function that we specify both its domain and its codomain.

**Example 14** The rule which assigns to each real number its cube, determines a function from the real numbers to the real numbers.

**Example 15** The best selling fiction book list published weekly by the New York Times can be viewed as a function from the set of integers \( \{1, 2, \ldots, 10\} \) to the set of all fiction books.

**Example 16** An algorithm which computes the length of a shortest path between two vertices on a weighted, directed graph can be thought of as a function from the set of all such graphs to the reals.

**Example 17** A balance scale determines a function which assigns to each object of mass, a real number, namely the mass of the object.

It is important to recognize that two functions are equal just in case they both have the same domain, the same codomain and the same value at each point in their common domain. Note that the function which assigns to each positive real number, its cube, is not quite the same as the function discussed in Example 14, since the two functions have different domains.

The composition of two functions \( g : S \to T \) and \( f : R \to S \), written \( gf \), is the function \( gf : R \to T \) defined by \( r \mapsto g(f(r)) \). Thus the composition of \( g \) with \( f \) is a “function of a function.” Note that the definition of \( gf \) makes sense only if domain \( g = \text{codomain } f \).

### 3.2.1 Linear Transformations

In order to define a linear transformation we first need two linear vector spaces, \( \mathcal{X} \) and \( \mathcal{Y} \), defined over the same field, \( \mathbb{K} \). A function \( f : \mathcal{X} \to \mathcal{Y} \) is then said to be a linear transformation if the superposition rule

\[
f(k_1x_1 + k_2x_2) = k_1f(x_1) + k_2f(x_2)
\]

holds for all \( x_1, x_2 \in \mathcal{X} \) and all \( k_1, k_2 \in \mathbb{K} \). The reason for requiring \( \mathcal{X} \) and \( \mathcal{Y} \) to be linear vector spaces is clear: in order to define \( f \) as a linear transformation, we need well-defined notions of addition and scalar multiplications in both \( f \)'s domain and codomain. Throughout these notes, linear transformations are usually denoted by capital Roman letters \( A, B, C, \ldots \). When we wish to distinguish sharply between a linear transformation \( A \) and a matrix \( A \), we shall usually express the latter in boldface; e.g., \( \mathbf{A} \). We often write \( Ax \) rather than \( A(x) \) to denote the value of \( A \) at \( x \).

**Example 18** Let \( \mathcal{P}^{n+1} \) denote the \((n+1)\)-dimensional linear vector space of all real-coefficient polynomials, of the form \( \alpha(t) = a_{n+1}t^{n+1} + a_n t^{n-1} + \cdots + a_2 t + a_1 \) together standard polynomial addition and multiplication by scalars in \( \mathbb{R} \). The function which assigns to each \( \alpha(t) \in \mathcal{P}^{n+1} \), its derivative \( \frac{d}{dt} \alpha(t) \), determines a linear transformation from \( \mathcal{P}^{n+1} \) to \( \mathcal{P}^n \).

**Example 19** Let \( \mathcal{Y} \) denote the vector space of all directed line segments in the real plane, drawn from a fixed origin. The function which assigns to each line segment, the same line segment rotated 60 degrees clockwise, is a linear transformation from \( \mathcal{Y} \) to \( \mathcal{Y} \).
Example 20 Let \( Z \) denote the set of all directed line segments in the real plane, drawn from all points in the plane. The function which assigns to each line segment in the vector space \( Y \) defined in Example 19, the same line segment translated two units to the right, is a function from \( Y \) to \( Z \). However this function is not quite a linear transformation because superposition does not hold.

Example 21 Let \( A_{n \times m} \) be a real matrix. The function which assigns to each vector \( x \) in \( \mathbb{R}^m \), the vector \( Ax \) in \( \mathbb{R}^n \) is a linear transformation from \( \mathbb{R}^m \) to \( \mathbb{R}^n \).

3.2.2 Operations With Linear Transformations

There are basically two operations one can perform with linear transformations, namely composition and addition. The composition of the linear transformations \( M : \mathcal{X} \rightarrow \mathcal{Y} \) with the linear transformation \( N : \mathcal{Y} \rightarrow \mathcal{W} \) is the function \( NM : \mathcal{X} \rightarrow \mathcal{W} \) defined so that \( NM(x) = N(Mx) \) \( \forall x \in \mathcal{X} \). Thus the composition of functions is defined in the same manner, whether the functions are linear or not. It turns out however that the composition of two linear functions is a linear function. The reader should try to verify this.

The sum of two linear transformations \( M : \mathcal{X} \rightarrow \mathcal{Y} \) and \( L : \mathcal{X} \rightarrow \mathcal{Y} \), written \( M + L \), is the function \( (M + L) : \mathcal{X} \rightarrow \mathcal{Y} \) defined so that \( (M + L)(x) = M(x) + L(x) \), \( \forall x \in \mathcal{X} \). The reader may wish to verify that the sum of two linear transformations is a linear transformation.

3.2.3 Representations of Linear Transformations

As we’ve already noted, is is sometimes useful to represent abstract vectors using column matrices. Doing this makes it possible to study properties of the original abstract vectors \{e.g., linear independence\} in terms
of their representations. The same is true of linear transformations. The idea is that a linear transformation can be "represented" or "characterized" by a matrix, once bases for the domain and codomain of the transformation have been settled upon. Matrix representations of linear transformations are defined so that the linear transformation operations of addition and composition translate into the matrix operations of addition and multiplication respectively. We shall expand on these ideas in the sequel.

Let \( L : \mathcal{X} \to \mathcal{Y} \) be a given linear transformation and let \( \{ x_j : j \in \{1, 2, \ldots, n\}\} \) and \( \{ y_i : i \in \{1, 2, \ldots, m\}\} \) be bases for \( \mathcal{X} \) and \( \mathcal{Y} \) respectively. Since \( L(x_j) \) is a vector in \( \mathcal{Y} \) and since \( \{ y_i : i \in \{1, 2, \ldots, m\}\} \) is a basis for \( \mathcal{Y} \), there are unique numbers \( k_{ij} \) such that

\[
L(x_j) = \sum_{i=1}^{m} k_{ij} y_i, \quad j \in \{1, 2, \ldots, n\}
\]  

(3.12)

The array

\[
L = [k_{ij}]_{m \times n} = \begin{bmatrix}
k_{11} & k_{12} & \cdots & k_{1n} \\
k_{21} & k_{22} & \cdots & \cdots \\
\vdots & \vdots & \ddots & \vdots \\
k_{m1} & \cdots & \cdots & k_{mn}
\end{bmatrix}
\]

is called the matrix representation of \( L \) in the chosen bases. Note that the size of \( L \), namely \( m \times n \), is determined by \( \dim(\mathcal{Y}) \) and \( \dim(\mathcal{X}) \).

**Example 22** Let \( P^3 \) denote the vector space of all polynomials \( \alpha(t) = a_3 t^3 + a_2 t^2 + a_1 \) as in Example 18. Let \( P^3 \to P^2 \) denote the linear transformation which assigns to \( \alpha(t) \), its derivative \( \frac{d}{dt} \alpha(t) \). That is \( L(\alpha(t)) \triangleq \frac{d}{dt} \alpha(t) \). Let \( \{1, t, t^2\} \) and \( \{1, 1 + t\} \) be bases for \( P^3 \) and \( P^2 \) respectively. Then

\[
\begin{align*}
L(1) & = 0 = 0(1) + 0(1 + t) \\
L(t) & = 1 = 1(1) + 0(1 + t) \\
L(t^2) & = 2t = -2(1) + 2(1 + t)
\end{align*}
\]

Hence the matrix of \( L \) in the chosen bases is

\[
L = \begin{bmatrix}
0 & 1 & -2 \\
0 & 0 & 2
\end{bmatrix}
\]

Let \( \mathcal{X}, \mathcal{Y}, \) and \( \mathcal{Z} \) be vector spaces with bases \( \{ x_i : i \in \{1, 2, \ldots, n\}\} \), \( \{ y_i : i \in \{1, 2, \ldots, m\}\} \) and \( \{ z_i : i \in \{1, 2, \ldots, q\}\} \) respectively. Suppose that \( L, M \) and \( N \) are matrix representations of linear transformations \( L : \mathcal{X} \to \mathcal{Y}, M : \mathcal{X} \to \mathcal{Y} \) and \( N : \mathcal{Y} \to \mathcal{Z} \) respectively, in these bases. It can be shown in a straightforward manner that in these same bases, the matrix representations of the sum \( L + M \) and the composed transformation \( NM \) are \( L + M \) and \( NM \) respectively. Because of this there is a close correspondence between matrix addition and multiplication one the one hand and linear transformation addition and composition on the other.

### 3.2.4 Coordinate Transformations

Suppose as above, that \( L = [k_{ij}] \) is the matrix representation of \( L : \mathcal{X} \to \mathcal{Y} \) in the bases \( \{ x_j : j \in \{1, 2, \ldots, n\}\} \) and \( \{ y_i : i \in \{1, 2, \ldots, m\}\} \). Suppose that \( \{ \bar{x} : i \in \{1, 2, \ldots, n\}\} \) and \( \{ \bar{y}_i : i \in \{1, 2, \ldots, m\}\} \) is another pair of bases for \( \mathcal{X} \) and \( \mathcal{Y} \) respectively, and that \( \bar{L} = [\bar{k}_{ij}] \) is the representation of \( L \) in these bases. Our aim is to explain how these two representations of \( L \) are related. Since \( \{ \bar{y}_i : i \in \{1, 2, \ldots, m\}\} \) and \( \{ x_j : j \in \{1, 2, \ldots, n\}\} \) are bases for \( \mathcal{Y} \) and \( \mathcal{X} \) respectively, there are numbers \( p_{ir} \) and \( q_{sj} \) such that

\[
y_r = \sum_{i=1}^{m} p_{ir} \bar{y}_i, \quad r \in \{1, 2, \ldots, m\}
\]  

(3.13)
\[
\bar{x}_j = \sum_{s=1}^{n} q_{sj} x_s, \quad j \in \{1, 2, \ldots, n\} \quad (3.14)
\]

Therefore for \(j \in \{1, 2, \ldots, n\}\)

\[
L(\bar{x}_j) = L \left( \sum_{s=1}^{n} q_{sj} x_s \right) \quad \text{(by (3.14))}
\]

\[
= \sum_{s=1}^{n} q_{sj} L(x_s) \quad \text{(by linearity of } L) \]

\[
= \sum_{s=1}^{n} q_{sj} \left( \sum_{r=1}^{m} k_{rs} y_r \right) \quad \text{(by (3.12))}
\]

\[
= \sum_{s=1}^{n} q_{sj} \left( \sum_{r=1}^{m} k_{rs} \left( \sum_{i=1}^{m} p_{ir} \bar{y}_i \right) \right) \quad \text{(by (3.13))}
\]

Thus for \(i \in \{1, 2, \ldots, m\}\) and \(j \in \{1, 2, \ldots, n\}\)

\[
\bar{k}_{ij} = \sum_{s=1}^{n} q_{sj} \left( \sum_{r=1}^{m} k_{rs} p_{ir} \right)
\]

Using this formula for the \(ij\)th entry of \(\bar{L}\), it is straightforward {but tedious} to verify that

\[
\bar{L} = P L Q \quad (3.15)
\]

where \(P \triangleq \left[ p_{ir} \right]_{m \times m}\) and \(Q \triangleq \left[ q_{sj} \right]_{n \times n}\). Moreover \(P\) and \(Q\) are nonsingular matrices because of Proposition 4. The reader should try to verify these claims.

What (3.15) shows is that two different representations of \(L\), resulting from two different choices of bases for \(X\) and \(Y\), are related by a left-right equivalence transformation. Matrices \(P\) and \(Q\) correspond to changes of bases for \(Y\) and \(X\) respectively. From this discussion we can conclude that the matrix representation of a linear transformation is uniquely determined only up to a left-right equivalence transformation.

### 3.2.5 Defining a Linear Transformation on a Basis

One of the useful consequences of linearity is that it is possible to define a linear transformation by merely specifying its action on the elements of a basis of its domain. To illustrate this, suppose \(\mathcal{X}\) is a \(\mathbb{K}\)-vector space with basis \(\{x_i : i \in \{1, 2, \ldots, n\}\}\). Let \(\mathcal{Y}\) be any \(\mathbb{K}\)-vector space and let \(\{y_i : i \in \{1, 2, \ldots, n\}\}\) be any given set of \(n\) vectors in \(\mathcal{Y}\). The specification \(x_i \mapsto y_i, \ i \in \{1, 2, \ldots, n\}\), then completely defines a linear transformation \(L : \mathcal{X} \to \mathcal{Y}\). For if \(x\) is any vector in \(\mathcal{X}\) { i.e.,}

\[
x = \sum_{i=1}^{n} k_i x_i
\]

where the \(k_i\) are the coordinates of \(x\) in the basis \(\{x_i : i \in \{1, 2, \ldots, n\}\}\) then

\[
L(x) = \sum_{i=1}^{n} k_i L(x_i) = \sum_{i=1}^{n} k_i y_i
\]

because of linearity and the definitions \(L(x_i) \triangleq y_i, \ i \in \{1, 2, \ldots, n\}\). In other words, because of linearity, defining \(L\) on a basis for \(\mathcal{X}\) automatically defines \(L\) on all of \(\mathcal{X}\).
Let \( L : \mathcal{X} \to \mathcal{Y} \) and \( \bar{y} \in \mathcal{Y} \) be given. Our aim is to develop conditions in terms of \( L \) and \( \bar{y} \) for the existence of a vector \( x \in \mathcal{X} \) which solves the linear equation

\[
L(x) = \bar{y}
\]  

(3.16)

As a first step, we introduce a special set of vectors called the *image* of \( L \):

\[
\text{image } L \triangleq \{ L(x) : x \in \mathcal{X} \}
\]

In other words, the image of \( L \) consists of all vectors \( y \) such that \( y = L(x) \) for some \( x \in \mathcal{X} \).

Let us note that image \( L \) is a subspace of \( \mathcal{Y} \). For if \( y_1, y_2 \in \text{image } L \), then there must be vectors \( x_1, x_2 \in \mathcal{X} \) such that \( y_i = L(x_i), \ i \in \{1, 2, \ldots, 2\} \). From this it follows that for any \( k_1, k_2 \in \mathbb{K} \),

\[
k_1 y_1 + k_2 y_2 = k_1 L(x_1) + k_2 L(x_2) = L(k_1 x_1 + k_2 x_2)
\]

so \( k_1 y_1 + k_2 y_2 \in \text{image } L \). Thus image \( L \) is closed under addition and scalar multiplication which proves that image \( L \) is a subspace of \( \mathcal{Y} \).

Returning to the problem under consideration, let us note that (3.16) will have a solution \( x = \bar{x} \) if and only if

\[
\bar{y} \in \text{image } L
\]  

(3.17)

The claim is almost self-evident: it follows at once from the definition of image \( L \). For theoretical development, (3.17) proves to be a very useful condition for existence of solutions to (3.16). For numerical computation, something a little more concrete is needed. We’ll discuss this further a little later.

What we want to do next is to develop conditions for the existence of a solution \( X : \mathcal{Z} \to \mathcal{X} \) to the more general linear equation

\[
LX = M
\]  

(3.18)

where \( L : \mathcal{X} \to \mathcal{Y} \) and \( M : \mathcal{Z} \to \mathcal{Y} \) are given linear transformations. First suppose that (3.18) holds. Then for each \( z \in \mathcal{Z} \)

\[
L(x) = M(z)
\]  

(3.19)

where \( x = X(z) \). Thus for each vector \( z \in \mathcal{Z} \) there is a vector \( x \in \mathcal{X} \) such that (3.19) holds. This can be true only if

\[
\text{image } M \subset \text{image } L
\]  

(3.20)

so (3.20) is a necessary condition for (3.18) to have a solution \( X \).

Now suppose that (3.20) holds and let \( \{ z_i : i \in \{1, 2, \ldots, q\} \} \) be a basis for \( \mathcal{Z} \). Then \( M(z_i) \in \text{image } L, \ i \in \{1, 2, \ldots, q\} \). Therefore because of the definition of image, there are vectors \( x_1, x_2, \ldots, x_q \) in \( \mathcal{X} \) such that

\[
M(z_i) = L(x_i), \ i \in \{1, 2, \ldots, q\}
\]

Since \( \{ z_i : i \in \{1, 2, \ldots, q\} \} \) is a basis for \( \mathcal{Z} \), we can define \( X \) by simply specifying its action on this basis. In particular, define \( X \) so that \( X(z_i) = x_i, \ i \in \{1, 2, \ldots, q\} \). Thus

\[
M(z_i) = LX(z_i), \ i \in \{1, 2, \ldots, q\}
\]

Since \( \{ z_i : i \in \{1, 2, \ldots, q\} \} \) is a basis for \( \mathcal{Z} \) it must therefore be true that (3.18) holds. We have proved the following.

\[
LX = M
\]  

(3.18)
Proposition 9 Let $L: \mathcal{X} \to \mathcal{Y}$ and $M: \mathcal{Z} \to \mathcal{Y}$ be given linear transformations. There exists a linear transformation $X: \mathcal{Z} \to \mathcal{X}$ such that

$LX = M$ \hspace{1cm} (3.21)

if and only if

image $M \subseteq$ image $L$ \hspace{1cm} (3.22)

Note that (3.22) will necessarily hold {and (3.21) will consequently always have a solution} for any $M: \mathcal{Z} \to \mathcal{Y}$, provided image $L = \mathcal{Y}$. Linear transformations with this property are important enough to be given a special name. $L: \mathcal{X} \to \mathcal{Y}$ is said to be an epimorphism$^2$ if image $L = \mathcal{Y}$. Thus a sufficient condition for (3.21) to have a solution is that $L$ be an epimorphism. The reader may wish to verify that $L$ will be an epimorphism just in case any matrix representation of $L$ has all rows linearly independent.

Condition (3.22) is quite convenient for theoretical development. For computational purposes, one would apply the matrix version of this condition to matrix representations of $L$ and $M$.

As a first step toward the development of a matrix condition for existence, let us agree to call the subspace spanned by the columns of a given matrix $A$, the image$^3$ of $A$. Now in terms of matrix representations (3.18) is equivalent to

$LX = M$ \hspace{1cm} (3.23)

It is straightforward, but a little tedious, to show that (3.22) is equivalent to

image $M \subseteq$ image $L$ \hspace{1cm} (3.24)

This containment can also be interpreted without regard to (3.18), as a necessary and sufficient condition for the matrix equation (3.23) to have a solution $X$.

To develop the matrix version of (3.24), we first write (3.24) in the equivalent form

image $M +$ image $L =$ image $L$ \hspace{1cm} (3.25)

Using the identity

image $M +$ image $L =$ image $[M \quad L]$ \hspace{1cm} (3.26)

(3.25) can be rewritten as

image $[M \quad L] =$ image $L$ \hspace{1cm} (3.27)

{The reader should verify (3.26) and the equivalence of (3.25) and (3.24).}

We claim that (3.27) is equivalent to

$\dim$ (image $[M \quad L]) = \dim$ (image $L$) \hspace{1cm} (3.28)

That (3.27) implies (3.28) is clear. The reverse implication is a consequence of the general subspace implication $\dim (U) = \dim (V) \Rightarrow U = V$ which is valid in the special case when $U \subseteq V$.

To proceed we need to make use of the following fact.

Proposition 10 If $L: \mathcal{X} \to \mathcal{Y}$ is a linear transformation, and $L$ is any matrix representation of $L$, then

$\dim$ (image $L) = \dim$ (image $L) = \text{rank } L$

$^2$Other names sometimes used are surjective linear transformation or onto linear transformation.

$^3$Sometimes this subspace is called the column span of $A$. 
A proof of this proposition can be derived using Propositions 4 and 5.

In view of the equivalence of (3.28) and (3.24) we can state the following.

**Proposition 11** Let $M_{n \times m}$ and $L_{n \times p}$ be given matrices over $\mathbb{K}$. A necessary and sufficient condition for the equation

$$LX = M$$

to have a solution $X$ over $\mathbb{K}$ is that

$$\text{rank } [M \ L]_{n \times (m+p)} = \text{rank } L$$

The proposition is true for any field $\mathbb{K}$ including $\mathbb{R}$, $\mathbb{C}$, rational numbers, $\text{GF}(2)$, etc.

### 3.2.7 Linear Equations: Uniqueness, Kernel, Monomorphism

Again let $L : \mathcal{X} \to \mathcal{Y}$ and $\bar{y} \in \mathcal{Y}$ be given. Suppose that $\bar{x}$ is a solution to

$$L(x) = \bar{y} \quad (3.29)$$

Our aim is to describe all possible solutions to (3.29). For this we define the following set of vectors called the **kernel** of $L$.

$$\text{ker } L \triangleq \{ x : L(x) = 0, x \in \mathcal{X} \}$$

In other words, the kernel of $L$ is the set of all vectors $x$ in $\mathcal{X}$ satisfying $L(x) = 0$. The reader should verify that kernel $L$ is a subspace of $\mathcal{X}$.

Returning to our problem, suppose that $\tilde{x}$ is any solution to (3.29); i.e., $L(\tilde{x}) = \bar{y}$. Since $L(\tilde{x}) = \bar{y}$ it must be that $L(\tilde{x}) - L(\bar{x}) = 0$ and thus because $L$ is linear, that $L(\tilde{x} - \bar{x}) = 0$. Therefore

$$\tilde{x} - \bar{x} \in \text{ker } L \quad (3.30)$$

Conversely if (3.30) holds and $\bar{x}$ is a solution to (3.29) then $L(\tilde{x} - \bar{x}) = 0$ and $L(\tilde{x}) = \bar{y}$ so $\tilde{x}$ is a solution to (3.29) as well. In other words the set of all solutions to (3.29) consists of all vectors of the form $\tilde{x} + z$ where $\bar{x}$ is a solution and $z$ is a vector in kernel $L$.

From these observations it is clear that (3.29) will have at most one solution just in case kernel $L = 0$. Linear transformations with this property are important enough to be given a special name. $L : \mathcal{X} \to \mathcal{Y}$ is said to be a **monomorphism**\(^4\) if kernel $L = 0$. Thus a necessary and sufficient condition for (3.29) to have at most one solution is that $L$ be a monomorphism. The reader may wish to verify that $L$ will be a monomorphism just in case any matrix representation of $L$ has all columns linearly independent.

In the sequel we will use the identity

$$\dim (\text{image } L) + \dim (\text{kernel } L) = \dim (\mathcal{X}) \quad (3.31)$$

The formula is not hard to verify. For a derivation see page 64 of [1].

Suppose that $L$ is a $m \times n$ matrix representation of $L$ and suppose that $L$’s rank is $r$. Then $\dim \mathcal{X} = n$, $\dim \mathcal{Y} = m$, and $\dim (\text{image } L) = r$. Then $\dim (\text{kernel } L) = n - r$, because of (3.31). It follows that the number of linearly independent solutions to

$$Lx = \bar{y}$$

is $n - r + 1$ provided at least one solution exists in the first place.

\(^4\)Other names sometimes used are injective linear transformation or one-to-one linear transformation.
A linear transformation $L : X \rightarrow Y$ is an isomorphism if it is both a monomorphism and an epimorphism. Note that any matrix representation of an isomorphism must have both linearly independent rows and linearly independent columns and hence must be a square, nonsingular matrix.

Two $\mathbb{K}$-vector spaces $U$ and $V$ are isomorphic if there exists an isomorphism $T : U \rightarrow V$ mapping one space into the other. Suppose that $T$ is such an isomorphism. Then kernel $T = 0$ and image $T = V$. It follows from (3.31) that $\dim V = \dim U$. In other words, if two finite dimensional vector spaces are isomorphic then they must both have the same dimension. Conversely, if $U$ and $V$ are any two vector spaces with the same finite dimension - say $n$ - then they must be isomorphic. To prove that this is so, let $\{u_i : i \in \{1, 2, \ldots , n\}\}$ and $\{v_i : i \in \{1, 2, \ldots , n\}\}$ be bases for $U$ and $V$ respectively and define $T : U \rightarrow V$ so that $T(u_i) = v_i$, $i \in \{1, 2, \ldots , n\}$. We leave it to the reader to verify (1) that $T$ is a monomorphism, and (2) that $T$ is a epimorphism {and thus an isomorphism} and therefore that $U$ and $V$ are isomorphic as claimed.

The fact that all $n$-dimensional vector spaces over $\mathbb{K}$ are isomorphic has an important consequence. What’s implied is {say for $\mathbb{K} = \mathbb{R}$} that all real $n$-dimensional spaces differ from $\mathbb{R}^n$ by at most an isomorphism. It turns out that for many practical problems, this difference is not too important. For example, if a problem begins with an abstract vector space such as the vector space of all real polynomials of degree less than $n$, then the problem can be transformed via isomorphism into an equivalent problem in $\mathbb{R}^n$ where analysis can be carried out in more concrete terms.

### 3.2.9 Endomorphisms and Similar Matrices

A linear transformation $A : X \rightarrow X$ which maps a vector space into itself is called an endomorphism of $X$. In representing $A$ as a matrix, one always chooses the same basis for $A$’s domain $X$ and $A$’s codomain $X$.

Suppose that $\{x_i : i \in \{1, 2, \ldots , n\}\}$ and $\{\bar{x}_i : i \in \{1, 2, \ldots , n\}\}$ are two different bases for $X$ and that $A$ and $\bar{A}$ are the matrix representations of $A$ in these respective bases. Our aim is to explain how these two representations of $A$ are related. In the light of the discussion in section 3.2.4, it is clear that

$$\bar{A} = PAQ$$

where $P = \begin{bmatrix} p_{ir} \end{bmatrix}_{n \times n}$, $Q = \begin{bmatrix} q_{sj} \end{bmatrix}_{n \times n}$, and the $p_{ir}$ and $q_{sj}$ are numbers such that

$$x_r = \sum_{i=1}^{n} p_{ir} \bar{x}_i, \quad j \in \{1, 2, \ldots , n\}$$

Thus

$$\bar{x}_j = \sum_{s=1}^{n} q_{sj} \bar{x}_i, \quad i \in \{1, 2, \ldots , n\}$$

so

$$\bar{x}_j = \sum_{s=1}^{n} b_{ij} \bar{x}_i, \quad j \in \{1, 2, \ldots , n\}$$

where

$$b_{ij} = \sum_{s=1}^{n} p_{is} q_{sj}, \quad i, j \in \{1, 2, \ldots , n\}$$

These last two equations respectively imply that $\begin{bmatrix} b_{ij} \end{bmatrix}$ is a matrix representation of the identity map $x \rightarrow x$ and also the matrix product $PQ$, i.e.,

$$PQ = I_{n \times n}$$
Therefore $Q = P^{-1}$ so
\[ \tilde{A} = PAP^{-1} \] (3.32)

This is an example of "matrix similarity." In particular, two $n \times n$ \textit{K}-matrices $A$ and $B$ are said to be \textit{similar} if there exists a nonsingular \textit{K}-matrix $T$ such that
\[ B = TAT^{-1} \]

and the function $A \mapsto TAT^{-1}$ is called a \textit{similarity transformation}. It is easy to verify that matrix similarity is an equivalence relation on the class of all $n \times n$ matrices over \textit{K}.

What (3.32) shows is that two different matrix representations of an endomorphism $A$, resulting from two different choices of a basis for $X$, are related by a similarity transformation. From this we can conclude that the matrix representation of an endomorphism is uniquely determined only up to a similarity transformation. In the sequel we will study in detail those properties of a matrix $A$ which remain unchanged or invariant under similarity transformations. Such coordinate-independent properties typically turn out to be the physically significant attributes of the mathematical model under consideration.

This finishes our discussion of basis linear algebra. There are two more topics in the area which we intend to discuss. The first deals with similarity transformations and the second with congruence transformations. Both topics will be treated later in the notes. Similarity transformations are what result when one makes changes of variables in ordinary linear equations.
Chapter 4

Basic Concepts from Analysis

In the next chapter we will begin to study ordinary differential equations. To do this it is necessary to have in hand a number of basic concepts from calculus and real analysis. The aim of this short chapter is to briefly review the specific concepts which are need.

4.1 Normed Vector Spaces

In order to be able to talk about solutions to differential equations, we need to extend various ideas from single-variable real analysis or calculus to vectors of $n$-variables. The device which makes these extensions possible is the concept of a “norm.”

Let $V$ be any vector space over $\mathbb{K}$ {i.e., $\mathbb{R}$ or $\mathbb{C}$}; here $V$ need not even be finite dimensional. A nonnegative-valued function mapping domain $V$ into codomain $\mathbb{R}$, written $|| \cdot ||$, is a norm on $V$ if the following rules all hold:

i. $||v_1 + v_2|| \leq ||v_1|| + ||v_2||, \quad \forall v_i \in V$

ii. $||\mu v|| = |\mu||v||, \quad \forall v \in V, \mu \in \mathbb{K}$

iii. $||v|| = 0 \iff v = 0$

A vector spaces equipped with a norm is called a normed vector space. A norm can be thought of as a generalization to $V$ of the familiar idea of absolute value as defined on $\mathbb{K}$.

Rule i. is the triangle inequality. The requirement that a norm be nonnegative - valued at each point in its domain means that a norm is a positive semi-definite function. Rule iii. mean that $||v||$ can be zero if and only if $v = 0$; positive-semidefinite functions with this property are often call positive-definite functions.

For the present we will need norms defined only on the vector spaces $\mathbb{R}^n$ for various values of $n$. For each such vector space there are a great many ways to define a norm. For fixed $n$, one class of norms is those of the form

$$||x||_p \overset{\Delta}{=} \left( \sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$


where $p$ is a positive integer and $x_i$ is the $i$th entry of $x \in \mathbb{R}^n$. The most common of these are the one-norm

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

and the two-norm

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x'y}$$

which is usually called the Euclidean Norm. Also common is the infinity norm

$$||x||_\infty \triangleq \max_i |x_i|$$

One way to extend these definitions to the linear space $\mathbb{R}^{n \times m}$ for $m > 1$ is to define the $p$-norm of each matrix $A \in \mathbb{R}^{n \times m}$ to be the same as the $p$-norm of the $nm$-column vector obtained by stacking the columns of $A$ on top of each other. While this would be a perfectly good definition, it would fail to have the “submultiplicative property” discussed below, for values of $p > 1$. A better way to define the $p$-norm of $A$ is suggested by the observation that the definition of the $p$-norm of $x \in \mathbb{R}^n$ implies that

$$||x||_p = \sup_{y \in \mathbb{R}^m} \frac{||xy||_p}{||y||_p}$$

where $sup$ is an abbreviation for “supremum.” For $m > 1$, we define the $p$-norm of $A \in \mathbb{R}^{n \times m}$ to be

$$||A||_p \triangleq \sup_{y \in \mathbb{R}^m} \frac{||Ay||_p}{||y||_p}$$

It can be shown that

$$||A||_1 = \max_j \sum_{i=1}^{n} |a_{ij}| \quad \text{and} \quad ||A||_\infty = \max_i \sum_{j=1}^{m} |a_{ij}|$$

where $A = [a_{ij}]$. Latter in these notes it will be shown that $||A||_2$ is the square root of largest “eigenvalue” of the symmetric matrix $A'A$; this number, in turn, is sometimes called the largest “singular value” of $A$.

As we’ve already noted, the reason for defining $p$-norms is this way is because each has the sub-multiplicative property that is for all $p > 1$, including $p = \infty$,

$$||AB||_p \leq ||A||_p \cdot ||B||_p, \quad \forall A \in \mathbb{R}^{n \times m}, \forall B \in \mathbb{R}^{m \times r}$$

(4.1)

Observe that for $r = 1$, (4.1) is an immediate consequence of the definition of a $p$-norm. It should be noted that the sub-multiplicative property does necessarily not hold for other norms defined on $\mathbb{R}^{n \times m}$ nor for norms defined on infinite dimensional spaces.

It can be verified that for any fixed positive integers $p$, $q$, $n$, $m$, including possible $p = \infty$, there are constants $c_{pq}$ and $c_{qp}$ depending on $n$ and $m$, such that

$$||X||_p \leq c_{pq} ||X||_q$$

and

$$||X||_q \leq c_{qp} ||X||_p$$

(4.2)

for all matrices $X \in \mathbb{R}^{n \times m}$. This means, for example, that if $f : S \rightarrow \mathbb{R}^n$ is bounded with respect to $|| \cdot ||_p$, then it is bounded with respect to $|| \cdot ||_q$ for every value of $q$. To this extent all such $p$-norms on $\mathbb{R}^{n \times m}$ are equivalent. For many purposes, one such $p$-norm on $\mathbb{R}^{n \times m}$ proves to be just as useful as another. In the sequel we will often drop the subscript $p$ from $|| \cdot ||_p$ when what’s being discussed is valid for any $p$.

---

1. A function $f : S \rightarrow \mathbb{R}^n$ defined on some set $S$ is bounded (with respect to the $p$-norm $|| \cdot ||_p$) if there is a finite number $C$ such that $||f(x)||_p < C$ for all $x \in S$. If $f$ is bounded, the smallest value of $C$ with the aforementioned property is denoted by $\sup_S ||f||_p$ and is called the supremum of $f$ over $S$ with respect to $|| \cdot ||_p$. If $f$ is not bounded, $\sup_S ||f||_p \equiv \infty$. Why is it necessary to introduce the term “supremum” when we already have the term “max?” In other words, what’s the difference between sup and max?

2. This is not necessarily so for norms defined on infinite dimensional vector spaces.
4.1.1 Open Sets

Recall that an open interval of the real line, written \((t_1, t_2)\) is the set of point \(t : t_1 < t < t_2\). Here \(t_1\) and \(t_2\) can be either finite numbers or \(-\infty\) and/or \(+\infty\) respectively. Similarly, for finite \(t_1\) and \(t_2\) we can define a closed interval \([t_1, t_2]\) to be the set \(\{t : t_1 \leq t \leq t_2\}\). The difference between \([t_1, t_2]\) and \((t_1, t_2)\) is thus that the former contains its end points \(t_1\) and \(t_2\) whereas the latter does not. Right half-open intervals and left-half open intervals, written \([t_1, t_2)\) and \((t_1, t_2]\) respectively, are defined in the obvious ways.

Open and closed intervals are examples of open and closed sets in \(\mathbb{R}\). Other examples of open sets in \(\mathbb{R}\) are finite unions of open intervals. Similarly the union of a finite number of closed intervals is an example of a closed set in \(\mathbb{R}\). Using norms it is possible to extend the ideas of open and closed sets to vector spaces of any dimension. Let \(\mathcal{V}\) be any vector space \{finite dimensional or not\} and let \(|| \cdot ||\) be a norm defined on \(\mathcal{V}\). A neighborhood or open ball of radius \(r\) about a vector \(\vec{v} \in \mathcal{V}\) is the set of points

\[
\{v : ||v - \vec{v}|| < r\}
\]

An open set \(S \subset \mathcal{V}\) is simply any subset of \(\mathcal{V}\) whose elements all have neighborhoods which lie completely inside of \(S\). Thus the interior of a rectangle drawn on this page is an open set \(R\) in any real \(p\)-normed two-dimensional vector within which the page resides.

Note that the set \(R\) consisting of all points within the rectangle and on its boundary is not open because it is impossible to construct an open neighborhood of any point on the boundary which lies totally within \(R\).

The rectangle plus its boundary is an example of a “closed-set,” which in turn is a generalization of the idea of a closed interval. For any vector space \(\mathcal{V}\), a subset \(W\) is a closed set if its complement in \(\mathcal{V}\) is an open set.

4.2 Continuous and Differentiable Functions

Let \(f(\cdot)\) be a real-valued, scalar function whose domain \(\Omega\) is a connected subset\(^3\) of \(\mathbb{R}^n\). We assume that the reader knows what it means for \(f\) to be continuous or differentiable at a point \(\bar{x} \in \Omega\). Recall that \(f\) is a continuous function on \(\Omega\), if it is continuous at each point in its domain. We say that \(f\) continuously differentiable if the row vector of partial derivatives

\[
\frac{\partial f}{\partial x} \triangleq \left[ \frac{\partial f}{\partial x_1} \quad \frac{\partial f}{\partial x_2} \quad \ldots \quad \frac{\partial f}{\partial x_n} \right]
\]

\(^3\)By connected we mean that it is possible to connect any two points in \(\Omega\) with a curve which does not leave \(\Omega\).
exists at all points in $\Omega$ and each partial derivative is a continuous function on $\Omega$. For example, for even values of $p$, $f(x) \equiv ||x||_p$ is a continuously differentiable function on $\mathbb{R}^n$. All $p$ norms are continuous.

The notions of continuity, differentiability and continuous differentiability extend to vector-valued functions of $n$ variables in a natural way. For our purposes it suffices to say that a function $g : \mathbb{R}^n \rightarrow \mathbb{R}^m$ is continuous (resp. differentiable, continuously differentiable) if each component of $g$ is a continuous (resp. differentiable, continuously differentiable) scalar-valued function. By the way, note that the derivative of an $m$-vector

$$g \doteq \begin{bmatrix} g_1 \\ g_2 \\ \vdots \\ g_m \end{bmatrix}$$

whose elements $g_i$ are scalar-valued functions of an $n$-vector

$$x \doteq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

is an $m \times n$ matrix of the form

$$\frac{\partial g}{\partial x} \doteq \begin{bmatrix} \frac{\partial g_1}{\partial x_1} & \frac{\partial g_1}{\partial x_2} & \cdots & \frac{\partial g_1}{\partial x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial g_m}{\partial x_1} & \frac{\partial g_m}{\partial x_2} & \cdots & \frac{\partial g_m}{\partial x_n} \end{bmatrix}.$$  

In the event that $x$ is a differentiable function on $\mathbb{R}$, then

$$\frac{dx(x(t))}{dt}$$

is a $n$-vector whose $i$th component is

$$\frac{dx_i(x(t))}{dt}.$$

Note that by the chain rule

$$\frac{dg(x(t))}{dt} = g_x \frac{dx(x(t))}{dt},$$

where $g_x$ is the $m \times n$ matrix

$$g_x \doteq \left. \frac{\partial g(\mu)}{\partial \mu} \right|_{\mu = x(t)}.$$

In the event that $g$’s domain $\Omega$ is a interval {i.e., a connected subset of $\mathbb{R}$} it is possible to generalize slightly the concept of continuity. We say that $g$ is a piecewise-continuous function if, on each subinterval of $\Omega$ of finite length, (1) $g$ has at most a finite number of points of discontinuity and (2) at each such point $g$ has unique, finite limits which approached from above or below. A standard square wave on the real time interval $0 \leq t < \infty$ is a good example of the graph of a piecewise-continuous function.

Suppose $g$’s domain $\Omega$ is a an interval with finite-valued endpoints $t_1$ and $t_2$. If $g$ is at least piecewise continuous, the integral of each of its components over this interval is well-defined. We use the notation

$$\int_{t_1}^{t_2} g(\mu) d\mu.$$
to denote the \( m \)-vector whose \( i \)th component is
\[
\int_{t_1}^{t_2} g_i(\mu) d\mu
\]

Using the definition of the Riemann Integral of a continuous function, as the limit of a Riemann Sum, it is not difficult to prove that for \( t_2 \geq t_1 \),
\[
\left\| \int_{t_1}^{t_2} g(\mu) d\mu \right\| \leq \int_{t_1}^{t_2} \| g(\mu) \| d\mu
\]

In other words, the \( p \)-norm of the integral of a piecewise continuous function is less than or equal to the absolute value integral of the \( p \)-norm of the function. We’ll make use of this inequality a little later.

### 4.3 Convergence

A given sequence of vectors \( x_1, x_2, \ldots, x_i, \ldots \) in \( \mathbb{R}^n \) is said to \textit{converge} to a \textit{limit} \( \bar{x} \in \mathbb{R}^n \) if the sequence of numbers
\[
\| x_i - \bar{x} \|
\]
converges to zero. In other words \( x_1, x_2, \ldots, x_i, \ldots \) converges to a limit \( \bar{x} \) if for any given positive tolerance \{i.e., number\} \( \epsilon > 0 \), there is an integer \( N \) which is sufficiently large so that all for all \( i \geq N \) the normed difference \( \| x_i - \bar{x} \| \) is within the given tolerance; i.e.,
\[
\| x_i - \bar{x} \| < \epsilon, \quad \forall i \geq N
\]

When this so we call \( x_1, x_2, \ldots, x_i, \ldots \) a \textit{convergent sequence} and we write
\[
\lim_{i \to \infty} x_i = \bar{x} \quad \text{or} \quad x_i \to \bar{x}
\]

The concept of a norm thus links the idea of a convergent sequence of vectors to the familiar idea of convergence of a sequence of numbers.

Using the concept of a convergent sequence it is possible to redefine a \textit{closed set} to be any set \( S \) within which all sequences converge to limits which are also in \( S \). This definition proves to be entirely equivalent to the definition of a closed set given earlier.

As a practical matter it is usually of greater interest to determine if a given sequence converges to a limit, than it is to determine what the sequence’s limit is, if there is a limit at all. An important idea which enables one to address the convergence question without having to compute a limit, is the concept of a “Cauchy Sequence.” A given sequence of vectors \( x_1, x_2, \ldots, x_i, \ldots \) in \( \mathbb{R}^n \) is said to be a \textit{Cauchy Sequence} if for any given tolerance \( \epsilon > 0 \) there is an integer \( N \) which is sufficiently large so that for all \( i \geq N \) and all \( j \geq N \) the normed difference \( \| x_i - x_j \| \) is within the given tolerance; i.e.,
\[
\| x_i - x_j \| < \epsilon, \quad \forall i, j \geq N
\]

Let us note that any convergent sequence must be a Cauchy Sequence; for if \( \epsilon > 0 \) is any tolerance and \( N \) is such that
\[
\| x_k - \bar{x} \| < \frac{\epsilon}{2}, \quad \forall k \geq N
\]
then for all \( i, j \geq N \)
\[
\| x_i - x_j \| = \| (x_i - \bar{x}) + (\bar{x} - x_j) \| \leq \| x_i - \bar{x} \| + \| \bar{x} - x_j \| \leq \epsilon
\]
Chapter 4. Basic Concepts from Analysis

The notions of limit, convergence and Cauchy Sequence are not limited just to \( \mathbb{R}^n \) or even to normed spaces which are finite dimensional. Moreover convergent sequences are always Cauchy Sequences, whether the vector space in question is finite dimensional or not. On the other hand, Cauchy Sequences are not necessarily convergent sequences. Normed linear spaces which have the property that every Cauchy Sequence is convergent are said to be \textit{complete} vector spaces. Such spaces are often called \textit{Banach Spaces} after the mathematician who first studied them. Determining what kinds of normed spaces are Banach Spaces is a topic within the area of functional analysis. It can be shown that \( \mathbb{R}^n \) equipped with any \( p \)-norm is a Banach Space. It can also be shown that for any closed finite-length interval \([t_1, t_2]\), the linear vector space of continuous functions \( f: [t_1, t_2] \to \mathbb{R}^n \) equipped with the norm

\[
||f||_{\text{cont}} \Delta \max_{t \in [t_1, t_2]} ||f(t)||
\]

is a Banach Space. Another way of saying this is as follows.

\textbf{Theorem 1 (Uniform Convergence)} Let \( \mathbb{R}^n \) be equipped with any \( p \)-norm \( || \cdot || \). Let \([t_1, t_2]\) be a closed, bounded interval and let \( f_1, f_2, \ldots, f_i, \ldots \) be a sequence of continuous functions \( f_i: [t_1, t_2] \to \mathbb{R}^n \). Suppose that for any tolerance \( \epsilon > 0 \), there is a number \( N \) sufficiently large so that for every \( i, j \geq N \)

\[
\max_{t \in [t_1, t_2]} ||f_i(t) - f_j(t)|| < \epsilon
\]

(4.5)

Then there is a continuous function \( \bar{f}: [t_1, t_2] \to \mathbb{R}^n\) such that

\[
\max_{t \in [t_1, t_2]} ||f_i(t) - \bar{f}(t)|| \to 0 \quad \text{as} \quad i \to \infty
\]

(4.6)

Condition (4.5) says in effect that \( f_1, f_2, \ldots, f_i, \ldots \) is a Cauchy Sequence in the space of continuous functions \( f: [t_1, t_2] \to \mathbb{R}^n \) with norm (4.4). Condition (4.6) says that for any tolerance \( \bar{\epsilon} \) there must be a number \( \bar{N} \), not depending on \( t \) such that

\[
||f_i(t) - \bar{f}(t)|| < \bar{\epsilon}, \quad \forall t \in [t_1, t_2], i \geq \bar{N}
\]

The modifier \textit{uniform} is used to emphasize the fact that \( \bar{N} \) need not depend on \( t \). The proof of this theorem can be found in most basic real analysis texts.
Chapter 5

Ordinary Differential Equations - First concepts

There is probably no type of equation of greater importance in the sciences and engineering than a differential equation. Most of the physical laws we know {e.g., Newton’s Laws, Maxwell’s Equations, …} are characterized in terms of differential equations. The aim of this chapter is to discuss “ordinary” differential equations.

5.1 Types of Equation

There are a number of different types of equations which one might encounter in the modeling and analysis of physical processes:

Algebraic Equations:

An algebraic equation is typically of the form

\[ f(x, y, z, \ldots) = 0 \]

where \( f(\cdot) \) is an “algebraic function” and \( x, y, z, \ldots \) are variables. Such an \( f \) might look like

\[ f(x, y, z) = c_1 x^2 + c_2 xy + c_3 \sqrt{xy} + c_4 \frac{y^2}{x^3 + 1} \]

where the \( c_i \) are given numbers or “coefficients.” Systems of algebraic equations consist of a family of simultaneous algebraic equations such as the proceeding, with each equation having the same variables.

Partial Differential Equation:

A partial differential equation is typically of the form

\[ f \left( y, \frac{\partial y}{\partial t}, \frac{\partial y}{\partial x}, \frac{\partial^2 y}{\partial t \partial x}, \ldots, y, \frac{\partial z}{\partial t}, \frac{\partial z}{\partial x}, \frac{\partial^2 z}{\partial t^2}, \ldots, t, x \right) = 0 \]
where \( x \) and \( t \) are independent variables and \( y \) and \( z \) are functions of \( x \) and \( t \) or dependent variables. Dependent variables are variables which go their own ways or vary independently without regard to the equations involved. Typical examples are time and space variables. Independent variables are variables whose evolutions with respect to the equation’s independent variables are governed by the particular equation under consideration. Partial differential equations can have more than two independent variables. They can also have any number of dependent variables. It is also possible to have simultaneous partial differential equations, each equation with the same dependent and independent variables.

**Ordinary Differential Equation:**

Partial differential equations with just one independent variable are sufficiently special to be given their own name: they are called ordinary differential equations. In this course we will deal exclusively with ordinary differential equations. Partial differential equations will be covered in the sequel to this course given in the spring.

One fairly general type of ordinary differential equation, with \( y \) a dependent scalar-valued variable and \( t \) as the independent variable, is of the form

\[
h\left( \frac{d^n y}{dt^n}, \frac{d^{(n-1)} y}{dt^{(n-1)}}, \ldots, \frac{d^2 y}{dt^2}, \frac{dy}{dt}, y, t \right) = 0
\]

where \( h(\cdot) \) is some given function of its arguments. The order of this equation is the largest value of \( i \) for which

\[
\frac{d^i y}{dt^i}
\]

appears explicitly. An example of such an equation, of order six, is

\[
\sin \left( y^2 \frac{d^6 y}{dt^6} \right) + \log(t) \frac{d^3 y}{dt^3} + \cos(t) = 0
\]

There are two ways in which the preceding generalizes. First, the number of dependent variables might be greater than one and second, the number of equations might be greater than one. For example, one might have the simultaneous differential equations

\[
\begin{align*}
\frac{dy^2}{dt^2} + y \frac{d^4 z}{dt^4} e^{-t} & = 0 \\
\frac{d^3 z}{dt^3} + y & = 0
\end{align*}
\]

**Ordinary Recursion Equation:**

An ordinary recursion equation or difference equation is very much like an ordinary differential equation except that the dependent variable \( t \) take on only discrete values \( \{ \text{i.e., } t = 1, 2, 3, \ldots \} \) and the derivative \( \frac{d}{dt}(\cdot) \) is replaced with shift \( \delta (\cdot) \). Such an equation would be of the form

\[
h \left( \delta^n \{ y \}, \delta^{(n-1)} \{ y \}, \ldots, \delta^2 \{ y \}, \delta \{ y \}, y, \delta^{-1} \{ y \}, \delta^{-2} \{ y \}, \ldots, t \right) = 0
\]

where

\[
\delta^i \{ y \} \overset{\Delta}{=} y(t + i)
\]
An example of a recursion equation is
\[
\sin \left(y^2 \delta^6 \{y\}\right) + \log(t) \delta^3 \{y\} + \delta^{-2} \{y\} + \cos(t) = 0
\]

### 5.2 Modeling

The problem of deciding which differential equation or recursion equation is appropriate to describe the physical behaviour in a particular application is usually called modeling. Modeling problems range from very easy to extremely difficult. Although this is not a course on modeling, we shall nevertheless give several examples of the modeling process.

In the sequel we shall adopt Newton’s dot notation
\[
\dot{x} \triangleq \frac{dx}{dt}, \quad \ddot{x} \triangleq \frac{d^2x}{dt^2}, \ldots
\]

For recursion equations where \( t \) take on only discrete values, we shall adopt the diamond notation
\[
\diamond x \triangleq \delta \{x\} = x(t + 1), \quad \diamond \diamond x \triangleq \delta^2 \{x\} = x(t + 2), \ldots
\]

In neither case need the independent variable \( t \) necessarily represent time.

**Example 23** Newton’s Gravitational Law applied to two masses yields the nonlinear, second order differential equation
\[
m\ddot{x} = -\frac{K}{x^2}
\]

where \( K \) is a constant.

![Diagram of two masses](image)

**Example 24** If \( m \) were a given decreasing function of time, then in place of mass times acceleration, one would use the time rate of change of momentum. Thus for this case the correct equation would be
\[
\frac{d}{dt} \{m\dot{x}\} = -\frac{K}{x^2}
\]
or
\[
m\ddot{x} + \dot{m}x + \frac{K}{x^2}
\]

Here \( x \) is a dependent variable and \( m \) and \( \dot{m} \) are time-varying coefficients.
Example 25  Consider the following mechanical system consisting of the series connection of a stiff \{nonlinear\} spring with force/displacement function

\[ f(\mu) \triangleq K_1 \mu^3, \]

a mass \(M_1\) which slides on a surface with friction coefficient \(B\), a linear spring with force/displacement constant \(K_2\), and a mass \(M_2\) which slides on a frictionless surface.

To write the equations which model the system’s motion, let \(y_1\) and \(y_2\) denote the displacements of masses \(M_1\) and \(M_2\) from their respective equilibrium positions. Application of Newton’s Law provides

\[
\begin{align*}
M_1 \ddot{y}_1 & = -B \dot{y}_1 - K_1 y_1^3 - K_2 (y_1 - y_2) \\
M_2 \ddot{y}_2 & = -K_2 (y_2 - y_1)
\end{align*}
\]  

(5.1)  

(5.2)

It is possible to combine these two equations into a single equation with one dependent variable. For example, by twice differentiating (5.1) one gets

\[
M_1 \frac{d^4 y_1}{dt^4} = -B \frac{d^3 y_1}{dt^3} - K_1 \left\{ 6 y_1 \dot{y}_1^2 + 3 \dot{y}_1 y_1^2 \right\} - K_2 (\ddot{y}_1 - \ddot{y}_2)
\]  

(5.3)

Elimination of the term \((y_1 - y_2)\) from (5.1) and (5.2) and then solving for \(\ddot{y}_2\) provides

\[
\ddot{y}_2 = \frac{M_1}{M_2} \ddot{y}_1 - \frac{B_1}{M_2} \dot{y}_1 - \frac{K_1}{M_2} y_1^3
\]

Substitution into (5.3) results in the 4th - order nonlinear differential equation

\[
M_1 \frac{d^4 y_1}{dt^4} = -B \frac{d^3 y_1}{dt^3} - K_1 \left\{ 6 y_1 \dot{y}_1^2 + 3 \dot{y}_1 y_1^2 \right\} - K_2 \ddot{y}_1 - K_1 \left\{ \frac{M_1}{M_2} \ddot{y}_1 + \frac{B_1}{M_2} \dot{y}_1 + \frac{K_1}{M_2} y_1^3 \right\}
\]  

(5.4)

Example 26  In 1202 Leonardo of Pisa, known as Fibonacci, discovered the sequence of numbers

\[ 1, 1, 2, 3, 5, 8, 13, \ldots \]

while studying the population growth patterns of rabbits. This list of numbers \(y(1), y(2), \ldots\), now known as the Fibonacci Sequence, can be generated by the linear recursion equation

\[
y(t + 2) = y(t + 1) + y(t), \quad t \in \{1, 2, \ldots\}
\]

with the initial conditions \(y(1) \triangleq 1\) and \(y(2) \triangleq 1\). The sequence has a number of curious properties. For example the limit of the ratio

\[
\frac{y(t + 1)}{y(t)}
\]
as \( t \) tends to infinity turns out to be the “golden mean.\(^1\)” Another curious fact is that

\[
y(12j) = (12)^j, \quad j \in \{1, 2, \ldots\}
\]

Properties such as these are exploited in a variety of ways in the development of practical computational algorithms for many purposes. Later in the course, we will know enough about recursion equations to verify that these properties are indeed correct. By the way, do you think the sequence of prime numbers 1, 2, 3, 5, 7, 11, 13 \ldots can be generated be a \{linear\} recursion equation?

**Example 27** Suppose for a given a real-valued scaler function \( f \) of a real variable \( x \), it is desired to solve the equation

\[
f(x) = 0
\]

One way to try to do this is to generate a sequence of successive approximations to a solution \( x = \bar{x} \) using Newton’s Method. Suppose \( x_i \) is the \( i \)th approximation of \( \bar{x} \). The heuristic idea upon which the generation of the next approximation \( x_{i+1} \) is based is as follows. If \( \bar{x} \) were a correct solution, then one would have

\[
f(\bar{x}) = 0 \quad \text{or} \quad f(x_i + \delta_i) = 0
\]

where

\[
\delta_i \triangleq \bar{x} - x_i
\]

Formally expanding the preceding in a Taylor series about \( x_i \) one would get

\[
f(x_i) + f_x(x_i)\delta_i + \{\text{higher order terms in } \delta_i\} = 0
\]

where

\[
f_x(x) \triangleq \frac{df(x)}{dx}
\]

Therefore to first approximation

\[
\delta_i \approx \frac{f(x_i)}{f_x(x_i)} \quad \text{or} \quad \bar{x} \approx x_i - \frac{f(x_i)}{f_x(x_i)}
\]

Thus a reasonable way to generate \( x_{i+1} \) would be via the equation

\[
x_{i+1} = x_i - \frac{f(x_i)}{f_x(x_i)}
\]

Of course without saying more about \( f \), there is no guarantee that this equation will generate a sequence of values \( x_1, x_2, x_3, \ldots \) which converges to a value \( \bar{x} \) for which \( f(\bar{x}) = 0 \). One definition for \( f \) for which convergence from any initial chosen value \( x_1 \) can be proved is

\[
f(x) \triangleq x^2 - d
\]

where \( d \) is a positive number. Indeed this particular choice for \( f \) leads to the highly efficient algorithm

\[
x_{i+1} = \frac{x_i}{2} + \frac{d}{2x_i}
\]

for recursively computing the square root of \( d \).

\(^1\)Geometrically the **golden mean** is the ratio of length to width of any rectangle which has the property that the ratio of its length to its width equals the ratio of its length plus its width to its length. Rectangles of this type were thought by the ancient Greeks to have the most pleasing proportions possible and consequently were often used in their architecture.
5.3 State Space Systems

Given a differential equation (or a recursion equation) there arise several key issues:

1. How might one go about solving the equation?
   (a) What is meant by a solution?
   (b) When does such a solution exist and when is it unique?
   (c) How might a solution be found using analytic methods?
   (d) How might such a solution be found using a computer?
   (e) How might a solution be approximated when exact solvability is hopelessly difficult?

2. How might one go about determining various properties of whole families of solutions without having to actually construct the family?
   (a) What are the “equilibrium points” of a differential equation?
   (b) How might one predict periodic behavior or “limit cycles”?
   (c) How might one predict “unstable” or “chaotic” behavior?

In order to address these issues in a systematic manner, it is especially useful to focus on differential equation represented in “state space” form. By a state space system or dynamical system of dimension $n$ is meant a system of coupled first-order differential equations of the form

$$
\begin{align*}
\dot{x}_1 &= f_1(x_1, x_2, \ldots, x_n, t) \\
\dot{x}_2 &= f_2(x_1, x_2, \ldots, x_n, t) \\
&\vdots \\
\dot{x}_n &= f_n(x_1, x_2, \ldots, x_n, t)
\end{align*}
$$

Here the $x_i$ are scalar-valued dependent variables and each $f_i$ is a given scalar-valued function of the $x_i$ and the independent variable $t$. We’ll often refer to $t$ as time, but it could just as easily represent another quantity such as distance. The $x_i$ are often called state variables. An analogous discrete dynamical system would be of the form

$$
\begin{align*}
\dot{x}_1 &= f_1(x_1, x_2, \ldots, x_n, t) \\
\dot{x}_2 &= f_2(x_1, x_2, \ldots, x_n, t) \\
&\vdots \\
\dot{x}_n &= f_n(x_1, x_2, \ldots, x_n, t)
\end{align*}
$$

It is possible to denote the preceding systems a little more concisely using the notations

$$
\dot{x} = f(x, t) \quad \text{and} \quad \dot{x}^\circ = f(x, t)
$$

respectively where

$$
\begin{align*}
x &\triangleq \begin{bmatrix}
x_1 \\
x_2 \\
&\vdots \\
x_n
\end{bmatrix} \quad \text{and} \quad f(x, t) \triangleq \begin{bmatrix}
f_1(x_1, x_2, \ldots, x_n, t) \\
f_2(x_1, x_2, \ldots, x_n, t) \\
&\vdots \\
f_n(x_1, x_2, \ldots, x_n, t)
\end{bmatrix}
\end{align*}
$$
5.3.1 Conversion to State-Space Form

Many ordinary differential equations and recursion equations can be converted to equivalent state-space forms. The conversion process is often straight-forward. Consider for example, the \( n \)th order scalar differential equation

\[
h \left( \frac{d^n y}{dt^n}, \frac{d^{(n-1)} y}{dt^{(n-1)}}, \ldots, \frac{d^2 y}{dt^2}, \frac{dy}{dt}, y, t \right) = 0
\]

Suppose that it is possible to express \( \frac{d^n y}{dt^n} \) as a function of lower derivatives and \( t \):

\[
\frac{d^n y}{dt^n} = g \left( \frac{d^{(n-1)} y}{dt^{(n-1)}}, \ldots, \frac{d^2 y}{dt^2}, \frac{dy}{dt}, y, t \right)
\]

To carry out the conversion, one could define

\[
x_1 \triangleq y, \quad x_2 \triangleq \frac{dy}{dt}, \ldots, \quad x_n \triangleq \frac{d^{n-1} y}{dt^{n-1}}
\]

This would imply that

\[
\dot{x}_1 = \frac{dx}{dt} = x_2 \\
\dot{x}_2 = \frac{d^2 y}{dt^2} = x_3 \\
\vdots \\
\dot{x}_{n-1} = \frac{d^{n-1} y}{dy^{n-1}} = x_n \\
\dot{x}_n = \frac{d^n y}{dt^n} = g(x_n, x_{n-1}, \ldots, x_1, t)
\]

Thus in vector form, these equations would look like

\[
\dot{x} = f(x, t)
\]

and

\[
y = C x
\]

where

\[
f(x, t) \triangleq \begin{bmatrix}
  x_2 \\
x_3 \\
\vdots \\
x_n \\
g(x_n, x_{n-1}, \ldots, x_1, t)
\end{bmatrix} \\
C \triangleq \begin{bmatrix}
  1 & 0 & \cdots & 0
\end{bmatrix}_{1 \times n}
\]

This particular representation is sometime called \textit{phase variable form}. There are lots of other ways to define the \( x_i \).

\textbf{Example 28} If \( n = 2 \) and

\[
h(\ddot{y}, \dot{y}, y, t) \triangleq q(\dot{y}, y, t)\ddot{y} + p(\dot{y}, y, t)
\]

where \( p \) and \( q \) are given functions, then \( g \) would be

\[
g(\dot{y}, y, t) \triangleq -\frac{p(\dot{y}, y, t)}{q(\dot{y}, y, t)}
\]
In phase variable form \( x \) would be a two-vector and \( f \) would be
\[
f(x, t) = \begin{bmatrix} x_2 \\ -p(x_2, x_1, t) \\ q(x_2, x_1, t) \end{bmatrix}
\]

**Example 29** The original set of equations describing the motion of the masses in the mechanical system discussed in Example 25 were
\[
\begin{align*}
M_1 \ddot{y}_1 &= -B \dot{y}_1 - K_1 \dot{y}_1^3 - K_2 (y_1 - y_2) \quad (5.5) \\
M_2 \ddot{y}_2 &= -K_2 (y_2 - y_1) \quad (5.6)
\end{align*}
\]

Before we explained how to combine these two equations into the single 4th-order nonlinear differential equation
\[
M_1 \frac{d^4 y_1}{dt^4} = -B \frac{d^3 y_1}{dt^3} - K_1 \left( 6y_1 \dot{y}_1^2 + 3y_1^2 \ddot{y}_1 \right) - K_2 \dot{y}_1 - K_1 \left( \frac{M_1}{M_2} \ddot{y}_1 + \frac{B}{M_2} \dot{y}_1 + \frac{K_1}{M_2} y_1^3 \right)
\]

It is also possible to convert (5.5) and (5.6) into an equivalent state space system. For example, this can be done by simply defining
\[
x_1 \triangleq y_1, \quad x_2 \triangleq \dot{y}_1, \quad x_3 \triangleq y_2, \quad x_4 \triangleq \dot{y}_2
\]
These definitions would result in the state space system
\[
\begin{bmatrix}
\dot{x}_1 \\
\dot{x}_2 \\
\dot{x}_3 \\
\dot{x}_4
\end{bmatrix} = \begin{bmatrix}
0 & 1 & 0 & 0 \\
-\frac{K_2}{M_1} & -\frac{B}{M_1} & \frac{K_2}{M_1} & 0 \\
0 & 0 & 0 & 1 \\
\frac{K_2}{M_2} & 0 & -\frac{K_2}{M_2} & 0
\end{bmatrix} \begin{bmatrix}
x_1 \\
x_2 \\
x_3 \\
x_4
\end{bmatrix} - \begin{bmatrix}
0 \\
x_1 \\
x_2 \\
x_3
\end{bmatrix} \begin{bmatrix}
\frac{K_1}{M_1} \\
0 \\
0 \\
0
\end{bmatrix}
\]

**Example 30** As a final example, consider the following lumped, linear electrical network driven by a voltage \( v \).
To model this network’s behavior, let \( y_1 \) and \( y_2 \) denote the current through the inductor and voltage across the capacitor respectively. Then \( y_1 + C \dot{y}_2 \) must be the current through resistor \( R_1 \). Kirkoff’s voltage law dictates that the \( v \) must equal the sum of the voltage drops across resistor \( R_1 \) and the capacitor. In other words,
\[
v = R_1(y_1 + C \dot{y}_2) + L \dot{y}_1
\]
(5.7)

In addition, the voltage drop across the inductor must equal the sum of the voltage drops across resistor \( R_2 \) and the capacitor; i.e.,
\[
L \dot{y}_1 = R_2 \dot{y}_2 + y_2
\]
(5.8)

It is not very difficult to rewrite these equations in the equivalent state space form
\[
\begin{bmatrix}
\dot{y}_1 \\
\dot{y}_2
\end{bmatrix}
= \begin{bmatrix}
\frac{-R_1 R_2}{(R_1 + R_2) L} & \frac{R_2}{(R_1 + R_2) L} \\
\frac{-R_1}{(R_1 + R_2) C} & \frac{-1}{(R_1 + R_2) C} - \frac{R_2}{(R_1 + R_2) L}
\end{bmatrix}
\begin{bmatrix}
y_1 \\
y_2
\end{bmatrix}
+ \begin{bmatrix}
\frac{R_2}{(R_1 + R_2) L} \\
\frac{1}{(R_1 + R_2) C}
\end{bmatrix}
v
\]

There are systematic methods for constructing state space models of this type for extremely complicated \{vlsi\} networks with many thousands of circuit elements. Without such models and systematic methods for analyzing and synthesizing them, the design and implementation very large scale integrated electrical networks \{i.e., computer chips!\} would be a hopelessly difficult task.

5.4 Initial Value Problem

What we want to do now is to focus attention on state space differential equations of the form
\[
\dot{x} = f(x, t)
\]
(5.9)

where \( f : \mathbb{R}^n \times [0, \infty) \rightarrow \mathbb{R}^n \) is at least a continuous function of \( x \) and a piecewise continuous function of \( t \).

To grasp the main ideas of the theory of ordinary differential equations, it is necessary to understand what it means for these equations to have a “solution,” when such a solution exists, and when it is unique.\footnote{It is possible to make precise what is meant by a solution to \( (14.5) \) in the case when \( f \) is not a continuous function of \( x \) \[2\]. Equations of this type arise, for example, if one is modeling static friction in a mechanical system or if one is interested in describing the behavior of an electromechanical system employing relays or other types of switches. An example a differential equation with a discontinuous right hand side is \( \dot{x} = -\text{sign} \, x \) where \( \text{sign} \, x \triangleq 1 \) if \( x \geq 0 \) and \( \text{sign} \, x = -1 \) otherwise. The reader may wish to consider what a “solution” might look like for this system assuming \( x(0) = 0 \).}

By a solution to \( (14.5) \) on a finite interval \( I \subset [0, \infty) \) of positive length, is meant a continuous function \( \phi : I \rightarrow \mathbb{R}^n \) which is differentiable at all but at most a finite number of points in \( I \) and satisfies
\[
\dot{\phi} = f(\phi, t)
\]

where ever its derivative \( \dot{\phi} \) exists. \( \phi \) is a solution to \( (14.5) \) on an interval \( I \) of infinite length if its restriction\footnote{The restriction of a function \( h : S \rightarrow T \) to a subset \( U \subset S \) is the function \( k : U \rightarrow T \) for which \( k(u) = h(u) \forall u \in U \). Note that \( h \) and \( k \) are different functions even though both use the same rule to map points from \( U \) to \( T \)} to each finite subinterval of \( I \) is a solution to \( (14.5) \).

Example 31 The continuous function
\[
\phi(t) \triangleq |t - 2|, \quad t \in [0, \infty)
\]
is a solution on \([0, \infty)\) to the differential equation
\[
\dot{x} = f(x, t)
\]
where \( f : \mathbb{R} \times [0, \infty) \rightarrow \mathbb{R} \) is the function

\[
f(x, t) = \begin{cases} 
1 & \text{if } t \leq 2 \\
-1 & \text{if } t > 2 
\end{cases}
\]

for all \( x \in \mathbb{R} \). Note that \( \phi \) is piecewise continuous but not continuous.

For our purposes, the \textit{initial value problem} is as follows: Given an initial time \( t_0 \in [0, \infty) \) and an initial state vector \( x_0 \in \mathbb{R}^n \) find, if possible, an interval \( \mathcal{I} \subset [0, \infty) \) of positive length which contains \( t_0 \), and a solution \( \phi \) to (14.5) on \( \mathcal{I} \) which passes through the “state” \( x_0 \) at \( t = t_0 \). Let’s look at some examples.

**Example 32** Certainly the most basic of all initial value problems is to find a solution to the first-order, linear differential equation

\[
\dot{x} = a(t)x
\]

for a given initial time \( t_0 \geq 0 \) and state \( x_0 \in \mathbb{R} \). Here \( a : [0, \infty) \rightarrow \mathbb{R} \) is a given piecewise continuous function. We claim that

\[
\phi(t) = x_0 e^{\int_{t_0}^t a(\mu) d\mu}
\]

solves this initial value problem on the whole interval \([0, \infty)\). This is because \( \phi \) is well-defined and continuous on the interval \([0, \infty)\), because

\[
\dot{\phi}(t) = a(t)x_0 e^{\int_{t_0}^t a(\mu) d\mu} = a(t)\phi(t), \quad t \in [0, \infty),
\]

and because

\[
\phi(t_0) = x_0
\]

Note that if \( a(t) \) were a constant, say \( a(t) = \lambda \), then \( \phi \) would be the exponential time function

\[
\phi(t) = x_0 e^{(t-t_0)\lambda}
\]

The initial value problem turns out to always have a solution \( \phi \) on some interval \( \mathcal{I} \) provided \( f \) is a continuous function of \( x \) and a piecewise-continuous function of \( t \). We will not prove this, in part because the kind of \( f \)'s typically encountered in science and engineering are more than continuous in \( x \), and by exploiting this one can prove things more easily. Moreover, it turns out that continuity with respect to \( x \) is not a strong enough condition to ensure uniqueness of solutions.

**Example 33** The differential equation

\[
\dot{x} = \sqrt{x}
\]

together with the initial data \( t_0 \triangleq 0 \) and \( x_0 \triangleq 0 \) demonstrates nonuniqueness. In particular,

\[
\phi(t) = \left( \frac{2}{3} \right)^{\frac{3}{2}} \quad \text{and} \quad \phi(t) = 0
\]

both satisfy the differential equation and the initial condition \( \phi(0) = 0 \).

To avoid unpleasant uniqueness questions, we shall impose a slightly stronger smoothness condition on \( f \) then mere continuity with respect to \( x \). The specific properties we shall require \( f \) to have are stated below. Fortunately, the \( f \)'s encountered in the study of most problems of interest to scientists and engineers have these properties.
Properties:

1. For each closed, bounded subset \( S \subset \mathbb{R}^n \) and each bounded interval \( I \subset [0, \infty) \) there exists a constant \( \lambda \) such that
\[
||f(x, t) - f(y, t)|| \leq \lambda ||x - y||
\]
for all \( x, y \in S \) and all \( t \in I \).

2. For each fixed \( x \in \mathbb{R}^n \), the function \( f(x, \cdot) : [0, \infty) \to \mathbb{R}^n \), \( t \mapsto f(x, t) \) has at most a finite number of points of discontinuity on each bounded subinterval of \([0, \infty)\), and at each such point, \( f(x, \cdot) \) has unique limits when approached from above or from below.

A function \( f \) with the preceding properties is said to be \textit{locally Lipschitz Continuous} on \( \mathbb{R}^n \) and piecewise-continuous on \([0, \infty)\); and \( \lambda \) is often called a \textit{Lipschitz Constant}. If Property 1 holds for \( S \) equal to all of \( \mathbb{R}^n \), then \( f \) is \textit{globally Lipschitz continuous} on \( \mathbb{R}^n \). Of course in this case \( \lambda \) may still depend on \( I \).

Example 34 The function \( f(x) = x^2 \) is locally Lipschitz on \( \mathbb{R} \) because for any given closed, bounded subset \( S \subset \mathbb{R} \)
\[
||x^2 - y^2|| \leq \lambda ||x - y||, \quad \forall x, y \in S
\]
where
\[
\lambda \overset{\triangle}{=} \max_{x, y \in S} ||x + y|| < \infty
\]
Here we’ve used the relations
\[
||x^2 - y^2|| = ||(x + y)(x - y)|| \leq ||x + y|| ||x - y||
\]
and the fact that the maximum value \( \{ \text{i.e., } L \} \) of a continuous function \( \{ \text{i.e., } ||x + y|| \} \) over a closed, bounded set \( \{ \text{i.e., } S \times S \} \) is finite. Of course \( \lambda \) depends on what \( S \) is. Note that \( f \) is not globally Lipschitz on \( \mathbb{R} \) because there is no finite number \( \lambda \) such that
\[
||x^2 - y^2|| \leq \lambda ||x - y||, \quad \forall x, y \in \mathbb{R}
\]

Let us note that any function of the form
\[
f(x) = A_{n \times n} x + b_{n \times 1}
\]
is globally Lipschitz on \( \mathbb{R}^n \). This is because
\[
||(Ax + b) - (Ay + b)|| = ||A(x - y)|| \leq ||A|| ||x - y||, \quad \forall x, y \in \mathbb{R}^n
\]

It can be shown that any continuously differentiable function is locally Lipschitz and that any locally Lipschitz function is continuous. On the other hand, the reverse implications do not hold in general. For example, the scalar function \( f(x, t) \overset{\triangle}{=} |ax + b| \) is locally Lipschitz on \( \mathbb{R} \) but not continuously differentiable on \( \mathbb{R} \). Similarly the function \( f(x) = \sqrt{x} \) is continuous but neither locally Lipschitz nor continuously differentiable on \( \mathbb{R} \). The reader should verify these claims.

In the sequel it will be shown that initial value problem always has a unique solution \( \phi \) on some interval \( I \) provided \( f \) is a locally Lipschitz function of \( x \) and a piecewise-continuous function of \( t \). The possibility that \( I \) might have to be something less than the whole real interval \([0, \infty)\) is illustrated by the following example.

---

\(^4\)A subset \( S \subset \mathbb{R}^n \) is \textit{bounded} if there is a finite number \( C \) such that \( ||x|| < C \) for all \( x \in S \). If \( S \) is bounded, the smallest value of \( C \) with the aforementioned property is called the \textit{supremum or sup} of \( S \).
finding a solution to (5.12). One way to try to accomplish the latter is to generate a sequence of functions

\[ \phi(t) = \int_0^t f(x(t), t) \, dt, \quad t \in [0, \tilde{t}] \]

where \( \tilde{t} \triangleq t_0 + \frac{1}{x_0} \)

The reader may wish to verify that \( \phi(t) \) satisfies the initial condition and also the differential equation except at the time \( t = \tilde{t} \). Note however that since \( \phi \) is not continuous at \( t = \tilde{t} \) it cannot be a solution to the initial value problem. On the other hand, if we define \( \theta \) to be the restriction of \( \phi \) to \([0, \tilde{t}]\), i.e. \( \phi : [0, \tilde{t}] \to \mathbb{R} \) and

\[ \phi(t) \triangleq \theta(t), \quad t \in [0, \tilde{t}] \]

then \( \phi \) must be a solution since it is continuous, and satisfies both the initial condition and the differential equation. A little thought reveals that \([0, \tilde{t}]\) is actually the largest possible interval on which a solution to the problem exists. Note: It is quite incorrect to say that a solution exists at all points in \([0, \infty)\) except at \( t = \tilde{t} \). The behavior of \( \theta \) for \( t > \tilde{t} \) is irrelevant to the initial value problem.

### 5.5 Picard Iterations

Consider again the differential equation.

\[ \dot{x} = x^2 \]

As noted before, \( f(x) = x^2 \) is a locally Lipschitz function on \( \mathbb{R} \). Because of this, for each set of initial time \( t_0 \geq 0 \) and each initial condition \( x_0 \), the differential equation must have exactly one continuous solution \( \phi \) defined on some interval \( I \) which passes through \( x_0 \) at \( t_0 \). Fix \( t_0 \in [0, \infty) \) and \( x_0 > 0 \) and define the function \( \theta : [0, \infty) \to \mathbb{R} \) so that

\[ \theta(t) \triangleq \begin{cases} \frac{x_0}{1 - (t - t_0)x_0} & \text{if } t \neq \tilde{t} \\ 0 & \text{if } t = \tilde{t} \end{cases} \]

where

\[ \tilde{t} \triangleq t_0 + \frac{1}{x_0} \]

Consider the first order differential equation

\[ \dot{x} = f(x, t) \quad (5.11) \]

Suppose that \( f : \mathbb{R}^n \times [0, \infty) \to \mathbb{R}^n \) is locally Lipschitz in \( x \) and piecewise continuous in \( t \). Suppose that for given \( t_0 \geq 0 \) and \( x_0 \), the function \( \phi : [t_0, t_1] \to \mathbb{R}^n \) solves the initial value problem on some closed interval of positive length \([t_0, t_1]\). Since \( \phi \) is required to be continuous, \( f(\phi(t), t) \) must be piecewise continuous on \([t_0, t_1]\). From this and the requirement that \( \phi \) must satisfy (5.11) wherever \( \phi \) exists, it follows that \( \phi \) must satisfy the integral equation

\[ \phi(t) = x_0 + \int_{t_0}^t f(\phi(\mu), \mu) \, d\mu, \quad t \in [t_0, t_1] \quad (5.12) \]

Conversely any function satisfying \( \phi \) satisfying (5.12) must be continuous and must satisfy \( \phi(t_0) = x_0 \). In addition (5.11) must hold wherever \( \phi \) exists because of the fundamental theorem of calculus. Therefore finding a continuous solution \( \phi \) to (5.11) on \([t_0, t_1]\) which satisfies the initial condition \( \phi(t_0) = x_0 \) is equivalent to finding a solution to (5.12). One way to try to accomplish the latter is to generate a sequence of functions \( x_1(t), x_2(t), \ldots, x_i, \ldots \) using the recursion formula

\[ x_{i+1}(t) \triangleq x_0 + \int_{t_0}^t f(x_i(\mu), \mu) \, d\mu, \quad t \in [t_0, t_1], \quad i \in \{0, 1, 2, \ldots\} \quad (5.13) \]

\(^5\)All that follows can also be done for \( \phi \) defined on a closed interval \([t_a, t_b] \subset [0, \infty)\) of positive length whose interior contains \( t_0 \) or whose right end point \( t_b = t_0 \). The developments for these generalizations are essentially the same.
starting with the initial function

\[ x_0(t) \overset{\Delta}{=} x_0, \quad t \in [t_0, t_1] \quad (5.14) \]

For if such a sequence were to converge to a continuous limit \( \bar{x}(t) \) then as we shall soon see, \( \bar{x}(t) \) would have to satisfy

\[ \bar{x}(t) = x_0 + \int_{t_0}^{t} f(\bar{x}(\mu), \mu)d\mu, \quad t \in [t_0, t_1] \quad (5.15) \]

In other words if we can insure that the above sequence of functions has a continuous limit \( \bar{x} \), then we can define \( \phi \) to be \( \bar{x} \). This procedure is called Picard Iteration.

### 5.5.1 Three Relationships

In view of Example 35, it should not be surprising that for Picard Iteration to work, a constraint must be placed on the value of \( t_1 \). Basically what we need to do is to pick this number close enough to \( t_0 \) so as to insure that the value of each \( x_i(t) \), at each time \( t \in [t_0, t_1] \), remains within some set \( X \) on which \( f \) satisfies the Lipschitz condition

\[ ||f(x) - f(y)|| \leq \lambda ||x - y||, \quad \forall x, y \in X, \quad \forall t \in [t_0, t_1] \quad (5.16) \]

for some nonnegative Lipschitz Constant \( \lambda \). In other words \( t_1 \) and \( X \subset \mathbb{R}^n \) must be chosen so that \( t_0 < t_1 \), \( x_0 \in X \), and (5.16) and

\[ x_i(t) \in X, \quad t \in [t_0, t_1], \quad i \in \{0, 1, 2, \ldots, n\} \quad (5.17) \]

both hold. To do this, two distinct cases need to be considered:

1. **\( f \) is locally Lipschitz in \( x \):** For this case pick any finite positive number \( b \), any finite time \( \bar{t} > t_0 \), and define

   \[ X \overset{\Delta}{=} \{ x : ||x - x_0|| \leq b \} \]

   Next pick \( F \) to be any finite positive number such that

   \[ F \geq \sup_{x \in X, t \in [t_0, \bar{t}]} ||f(x, t)|| \]

   Note that such a finite \( F \) must exist because \( X \) and \( [t_0, \bar{t}] \) are both closed, bounded sets and because \( f \) is continuous in \( x \) and piecewise continuous in \( t \). Now define

   \[ t_1 \overset{\Delta}{=} \min\{ \bar{t}, t_0 + \frac{b}{F} \} \]

   The definitions imply that \( t_0 < t_1 \leq \bar{t} \), that

   \[ ||f(x, t)|| \leq F, \quad x \in X, t \in [t_0, t_1] \quad (5.18) \]

   and that

   \[ (t - t_0)F \leq b, \quad t \in [t_0, t_1] \quad (5.19) \]

   We claim that with these choices, the Picard Iterates \( x_i \) generated by (5.13) and (5.14) must satisfy (5.17). Now this is clearly true for \( i = 0 \) because of the definitions of \( x_0(t) \) and \( X \). Suppose therefore that for some \( j \geq 0 \), (5.17) hold for all \( i \in \{0, 1, \ldots, j\} \). Thus we can use integral inequality (4.3) together with (5.13), (5.18) and (5.19) to deduce, for \( t \in [t_0, t_1] \), that

   \[ ||x_{j+1}(t) - x_0|| = \left\| \int_{t_0}^{t} f(x_j(\mu), \mu)d\mu \right\| \leq \int_{t_0}^{t} ||f(x_j(\mu), \mu)||d\mu \leq \int_{t_0}^{t} Fd\mu = (t - t_0)F \leq b \]

   In other words, \( x_{j+1} \in X \) for \( t \in [t_0, t_1] \). By induction, (5.17) must therefore hold for all \( i \geq 0 \).
2. \( f \) is globally Lipschitz in \( x \): In this case pick \( t_1 \) to be any finite number greater than \( t_0 \) and define \( X \triangleq \mathbb{R}^n \). The \( x_i \) then satisfy (5.17). Moreover, since \( f \) is globally Lipschitz in \( x \) there must be a finite nonnegative Lipschitz Constant \( \lambda \) \{possibly depending on \( t_1 \} \) such that (5.16) holds.

To proceed, we need one additional relationship, namely

\[
||x_1(t) - x_0(t)|| \leq b, \quad t \in [t_0, t_1]
\] (5.20)

This holds for case 1, when \( f \) is only locally Lipschitz in \( x \), because of (5.17) and the definitions of \( x_0(t) \) and \( X \). We can also make (5.20) hold for case 2 by simply defining \( b \triangleq (t_1 - t_0)M \), where

\[
M \triangleq \sup_{t \in [t_0, t_1]} ||f(x_0, t)||
\]

This is because for case 2

\[
||x_1(t) - x_0|| = \left\| \int_t^{t_0} f(x_0(\mu), \mu)d\mu \right\| \leq \int_t^{t_0} ||f(x_0, \mu)||d\mu \leq \int_t^{t_0} M d\mu = (t - t_0)M \leq (t_1 - t_0)M
\]

for all \( t \in [t_0, t_1] \). Note that \( M \) must be finite because \( f(x_0, t) \) is piecewise continuous and \([t_0, t_1] \) is a closed, bounded interval.

5.5.2 Convergence

To recap, for either case we’ve been able to define \( t_0, t_1, b \) and \( X \) so that (5.16), (5.17), and (5.20) all hold. What we aim to do next is to use these three relations to prove for either case 1 or case 2, that the Picard Iterates \( x_1(t), x_2(t), \ldots, x_i(t) \) form a uniformly convergent sequence. As a first step toward this end we shall establish for each \( i \geq 0 \), that

\[
||x_{i+1}(t) - x_i(t)|| \leq b \frac{(t - t_0)\lambda^{i}}{i!}, \quad t \in [t_0, t_1]
\] (5.21)

where \( i! \) denotes \( i \) factorial\(^6\).

In view of (5.20) we already know that (5.21) holds for \( i = 0 \). Suppose therefore that for some \( j \geq 1 \), (5.21) holds for all \( i \in \{0, 1, \ldots, j - 1\} \). Because of (5.13) and (5.14) we can write

\[
||x_{j+1}(t) - x_j(t)|| = \left\| \int_t^{t_0} (f(x_j(\mu)), \mu) - f(x_{j-1}(\mu), \mu))d\mu \right\|
\]

\[
\leq \int_t^{t_0} ||f(x_j(\mu), \mu) - f(x_{j-1}(\mu), \mu)||d\mu
\]

Hence by (5.16) \{which holds because of (5.17)\}

\[
||x_{j+1}(t) - x_j(t)|| \leq \int_t^{t_0} \lambda ||x_j(\mu) - x_{j-1}(\mu)||d\mu
\]

But since (5.21) holds at \( i = j \),

\[
||x_{j+1}(t) - x_j(t)|| \leq \int_t^{t_0} \lambda b \frac{((\mu - t_0)\lambda)^j}{j!} d\mu = b \frac{(t - t_0)\lambda(j + 1)!}{(j + 1)!}
\]

\(^6\)Recall that \( 0! \triangleq 1 \).
Thus by induction, (5.21) holds for all $i \geq 0$.

Next we want to do is to prove for any integers $i \geq 0$ and $j \geq 1$ that

$$||x_{i+j}(t) - x_i(t)|| \leq b \frac{(t-t_0)^i}{i!} e^{(t-t_0)\lambda}$$

(5.22)

To do this we write

$$x_{i+j}(t) - x_i(t) = \sum_{k=0}^{j-1} (x_{i+k+1}(t) - x_{i+k}(t))$$

Using this, the triangle inequality and (5.21) we get

$$||x_{i+j}(t) - x_i(t)|| = \left|\sum_{k=0}^{j-1} (x_{i+k+1}(t) - x_{i+k}(t))\right|$$

$$\leq \sum_{k=0}^{j-1} ||x_{i+k+1}(t) - x_{i+k}(t)||$$

$$\leq b \sum_{k=0}^{j-1} \frac{(t-t_0)\lambda^{i+k}}{(i+k)!}$$

But

$$\frac{1}{(i+k)!} \leq \frac{1}{i!} \frac{1}{k!}$$

so

$$||x_{i+j}(t) - x_i(t)|| \leq b \frac{(t-t_0)^i}{i!} \sum_{k=0}^{j-1} \frac{(t-t_0)\lambda^k}{k!}$$

$$\leq b \frac{(t-t_0)^i}{i!} \sum_{k=0}^{\infty} \frac{(t-t_0)\lambda^k}{k!}$$

$$= b \frac{(t-t_0)^i}{i!} e^{(t-t_0)\lambda}$$

Thus (5.22) is true.

We are now ready to establish the convergence of the $x_i$. First, let us note from (5.22) that for $i \geq 0$ and $j \geq 0$

$$\max_{t \in [t_0, t_1]} ||x_{i+j}(t) - x_i(t)|| \leq b \frac{(t_1-t_0)^i}{i!} e^{(t_1-t_0)\lambda}$$

(5.23)

Let $\epsilon > 0$ be any positive tolerance. Pick an integer $N$ so large that

$$b \frac{(t_1-t_0)^i}{i!} e^{(t_1-t_0)\lambda} \leq \epsilon$$

for $i \geq N$. Then (5.23) enables us to write that

$$\max_{t \in [t_0, t_1]} ||x_k(t) - x_i(t)|| \leq \epsilon, \quad \forall k \geq N, \text{ and } \forall i \geq N$$

Since this is condition (4.5) of the Uniform Convergence Theorem, it must be that $x_0(t), x_1(t), \ldots$, is indeed a uniformly convergent sequence with a continuous limit; we henceforth denote this limit by $\bar{x}(t)$. 


Our next objective is to show that \( \bar{x} \) satisfies (5.15) which, for ease of reference we rewrite here:

\[
\bar{x}(t) = x_0 + \int_{t_0}^{t} f(\bar{x}(\mu), \mu) d\mu, \quad t \in [t_0, t_1]
\]  

(5.24)

For this let us first note that for \( i \geq 0 \) and \( t \in [t_0, t_1] \)

\[
\left\| \int_{t_0}^{t} f(x_i(\mu), \mu) d\mu - \int_{t_0}^{t} f(\bar{x}(\mu), \mu) d\mu \right\| \leq \int_{t_0}^{t} ||f(x_i(\mu), \mu) - f(\bar{x}(\mu), \mu)|| d\mu
\]

\[
\leq \lambda \int_{t_0}^{t} ||x_i(\mu) - \bar{x}(\mu)|| d\mu
\]

\[
\leq \lambda \max_{\tau \in [t_0, t_1]} ||x_i(\tau) - \bar{x}(\tau)|| \left( \int_{t_0}^{t} d\mu \right)
\]

\[
\leq (t_1 - t_0)\lambda \max_{t \in [t_0, t_1]} ||x_i(t) - \bar{x}(t)||
\]

Since

\[
\lim_{i \to \infty} \max_{t \in [t_0, t_1]} ||x_i(t) - \bar{x}(t)|| = 0
\]

it must therefore be true that

\[
\lim_{i \to \infty} \left\| \int_{t_0}^{t} f(x_i(\mu), \mu) d\mu - \int_{t_0}^{t} f(\bar{x}(\mu), \mu) d\mu \right\| = 0, \quad \forall t \in [t_0, t_1]
\]

Therefore for \( t \in [t_0, t_1] \)

\[
\int_{t_0}^{t} f(x_i(\mu), \mu) d\mu \to \int_{t_0}^{t} f(\bar{x}(\mu), \mu) d\mu \quad \text{and} \quad x_{i+1}(t) \to \bar{x}(t)
\]

as \( i \to \infty \). Since the \( x_i \) satisfy

\[
x_{i+1}(t) \Delta = x_0 + \int_{t_0}^{t} f(x_i(\mu), \mu) d\mu, \quad t \in [t_0, t_1], \quad i \in \{0, 1, 2, \ldots \},
\]

in the limit (5.24) must hold. \( \square \)

### 5.5.3 Uniqueness

In this section it will be shown that there is no continuous function defined on \([t_0, t_1]\), other than \( \bar{x} \), which satisfies (5.24). To prove that this is so, we will make use of a special technical result which also finds application in a number of other areas including most especially the stability analysis of ordinary differential equations.

**Lemma 1 (Bellman-Gronwall)** If some numbers \( t_b \geq t_a \), some constant \( c \geq 0 \) and some nonnegative, piecewise-continuous function \( \alpha : [t_a, t_b] \to \mathbb{R} \), \( w : [t_a, t_b] \to \mathbb{R} \) is a continuous function satisfying

\[
w(t) \leq c + \int_{t_a}^{t} \alpha(\mu) w(\mu) d\mu, \quad t \in [t_a, t_b]
\]

(5.25)

then

\[
w(t) \leq ce^{\int_{t_a}^{t} \alpha(\mu) d\mu}, \quad t \in [t_a, t_b]
\]

(5.26)
The Bellman-Gronwall Lemma becomes quite plausible as soon as one recognizes that the solution to the scalar differential equation,
\[ \dot{w} = \alpha w \]
\[ w(t_a) = c \]
or equivalent integral equation
\[ w(t) = c + \int_{t_a}^{t} \alpha(\mu)w(\mu)d\mu \]
is
\[ w(t) = ce^{\int_{t_a}^{t} \alpha(\mu)d\mu} \]
The lemma remains true if the right and/or left end point is removed from \([t_a, t_b]\).

**Proof:** Set
\[ v(t) = w(t)e^{-\int_{t_a}^{t} \alpha(\mu)d\mu} \]
\[ u(t) = \left( c + \int_{t_a}^{t} \alpha(\mu)w(\mu)d\mu \right) e^{-\int_{t_a}^{t} \alpha(\mu)d\mu} \]
and note from (5.25) that
\[ v(t) \leq u(t), \quad t \in [t_a, t_b] \quad (5.27) \]
Differentiating the expression for \( u \) and then replacing \( w \) with
\[ v(t)e^{\int_{t_a}^{t} \alpha(\mu)d\mu} \]
one gets
\[ \dot{u} = -\alpha u + \alpha we^{-\int_{t_a}^{t} \alpha(\mu)d\mu} = \alpha(v - u) \]
Hence by (5.27) \( \dot{u} \leq 0 \) so \( u(t) \leq u(t_0) = c, \quad t \in [t_a, t_b] \). Therefore by (5.27),
\[ v(t) \leq c, \quad t \in [t_a, t_b] \]
Multiplying through by \( e^{\int_{t_a}^{t} \alpha(\mu)d\mu} \) and then replacing \( v(t)e^{\int_{t_a}^{t} \alpha(\mu)d\mu} \) by \( w(t) \) yields (5.26) which is the desired result.

We now turn to the uniqueness question. For this, fix \( y_0 \in \mathbb{R}^n \) and suppose that \( y : [t_0, t_1] \to \mathbb{R}^n \) is any continuous function satisfying the initial condition
\[ y(t_0) = y_0 \quad (5.28) \]
and the equation
\[ \dot{y}(t) = f(y(t), t) \quad (5.29) \]
wherever \( \dot{y} \) exists. As we’ve already seen this is equivalent to \( y \) satisfying
\[ y(t) = y_0 + \int_{t_0}^{t} f(\bar{x}(\mu), \mu)d\mu, \quad t \in [t_0, t_1] \quad (5.30) \]
To establish \( \bar{x}'s \) uniqueness it is enough to show that if \( y_0 = x_0 \) then \( y(t) = x(t), \quad t \in [t_0, t_1] \).

To begin, let us first note that since \( \bar{x} \) and \( y \) are continuous functions and \([t_0, t_1]\) is a closed bounded interval, both \( \bar{x} \) and \( y \) must be bounded on \([t_0, t_1]\). Thus there must be a closed, bounded subset \( \bar{X} \subset \mathbb{R}^n \)
such that $\bar{x}(t) \in \bar{X}$ and $y(t) \in \bar{X}$ for all $t \in [t_0, t_1]$. Since $f$ is at least locally Lipschitz in its first argument, this means that there must be a finite constant $\lambda \geq 0$ such that

$$||f(y(t), t) - f(\bar{x}(t), t)|| \leq \lambda ||y(t) - \bar{x}(t)||, \quad t \in [t_0, t_1]$$

From this, (5.30), and (5.24) it thus follows that for $t \in [t_0, t_1]$

$$||y(t) - \bar{x}(t)|| = \left\| y_0 + \int_{t_0}^{t} f(y(\mu), \mu) d\mu - x_0 - \int_{t_0}^{t} f(\bar{x}(\mu), \mu) d\mu \right\|$$

$$\leq ||y_0 - x_0|| + \int_{t_0}^{t} ||f(y(\mu), \mu) - f(\bar{x}(\mu), \mu)|| d\mu$$

$$\leq ||y_0 - x_0|| + \lambda \int_{t_0}^{t} ||y(\mu) - \bar{x}(\mu)|| d\mu$$

Therefore by defining $c \triangleq ||y_0 - x_0||$, $t_a \triangleq t_0$, $t_b \triangleq t_1$, $w \triangleq ||y(t) - \bar{x}(t)||$, $\alpha \triangleq \lambda$ and applying the Bellman-Gronwall Lemma, we get

$$||y(t) - \bar{x}(t)|| \leq ||y_0 - x_0|| e^{(t-t_0)\lambda}, \quad t \in [t_0, t_1]$$

(5.31)

Therefore, if $y_0 = x_0$ it must be true that $||y(t) - \bar{x}(t)|| = 0, \quad t \in [t_0, t_1]$ and thus that $y(t) = \bar{x}(t)$ for $t \in [t_0, t_1]$.

5.5.4 Summary of Findings

The last few sections covered a lot of material. Here is a brief summary:

1. The problem of interest has been to find a solution to the initial value problem: Given is a differential equation

$$\dot{x} = f(x,t)$$

(5.32)

where $f : \mathbb{R}^n \times [0, \infty) \rightarrow \mathbb{R}^n$ is a function with the following properties:

**Locally Lipschitz Continuity in $x$:** For each closed, bounded subset $S \subset \mathbb{R}^n$ and each bounded interval $I \subset [0, \infty)$ there exists a constant $\lambda$ such that

$$||f(x,t) - f(y,t)|| \leq \lambda ||x - y||$$

(5.33)

for all $x, y \in S$ and all $t \in I$.

**Piecewise Continuity in $t$:** For each fixed $x \in \mathbb{R}^n$, the function $f(x, \cdot) : [0, \infty) \rightarrow \mathbb{R}^n$, $t \mapsto f(x, t)$ has at most a finite number of points of discontinuity on each bounded subinterval of $[0, \infty)$, and at each such point, $f(x, \cdot)$ has unique, finite limits when approached from above or from below.

Given also is an initial time $t_0$ and an initial state $x_0 \in \mathbb{R}^n$. The problem then is to find an interval $I \subset [0, \infty)$ of positive length which contains $t_0$, and a continuous function $\phi : I \rightarrow \mathbb{R}^n$ which passes through $x_0$ at $t = t_0$ and satisfies the equation

$$\dot{\phi}(t) = f(\phi(t), t)$$

(5.34)

at all points in $I$ where $\dot{\phi}$ exists.
2. We explained how to construct such a \( \phi \) for the case when \( I \) is a closed interval of positive but finite length with \( t_0 \) as its left end point. In particular, we showed how to pick a finite number \( t_1 > t_0 \) in such a way that the sequence of Picard Iterates

\[
x_{i+1}(t) \overset{\Delta}{=} x_0 + \int_{t_0}^{t} f(x_i(\mu), \mu) d\mu, \quad t \in [t_0, t_1], \quad i \in \{0, 1, 2, \ldots\}
\]

initialized with the function \( x_0(t) \overset{\Delta}{=} x_0, \; t \in [t_0, t_1], \) converges uniformly to such a \( \phi \). If \( t_0 \) is strictly greater than 0, then using essentially the same ideas, it is possible to pick a second number \( t_2 \in [0, t_0) \), so that Picard Iterates defined by \( x_0(t) \overset{\Delta}{=} x_0, \; t \in [t_2, t_0] \) and

\[
x_{i+1}(t) \overset{\Delta}{=} x_0 + \int_{t_0}^{t} f(x_i(\mu), \mu) d\mu, \quad t \in [t_2, t_0], \quad i \in \{0, 1, 2, \ldots\}
\]

also converge uniformly.

3. For the case when \( f \) is globally Lipschitz in \( x \) {i.e., when for each finite length interval \( I \subset [0, \infty) \) there is a constant \( \lambda \) for which (5.33) holds for all \( x, y \in \mathbb{R}^n \) and all \( t \in I \) all that is required of \( t_1 \) for the preceding to be true is that \( t_1 \) be finite and greater than \( t_0 \). It is also true that if \( t_0 > 0 \), then the only constraint on the time \( t_2 \) mentioned above, is that it be in the interval \([0, t_0)\).

4. We then used the Bellman-Gronwall Lemma to prove that if \( y_0 \in \mathbb{R}^n \) is arbitrary and \( y : [t_0, t_1] \to \mathbb{R}^n \) is any continuous function satisfying the initial condition \( y(t_0) = y_0 \), and the differential equation

\[
\dot{y}(t) = f(y(t), t)
\]

then

\[
||y(t) - x(t)|| \leq ||y_0 - x_0||e^{(t-t_0)\lambda}, \quad t \in [t_0, t_1]
\]

5. From the preceding we concluded that \( \phi \) is the only continuous function defined on \([t_0, t_1]\) which satisfies the initial condition \( \phi(t_0) = x_0 \) and the differential equation (5.34).

From the preceding it is possible to draw a number of useful conclusions:

I. More Uniqueness:

- Suppose \( \phi : [t_0, t_1] \to \mathbb{R}^n \) is the solution to the initial value problem which passes through \( x_0 \) at \( t = t_0 \).
- Suppose \( \bar{\phi} : [\bar{t}_0, \bar{t}_1] \to \mathbb{R}^n \) is the solution to the initial value problem which passes through \( \bar{x}_0 \) at \( t = \bar{t}_0 \).

If there is a time \( t^* \) in the interval \([\max\{t_0, \bar{t}_0\}, \min\{t_1, \bar{t}_1\}]\) at which \( \phi(t^*) = \bar{\phi}(t^*) \) then, as a consequence of uniqueness,

\[
\phi(t) = \bar{\phi}(t), \quad t \in [\max\{t_0, \bar{t}_0\}, \min\{t_1, \bar{t}_1\}]
\]

Moreover the the function \( \phi^* : [\min\{t_0, \bar{t}_0\}, \max\{t_1, \bar{t}_1\}] \to \mathbb{R}^n \)

\[
\phi^*(t) \overset{\Delta}{=} \begin{cases} 
\phi(t) & \text{for } t \in [t_0, t_1] \\
\bar{\phi}(t) & \text{for } t \in [\bar{t}_0, \bar{t}_1]
\end{cases}
\]

is unambiguously defined, passes through \( x_0 \) at \( t = t_0 \), and satisfies the differential equation (5.32) on \([\min\{t_0, \bar{t}_0\}, \max\{t_1, \bar{t}_1\}]\). This too is an immediate consequence of uniqueness.
II. Maximal Interval of Existence

For fixed \( t_0 \) and \( x_0 \) we know that the set

\[
S_1 \triangleq \{ t_1 : \text{a solution with value } x_0 \text{ at } t = t_0 \text{ exists on } [t_0, t_1] \text{ and } t_0 < t_1 < \infty \}
\]

is nonempty. Similarly if \( t_0 > 0 \), the set

\[
S_2 \triangleq \{ t_2 : \text{a solution with value } x_0 \text{ at } t = t_0 \text{ exists on } [t_2, t_0] \text{ and } 0 \leq t_2 < t_0 \}
\]

is also nonempty. Let us define\(^7\)

\[
\tilde{t}_1 \triangleq \sup S_1
\]

and

\[
\tilde{t}_2 \triangleq \begin{cases} 
0 & \text{if } t_0 = 0 \\
\inf S_2 & \text{otherwise}
\end{cases}
\]

Let us also define

\[
\mathcal{I}_{(t_0, x_0)} \triangleq \begin{cases} 
[0, \tilde{t}_1) & \text{if } t_0 = 0 \\
(\tilde{t}_2, \tilde{t}_1) & \text{otherwise}
\end{cases}
\]

Note whether \( \tilde{t}_1 \) is finite or not there must be a continuous function \( \tilde{\phi} : \mathcal{I}_{(t_0, x_0)} \to \mathbb{R}^n \) which satisfies (5.32) and passes through the initial state \( x_0 \) at \( t = t_0 \). \( \tilde{\phi} \) is called a *maximal solution* because \( \mathcal{I}_{(t_0, x_0)} \) is the largest possible interval for which there exists a solution \( \phi \) which passes through \( x_0 \) at time \( t_0 \). \( \mathcal{I}_{(t_0, x_0)} \) is often called the *maximal interval of existence* for the initial value problem under consideration. The uniqueness results discussed earlier, also apply to maximal solutions.

II. Global Existence: There are two important instances when \( \mathcal{I}_{(t_0, x_0)} = [0, \infty) \)

1. **If \( f \) is Globally Lipschitz:** In this case we know that a solution must exist on every closed interval \([t_2, t_1] \subset [0, \infty)\) of finite length which contains \( t_0 \). Thus in this case \( t_2 \) can be taken as \( t_2 = 0 \) and the supremum of such values of \( t_1 \) must be \( \infty \).

2. **If the Maximal Solution is Bounded:** Suppose that we happen to know that the maximal solution \( \tilde{\phi} \) is bounded. Then because of continuity, the limits

\[
\tilde{x}_i \triangleq \lim_{t \to \tilde{t}_i} \tilde{\phi}(t) \quad i \in \{1, 2, \ldots, 2\}
\]

must exist. Now if \( \tilde{t}_1 \) were finite there would have to be some interval \([\tilde{t}_1, \tilde{t}]\) of positive length on which there were a continuous solution \( \tilde{\phi} \) to (5.32) which passed through \( \tilde{x} \) at \( t = \tilde{t}_1 \). This in turn would imply that the concatenated function \( \theta : \{ \mathcal{I}_{(t_0, x_0)} \cup [\tilde{t}_1, \tilde{t}] \} \to \mathbb{R}^n \) defined so that

\[
\theta(t) = \tilde{\phi}(t), \quad t \in \mathcal{I}_{(t_0, x_0)}; \quad \text{and} \quad \theta(t) = \tilde{\phi}(t), \quad t \in [\tilde{t}_1, \tilde{t}]
\]

would be a continuous solution to (5.32) passing through \( x_0 \) at time \( t_0 \). But this cannot be so because \( t > \tilde{t}_1 \) and \( \mathcal{I}_{(t_0, x_0)} \) is the maximal interval of existence. In other words, \( \tilde{t}_1 \) cannot be finite. Similar reasoning can be used to prove that \( t_2 \) must be \( 0 \).

---

\(^7\)The *inf or infimum of a subset \( S \subset \mathbb{R}^n \) is the smallest number \( \mu \) such that \( ||x|| \leq \mu \) for all \( x \in S \).
The following theorems summarize what we’ve uncovered so far.

**Theorem 2 (Local Existence)** Suppose that \( f : \mathbb{R}^n \times [0, \infty) \rightarrow \mathbb{R}^n \) is a function which is locally Lipschitz in \( x \) and piecewise continuous in \( t \) as defined previously. For each \( x_0 \in \mathbb{R}^n \) and each \( t_0 \in [0, \infty) \) there exists a largest interval \( \mathcal{I}_{(t_0,x_0)} \subset [0, \infty) \) and exactly one continuous solution \( \phi \) to

\[
\dot{x} = f(x, t)
\]

on \( \mathcal{I}_{(t_0,x_0)} \) which passes through \( x_0 \) at \( t = t_0 \). Moreover, the restriction of \( \phi \) to any given subinterval \( \mathcal{I} \subset \mathcal{I}_{(t_0,x_0)} \) containing \( t_0 \), is the only continuous function which passes through \( x_0 \) as time \( t_0 \) and is a solution to (5.37) on \( \mathcal{I} \).

**Theorem 3 (Global Existence)** Suppose that under the conditions of Theorem 2, \( f \) is is globally Lipschitz in \( x \). Then for each \( x_0 \in \mathbb{R}^n \) and each \( t_0 \in [0, \infty) \), the maximal interval of existence is \( [0, \infty) \).

Linear differential equations of the form

\[
\dot{x} = A(t)x + b(t)
\]

are equations for which \( f \triangleq A(t)x + b(t) \). Since such \( f \)'s are globally Lipschitz \( x \), solutions to such linear differential equations always exist on \([0, \infty)\), assuming of course that \( A(t) \) and \( b(t) \) are piecewise continuous matrices. As we have already seen, the other instance when \( \mathcal{I}_{(t_0,x_0)} = [0, \infty) \) for a given maximal solution \( \phi \) to the initial value problem, is when the maximal solution in question is known to be bounded wherever it exists.

**Theorem 4 (Global Boundedness)** If, under the conditions of Theorem 2, the maximal solution to the initial value problem which passes through \( x_0 \) at \( t = t_0 \) is bounded, then the maximal interval of existence is \([0, \infty)\).

### 5.6 The Concept of State

Consider again the differential equation

\[
\dot{x} = f(x, t)
\]

where \( f : \mathbb{R}^n \times [0, \infty) \rightarrow \mathbb{R}^n \). For the remainder of these notes, unless otherwise stated we will assume that \( f(x, t) \) is at least locally Lipschitz Continuous in \( x \) and piecewise continuous in \( t \).

It is sometimes useful to think of (5.38) as a dynamical system \( \Sigma \) whose possible “states” are the vectors in \( \mathbb{R}^n \). \( \Sigma \) is said to be in state \( z \) at time \( t = \tau \) if \( z \) is the value at time \( \tau \) of the solution to (5.38) under consideration. \( \Sigma \)’s state space \( \mathbb{R}^n \) is thus the space in which solutions to (5.38) take values.

**Example 36** There are dynamical systems whose state spaces have more structure than \( \mathbb{R}^n \). For example, the two scalar differential equations

\[
\begin{align*}
\dot{x}_1 &= x_2 g(x_1, x_2) \\
\dot{x}_2 &= -x_1 g(x_1, x_2)
\end{align*}
\]

with \( g \) any scalar-valued function, can be viewed as a dynamical system whose state space is the unit circle \( \mathcal{C} \) centered at 0 in \( \mathbb{R}^2 \). This is because the differential equations imply that

\[
\frac{d}{dt} (x_1^2 + x_2^2) = 0
\]
and thus that $x^2 + x_2^2 = c^2$, for some constant $c$. This means that $\Sigma$’s state vector $x$ evolves on the unit circle - assuming of course $x$ started out there in the first place.

Note that the concept of state as defined here is consistent with what’s usually meant by the term in a non-technical context. In particular, knowledge of $\Sigma$’s state at time $\tau$ is sufficient to uniquely determine the future behaviour of $x$ for all $t \geq \tau$. This of course is because of uniqueness - i.e., there can be exactly one maximal solution to (5.38) passing through state $z$ at time $\tau$.

For each $z \in \mathbb{R}^n$ and each $\mu \geq 0$, let $\zeta(\cdot, \mu, z) : \mathcal{I}(\mu, z) \to \mathbb{R}^n$ denote the maximal solution to (5.38) which passes through state $z$ at time $\mu$. Then because of uniqueness it clearly must be true that for each $t_0 \in [0, \infty)$ and each $x_0 \in \mathbb{R}^n$

$$\zeta(t, t_0, x_0) = \zeta(t, \tau, \zeta(\tau, t_0, x_0)), \quad \forall \, t, \tau \in \mathcal{I}(t_0, x_0)$$

This is sometimes called the *composition rule* for dynamical systems. It says that the state of $\Sigma$ at time $t$, assuming $x_0$ was $\Sigma$’s state at time $t_0$, is the same as $\Sigma$’s state at time $t$ assuming $\Sigma$ was in state $\zeta(\tau, t_0, x_0)$ at time $\tau$. Thus the value of $\Sigma$’s state at time $\tau$ is sufficient to determine the value of $\Sigma$’s state at any future time $t \in \mathcal{I}(t_0, x_0)$.

### 5.6.1 Trajectories, Equilibrium Points and Periodic Orbits

The set of states which $\zeta(t, \tau, z)$ passes through at $t$ ranges over the interval $\mathcal{I}(\mu, z)$ from $\tau$ onward to its right end point is often called the *trajectory* of $\Sigma$ which leaves state $z$ at time $\tau$. An important consequence of uniqueness is that for fixed $\tau$, there is exactly one trajectory leaving each state $z \in \mathbb{R}^n$. Moreover, trajectories starting at the same time but from different states cannot cross in finite time.

The simplest example of a trajectory is an “equilibrium point.” A vector $z$ in $\mathbb{R}$ at which

$$f(z, t) = 0, \quad \forall t \in [0, \infty)$$

is called an *equilibrium point* of $\Sigma$. A given $f$ can have no, one, a finite number or even a continuum of equilibrium points. If $z$ is an equilibrium point, then clearly for each $t_0 \geq 0$, the constant function

$$\zeta(t, t_0, z) \triangleq z, \quad \forall t \in [0, \infty)$$

is a maximal solution to (5.38). Thus each equilibrium point of (5.38) is a trajectory. Since trajectories starting at the same time but in different states cannot cross in finite time, no equilibrium point can be reached from any other state in finite time.

**Example 37** The differential equation

$$\dot{x} = x^2 + 1$$

has no equilibrium points, whereas the differential equation

$$\begin{align*}
\dot{x} &= x^2 + xy \\
\dot{y} &= y^2 + xy
\end{align*}$$

has infinitely many, namely all points for which $x + y = 0$. The differential equation

$$\dot{x} = x^2 + 3x + 2$$
5.6. THE CONCEPT OF STATE

has two equilibrium points, namely $-1$ and $-2$. For $A$ a $n \times n$ nonsingular matrix the equation

$$\dot{x} = Ax$$

has only $x = 0$ as an equilibrium point.

An equilibrium point $z$ of (5.38) is an attractor if trajectories starting “close” enough to $z$ tend to $z$ as $t \to \infty$. In other words, $z$ is an attractor if there is a number $\epsilon > 0$ such that for all $t_0 \geq 0$ and all $x_0 \in \mathbb{R}^n$ satisfying $||x_0 - z|| \leq \epsilon$,

$$\lim_{t \to \infty} \zeta(t, t_0, x_0) = z$$

To determine if an equilibrium point $z$ is attractive, one looks at the behavior of trajectories in the vicinity of $z$. There are several ways to do this. One way, which is applicable in the case when $f$ is not dependent on $t$, involves examining the properties of

$$\frac{\partial f(x)}{\partial x} \bigg|_{x=z}$$

We will return to this topic later.

Example 38 The equilibrium point $x = y = 0$ of the dynamical system

$$\dot{x} = -x$$
$$\dot{y} = -y$$

is {globally} attractive because all solutions tend to zero at $t \to \infty$. On the other hand, for any nonnegative number $\lambda$, the equilibrium point $x = y = 0$ of the dynamical system

$$\dot{x} = \lambda x$$
$$\dot{y} = -y$$

is not attractive since no trajectory starting at $y \neq 0$ ever tends to the zero state.

A maximal solution $\zeta(t, t_0, x_0)$ to (5.38) is said to be periodic if it is either a constant or if there exists a finite number $T > 0$ such that $\zeta(t+T, t_0, x_0) = \zeta(t, t_0, x_0)$ for all $t \in I_{(t_0, x_0)}$. The smallest value of $T$ for which this is true is the period of $\zeta$. Note that if $\zeta(t, t_0, x_0)$ is periodic, it must be bounded. Thus in this case, because of the Global Boundednes Theorem, the right end point of $I_{(t_0, x_0)}$ must be infinity; i.e., periodic solutions go on forever.

Note that the trajectory of a periodic solution must be a closed curve or orbit in $\mathbb{R}^n$. The reverse is also true if $f$ does not depend on $t$. In other words, if $f$ is independent of $t$, then maximal solutions to (5.38), whose trajectories are closed orbits, are periodic solutions. The reader should try to prove this by exploiting both the time invariance of $f$ and the uniqueness of solutions to (5.38).

Example 39 The differential equation

$$\dot{x} = 3y$$
$$\dot{y} = -5x$$

has one equilibrium point, namely $x = y = 0$, but it is not attractive. On the other hand all maximal solutions are periodic. Each generates a closed orbit which is an ellipse in $\mathbb{R}^2$. 
Consider the time-independent or autonomous differential equation

\[ \dot{x} = f(x) \]  

(5.40)

where \( f : \mathbb{R}^n \to \mathbb{R}^n \) is a function whose first partial derivatives \( \frac{\partial f}{\partial x^i} \), \( i \in \{1, 2, \ldots, n\} \) are continuous functions on \( \mathbb{R}^n \). As before, let \( \zeta(t, \tau, z) \) denote the maximal solution to (5.40) which passes through state \( z \) at \( t = \tau \). The first consequence of time invariance is that for each \( t_0 \geq 0 \) and each \( x_0 \in \mathbb{R}^n \)

\[ \zeta(t, 0, z) = \zeta(t + t_0, t_0, z), \quad t \in \mathcal{I}_{(0, z)} \]  

(5.41)

In other words, the value at each time \( t \) of maximal solution starting in state \( z \) at time 0 is the same as the value at time \( t + t_0 \) of the maximal solution which passes through \( z \) at time \( t_0 \). {The reader should verify that the time-independence of \( f \) and the uniqueness of solutions to (5.40) imply that this is so.} Because of this property, there is no reason to not take \( t_0 = 0 \). And so for the time-invariant case we always will.

A set \( V \subset \mathbb{R}^n \) is called is said to be invariant with respect to (5.40) if every maximal solution to (5.40) which starts at a state in \( V \), remains in \( V \) forever. In other words, for \( V \) to be an invariant set, all solutions starting at states in \( V \) must exist for all time and moreover, each state along such a solution must lie on a trajectory which is a subset of \( V \).

**Example 40** The set of points in \( \mathbb{R}^2 \) for which \( y = 0 \) is an invariant subset \( V \) for the differential equation

\[ \begin{align*}
\dot{x} &= -x \\
\dot{y} &= -y 
\end{align*} \]

In this case \( V \) is actually a subspace.

Equilibrium points, closed orbits, and trajectories of solutions which exist for all time are also examples of invariant sets. A more interesting class of invariant sets can be characterized as follows.

Call a function \( F : \mathbb{R}^n \to \mathbb{R} \) a first integral of (5.40) if

\[ \frac{\partial F}{\partial x} f(x) = 0, \quad \forall x \in \mathbb{R}^n \]

If \( F \) is a first integral and \( \phi \) is a solution to (5.40), then

\[ \frac{d}{dt} F(\phi(t)) = 0, \quad t \geq 0 \]

What this means is that if \( F \) is a first integral, then \( F(x) \) is constant along trajectories in \( \mathbb{R}^n \).

If \( F \) is a first integral, then for each constant \( c \), the set

\[ S_c \overset{\Delta}{=} \{ x : F(x) = c, \ x \in \mathbb{R}^n \} \]

is called an integral manifold. Note that any such integral manifold is an invariant set provided the maximal solutions which generate the trajectories in the manifold exist for all time.
Example 41 The function $F(x, y) \triangleq x^2 + y^2$ is a first integral for
\[
\begin{align*}
\dot{x} &= y \\
\dot{y} &= -x
\end{align*}
\]
The integral manifold determined by
\[ F(x, y) = 2 \]
is a circle of radius 2 in $\mathbb{R}^2$.

5.7 Simulation

As we have already seen, it is possible to use Picard Iteration to compute an approximate solution to an initial value problem on a prespecified interval of existence $[t_0, t_1]$. For this, one needs a good numerical procedure {e.g., Simpson’ Rule} for calculating the integral, as a function of $t \in [t_0, t_1]$, of a function defined on $[t_0, t_1]$. Such procedures are readily available and can be found in most numerical software packages; in Matlab one could use QUAD or one of its variants.

It turns out that Picard Iteration is usually not the best way to go about computing a solution to an initial value problem. One important reason for this is that Picard Iteration requires one to specify $t_1$ prior to the initiation of the computation. For many problems of interest, the length of time over which one seeks a solution in open-ended and is only decided upon by studying the behavior of the solution as it evolves in time. An alternative approach which avoids this limitation is to compute an approximate solution by iterating with respect to the independent variable time. The basic idea is quite simple and can be illustrated with $x$ a scalar.

5.7.1 Numerical Techniques

Suppose $f : \mathbb{R} \times [0, \infty) \to \mathbb{R}$ is a given scalar valued function and it is desired to compute a solution to
\[
\dot{x} = f(x, t) \tag{5.42}
\]
with initial state $x_0$ at $t = 0$. Let us pick a small positive number $\Delta$ called a step size. Suppose that $\phi(t)$ is the exact solution to (5.42) with initial value $\phi(0) = x_0$. What we want to do is to develop a procedure for computing numbers $y(1), y(2), y(3), \ldots$ such that
\[
y(i) \approx \phi(i\Delta), \quad i \geq 0
\]

To start things off correctly, let us define
\[
y(0) \triangleq x_0
\]
We approach the problem of computing the $y(i), \ i \geq 1$ in much the same way as one approaches a proof by induction. We presume that for some $j \geq 0$, we’ve already computed an approximation
\[
y(j) \approx \phi(j\Delta) \tag{5.43}
\]

On the other hand, for “two-point” boundary value problems where some of the components of $x$ are given at $t_0$ while others are given at $t_1$, Picard Iteration or something similar may be be preferable.
and we seek a procedure for computing \( y(j+1) \). There are a great many different ways to go about doing this. Perhaps the most straightforward is “Euler’s Method” which is a simple procedure based on a one term Taylor series expansion of \( \phi(\Delta(j+1)) \) about the time \( t = \Delta j \). In particular \( \phi(\Delta(j+1)) \) admits the expansion

\[
\phi(\Delta(j+1)) = \phi(\Delta j) + \Delta \dot{\phi}(\Delta j) + O_j(\Delta)
\]

where \( O_j(\Delta) \) is a sum of higher power terms in \( \Delta \) with the property that

\[
\lim_{\Delta \to 0} \frac{O(\Delta)_j}{\Delta} = 0
\]

Since \( \phi \) satisfies

\[
\dot{\phi}(t) = f(\phi(t), t)
\]

we can also write

\[
\phi(\Delta(j+1)) = \phi(\Delta j) + \Delta f(\phi(\Delta j), \Delta j) + O_j(\Delta)
\]

This and (5.43) suggest that a reasonable way to try to approximate \( \phi(\Delta(j+1)) \) with \( y(j+1) \) might be to use the formula

\[
y(j+1) = y(j) + \Delta f(y(j), \Delta j)
\]

This equation can be recursively solved thereby generating a sequence \( y(0), y(1), y(2), \ldots \) which in some sense approximates \( \phi(0), \phi(1), \phi(2), \ldots \). This particular approximation is called Euler’s method. As a practical matter it is too rudimentary to work very well, except in very special situations.

There are very much better techniques based on more sophisticated approximations. Note that Euler’s Method really stems from the approximation

\[
\phi(\Delta + t) - \phi(t) \approx \Delta \dot{\phi}(t + \Delta)
\]

Subtracting this from (5.45) and multiplying through by \( \frac{\Delta}{2} \) yields

\[
\phi(t + \Delta) - \phi(t) \approx \frac{\Delta}{2} \{ \dot{\phi}(t + \Delta) \}
\]

Replacing \( \dot{\phi}(t) \) and \( \dot{\phi}(t + \Delta) \) by \( f(\phi(t), t) \) and \( f(\phi(t + \Delta), t + \Delta) \) thus provides

\[
\phi(t + \Delta) - \phi(t) \approx \frac{\Delta}{2} \{ f(\phi(t), t) + f(\phi(t + \Delta), t + \Delta) \}
\]

(5.46)

Using the Euler approximation

\[
\phi(t + \Delta) \approx \phi(t) + \Delta \dot{\phi}(t) = \phi(t) + \Delta f(\phi(t), t)
\]

one can write

\[
f[\phi(t + \Delta), t + \Delta] \approx f[\phi(t) + \Delta f(\phi(t), t), t + \Delta]
\]

Using this approximation in (5.46) we obtain

\[
\phi(t + \Delta) - \phi(t) \approx \frac{\Delta}{2} \{ f(\phi(t), t) + f[\phi(t) + \Delta f(\phi(t), t), t + \Delta] \}
\]

This is the basis for the recursion algorithm

\[
y(i+1) = y(i) + \frac{\Delta}{2} \{ f(y(i), \Delta i) + f[y(i) + \Delta f(y(i), \Delta i), t + \Delta] \}
\]
which is one form of the Runge-Kutta method. A more elaborate version of the Runge-Kutta method is described by the equation

\[ y(i+1) = y(i) + \frac{\Delta}{6} \left( z_1(i) + 2z_2(i) + 2z_3(i) + z_4(i) \right) \]

where

\[ z_1(i) = f[y(i), \Delta i] \quad z_3(i) = f[y(i) + \Delta z_2(i), \Delta(i + \frac{1}{2})] \]
\[ z_2(i) = f[y(i) + \Delta z_1(i), \Delta(i + \frac{1}{2})] \quad z_4(i) = f[y(i) + 2\Delta z_3(i), \Delta(i + 1)] \]

This particular version of the Runge-Kutta Method is widely used in many software packages including Matlab.

Much is known about recursive techniques such as the ones we’ve been discussing. Often a variable step size \( \Delta(i) \) is used to control the accumulation of errors. A course on numerical analysis would typically deal with this topic in depth. We will not pursue the subject further here.

### 5.7.2 Analog Simulation

An alternative way to deal with initial value problems is via “simulation” on an “analog computer.” Analog computers were widely for simulation years ago (especially in real-time applications) prior to the development of modern digital computers. Although still in use, analog computers are at present much less popular than digital computers for two reasons:

1. Analog computers are limited by the accuracy of their components (e.g., resistors and capacitors) and for this reason cannot usually match the precision obtainable with digital computers.
2. Because of the limited dynamic ranges of their components, electronic analog computers are much more difficult to program and to keep from saturating than their digital counterparts.

Despite these shortcomings analog computation is worth knowing about for several reasons:

1. Analog computers are inherently parallel machines. Partly for this reason, analog computers are often faster than their digital counterparts, all other things being equal.
2. The detailed study of analog vlsi chips is currently begin pursued and there is some reason to believe that analog techniques will re-emerge as the preferred approach for dealing with certain types of problems, especially those for which speed is of much greater concern than accuracy.
3. Perhaps the most important reason for understanding the analog approach is because it provides an intuitively appealing conceptual framework for visualizing how a computing machine might go about solving a complicated system of ordinary differential equations. As we shall see, this framework is especially useful even if the the machine is digital.

An analog computer is an electronic device composed primarily of precision resistors, precision capacitors, high gain linear (i.e., operational) amplifiers, and diodes. Solutions to differential equations are represented on an analog computer as time-dependent voltage signals. Interconnections of an analog computer’s basic components yields higher level analog building blocks called *summers*, *inverters*, *integrators*, *multipliers*, *gains* and *nonlinear function generators*. Each of these building blocks performs the function corresponding to its name. For example, an *integrator* acts on its input signal \( x(t) \) by generating an output signal \( y(t) \) which is the real-time integral of \( x(t) \). Analog computers are capable of performing simultaneously many different
time integrations, multiplications, summations, etc. In other words, an analog computer is an inherently parallel machine.

An analog computer is programmed by physically interconnecting the inputs and outputs of its higher level building blocks thereby creating an analog program. A simulation is initiated by turning on the integrators and letting the computer run. The time evolutions of signals are monitored using standard analog displays including chart recorders and scopes.

In the sequel we will explain how an analog computer is programmed. We will do this using the building block symbols found in Matlab’s Simulink Package since these are the symbols normally used in an analog computer programming diagram and since Simulink is programmed as if it were the front end of an analog computer. Of course Simulink is not truly the front end of an analog computer. Rather it is an especially useful graphical interface for programming a digital computer to solve ordinary differential equations. Once Simulink is programmed in an analog fashion, what the computer actually does is to translate the program into digital code and then execute it serially using standard numerical algorithms such as the Runge-Kutta method. From the users perspective, the computer appears to be doing an analog simulation even though this is not really what’s happening.

### 5.7.3 Programming Diagrams

To access Simulink within Matlab, one enters the command `simulink` in the Command Window. In response there appears a mouse-driven window called Simulink which looks like this.

![Simulink Block Library](image)

As a first step one would typically open a new window by selecting `new` under the `file` menu. Double clicking on any of the above windows, e.g., Sources opens it. Any item can then be dragged into the window `new` opened previously. Items in columns 1 through 4 below come from windows Sources, Sinks, Linear and Nonlinear respectively. The one exception to this is the Unit Delay in the lower right hand corner which comes from the Discrete window.
These are some of the most important building blocks in Simulink for solving ordinary differential equations. Each building block does just what you’d expect:

**Signal Generator:** Generates a time signal whose specifics are set by opening the Signal Generator window.

**Sine Wave:** Generates a sine-wave time signal whose amplitude and frequency are set by opening the Sine-Wave window.

**Step Input:** Generates a step time signal whose height and time of occurrence are set by opening the Step Input window.

**Scope:** Plots its input signal(s) versus time.

**Graph:** Plots its input signals(s) on one graph versus time.

**XY Graph:** Plots one input signal on a graph versus the other - good for phase plane trajectory plots.

**Sum:** Generates an output time signal which is the sum of its input signals, each input being weighted by a specified +1 or −1; handles more than two inputs.

**Integrator:** Generates an output signal which is the time integral of its input plus a specified initial condition constant.

**Gain:** Generates an output signal which is its input signal times a specified gain constant.

**Product:** Generates an output signal which is the product of its input signals.

**Fcn:** Generates an output signal which is any specified function of its input signal.

**Unit Delay:** Generates an output signal which is the input signal delayed by a prespecified amount of time; used in place of an integrator for a discrete time simulation.

Arrows connecting building blocks can be drawn by holding down the mouse button and dragging within the new window.

**Example 42** The differential equation

\[
\ddot{y} + .5\dot{y} + y - \frac{4.5}{y^2} = \dot{u} + u
\]

can be represented by the state space equations

\[
\begin{align*}
\dot{x}_1 &= -.5x_1 + x_2 + u \\
\dot{x}_2 &= -x_1 + \frac{4.5}{x_1^2} + u
\end{align*}
\]

where

\[y = x_1\]

Note, by the way, that \(\dot{u}\) does not appear in the state space model, even though it appears in the original differential equation. This would be especially important if \(u\) were a discontinuous input like a step. Figure 42 shows a Simulink simulation diagram for this system with \(x_1\) and \(x_2\) being the signals coming out of integrators 1 and 2 respectively. Note that \(-.5x_1 + x_2 + u\) is the the sum of the signals going into integrator 1 and \(-x_1 + \frac{4.5}{x_1^2} + u\) is the corresponding sum going into integrator 2. As a rule, simulation diagrams for state space models can always be set up with the outputs of the integrators being state variables. Thus an \(n\)-dimensional state space model requires \(n\)-integrators for simulation.
Below is a graph of $y$ for a simulation in which $x_1(0) = .5$, $x_2(0) = 0$, and $u$ is five times a unit step applied at $t = 15$. 
Chapter 6

Linear Differential Equations

By an \( n \)-dimensional linear state differential equation is meant an equation of the form

\[
\dot{x} = A(t)x + b(t)
\]

(6.1)

where \( A(t) \) is a \( n \times n \) matrix and \( x \) and \( b(t) \) are \( n \) - vectors. If \( A(t) \) is constant, (6.1) is constant coefficient; if \( A(t) \) is constant and \( t \) is time, we sometimes say that (6.1) is time invariant or stationary or autonomous. The vector \( b(t) \) is often called a forcing function; in the event that it is the zero function, (6.1) is called unforced or homogeneous.

We will assume throughout that \( A(t) \) and \( b(t) \) are at least piecewise continuous functions. Then \( f(x, t) \triangleq A(t)x + b(t) \) is globally Lipschitz in \( x \) and piecewise continuous is \( t \). In view of our earlier findings, this means that for each initial time \( t_0 \in [0, \infty) \) and initial state \( x_0 \in \mathbb{R}^n \), there is exactly one solution \( \phi \) to (6.1) on \( [0, \infty) \) which passes through state \( x_0 \) at \( t = t_0 \). Sometimes we will simply write \( x \) instead of \( \phi \).

The aim of this chapter is to study the basic properties of such equations. We begin by explaining one of the most common ways in which linear differential equations arise.

6.1 Linearization

More often than not when modeling a physical process, one encounters ordinary differential equations which are not linear and which have “inputs”. Suppose that

\[
\dot{x} = f(x, t, u)
\]

(6.2)

is such an equation where \( f : \mathbb{R}^n \times [0, t) \times \mathbb{R}^m \to \mathbb{R}^n \) is a continuously differentiable function of \( x \) and a piecewise continuous function of \( t \) and a continuously differentiable function of \( u \). Suppose in addition that for some given initial time \( t_0 \) and initial state \( x_0 \) and piecewise continuous function \( v : [0, \infty) \to \mathbb{R}^m \) called an input, that \( \phi \) is the solution to (6.2) \{assuming \( u \triangleq v \}\}, which passes through \( x_0 \) at \( t = t_0 \). In many practical situations it is important to have a good description of how a this solution might change in response to a small change in \( x_0 \) or a small change in the input function \( u \). If the presumed changes are small enough, then their effect can be studied by using a linear model derived from (6.2). The process of constructing this model is called linearization and is as follows.

Let us suppose that \( y_0 \in \mathbb{R}^n \) is any vector with a “small” norm and that \( w : [0, \infty) \to \mathbb{R}^m \) is any piecewise continuous function whose norm \( \|w(t)\| \) is also “small” for each fixed \( t \in [0, \infty) \). Let \( \theta \) denote a
solution to (6.2) \{assuming \( u \hat{=} v + w \)\} which passes through \( x_0 + y_0 \) at \( t = t_0 \). What we want to do is to derive a linear differential equation whose solution approximates the difference \( \theta - \phi \). Now what we know is that

\[
\begin{align*}
\dot{\phi}(t) &= f(\phi(t), t, v(t)) \\
\phi(t_0) &= x_0 \\
\dot{\theta}(t) &= f(\theta(t), t, v(t) + w(t)) \\
\theta(t_0) &= x_0 + y_0
\end{align*}
\] (6.3) (6.4) (6.5) (6.6)

If for each fixed \( t \) we now expand the right side of (6.5) in a Taylor Series about the point \((\phi(t), v(t))\) we get

\[
f(\theta(t), t, v(t) + w(t)) = f(\phi(t), t, v(t)) + A(t)\{\theta(t) - \phi(t)\} + B(t)w(t) + \text{higher order terms in } \{\theta(t) - \phi(t)\} \text{ and } w(t)
\]

where

\[
A(t) \triangleq \left. \frac{\partial f(x, t, u)}{\partial x} \right|_{(x, t, u) = (\phi(t), t, v(t))} \quad \text{and} \quad B(t) \triangleq \left. \frac{\partial f(x, t, u)}{\partial u} \right|_{(x, t, u) = (\phi(t), t, v(t))}
\]

This and (6.3) imply that

\[
\frac{d}{dt}\{\theta(t) - \phi(t)\} = A(t)\{\theta(t) - \phi(t)\} + B(t)w(t) + \text{higher order terms in } \{\theta(t) - \phi(t)\} \text{ and } w(t)
\]

This suggests that by solving the linear differential equation

\[
\dot{y} = A(t)y + B(t)w(t)
\] (6.7)

with initial state \( y_0 \) at time \( t = t_0 \), one should get a solution \( \psi(t) \) which to first approximation satisfies

\[
\psi \approx \theta - \phi
\]

The point here is that (6.7) is a linear differential equation and consequently an equation amenable to analysis. We call (6.7) the linearization of (6.3) about the nominal solution \( \phi \) determined by nominal initial condition \( x_0 \) and nominal input \( v(t) \).

One of the most common applications of the linearization process arises in situations when \( f(x, t, u) \) does not depend on \( t \) \{i.e., \( f(x, t, u) = f(x, u) \)\} and when the nominal input \( v(t) \) and nominal solution \( \phi(t) \) are both constant. Note that for \( v(t) \) a constant \( \bar{v} \), \( \phi(t) \) will be the constant solution \( \phi(t) = x_0 \) provided

\[
\bar{f}(x_0, \bar{v}) = 0
\]

Note also that under these conditions both \( A(t) \) and \( B(t) \) will be constant matrices.

Example 43 Consider a satellite of mass \( m \) in orbit about the earth as shown in the following diagram.
Assume that the satellite can be thrusted in the radial direction with thrust $u_1$ and in the tangential direction with thrust $u_2$. In polar coordinates the satellite’s equations of motion are

\begin{align*}
\ddot{r} &= \dot{\theta}^2 r - \frac{k}{r^2} + u_1 \\
\ddot{\theta} &= -2\frac{\dot{r}\dot{\theta}}{r} + \frac{u_2}{r}
\end{align*}

where $\dot{\theta}^2 r$ and $\frac{k}{r^2}$ model the effects of centrifugal force and gravity respectively. To develop an equivalent state space model, define

\begin{align*}
x_1 &\triangleq r, & x_2 &\triangleq \dot{r}, & x_3 &\triangleq \theta, & x_4 &\triangleq \dot{\theta}
\end{align*}

Then

\begin{align*}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= x_2^2 x_1 - \frac{k}{x_1^2} + u_1 \\
\dot{x}_3 &= x_4 \\
\dot{x}_4 &= -2\frac{x_2 x_4}{x_1} + \frac{u_2}{x_1}
\end{align*}

and

\begin{align*}
r &= x_1 \\
\theta &= x_3
\end{align*}

Let us note that for $u_1 = u_2 = 0$, these equations admit solutions of the form

\begin{align*}
x_1 &= r_0, & x_2 &= 0, & x_3 &= \omega_0 t, & x_4 &= \omega_0
\end{align*}

where $r_0$ and $\omega_0$ are positive constants satisfying $\omega_0^2 r_0^3 = k$. Any such solution describes a circular orbit of radius $r_0$ with angular velocity $\omega_0$. Our aim is to linearize the above equations about such an orbit. In other words we want to develop a system of linear differential equations for deviation variables $z_1$, $z_2$, $z_3$, $z_4$ such that any solution $x_1$, $x_2$, $x_3$, $x_4$ to the above equations satisfies

\begin{align*}
x_1 \approx r_0 + z_1, & & x_2 \approx z_2, & & x_3 \approx \omega_0 t + z_3, & & x_4 \approx \omega_0 + z_4
\end{align*}
assuming \( u_1 \) and \( u_2 \) are small thrust signals and the initial values of \( x_1 \), \( x_2 \), \( x_3 \), and \( x_4 \) are close to \( r_0 \), 0, \( \omega_0 t \) and \( \omega_0 \) respectively. Carrying out the linearization one gets the equations

\[
\begin{align*}
\dot{z}_1 &= z_2 \\
\dot{z}_2 &= \left( \frac{\omega_0^2}{r_0^3} + \frac{3k}{r_0} \right) z_1 + 2\omega_0 r_0 z_4 + u_1 \\
\dot{z}_3 &= z_4 \\
\dot{z}_4 &= -2\frac{\omega_0}{r_0} z_2 + \frac{1}{r_0} u_2
\end{align*}
\]

which can be written more concisely as

\[
\dot{z} = Az + Bu
\]

where

\[
z \triangleq \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix}, \quad A \triangleq \begin{bmatrix} 0 & 0 & 0 & 0 \\ \omega_0^2/3r_0 & 0 & 0 & 2\omega_0 r_0 \\ 0 & 0 & 0 & 1 \\ 0 & -2\omega_0/r_0 & 0 & 0 \end{bmatrix}, \quad B \triangleq \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1/r_0 \end{bmatrix}, \quad u \triangleq \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}
\]

Note that for this example, the linearized equations are time-invariant even though the solution about which the linearization is made is not.

---

### 6.2 Linear Homogenous Differential Equations

In this section we shall investigate the fundamental properties of the linear, homogeneous, \( n \)-dimensional differential equation

\[
\dot{x} = A(t)x
\]  

(6.8)

where \( A(t) \) is a \( n \times n \) matrix of piecewise continuous time functions defined on the interval \([0, \infty)\). Let us note right away that for each finite interval \([t_0, t_1] \subset [0, \infty)\) we can write a Lipschitz inequality of the form

\[
\|A(t)x - A(t)y\| \leq \lambda\|x - y\|, \quad x, y \in \mathbb{R}^n
\]

where

\[
\lambda \triangleq \sup_{t \in [t_0, t_1]} \|A(t)\| < \infty
\]

Thus solutions to (6.8) exist globally and are unique.

**Theorem 5** Suppose that \( A : [0, \infty) \to \mathbb{R}^{n \times n} \) is piecewise continuous. For each \( x_0 \in \mathbb{R}^n \) and each \( t_0 \in [0, \infty) \) there exists exactly one continuous function \( \phi \) which satisfies (6.8) on \([0, \infty)\) and passes through the state \( x_0 \) at \( t = t_0 \).
6.2.1 State Transition Matrix

In the sequel we shall write $\zeta(t, \tau, z)$ for the unique solution to (6.8) which passes through state $z$ at $t = \tau$. Thus $\zeta: [0, \infty) \times [0, \infty) \times \mathbb{R}^n \to \mathbb{R}^n$ is a continuous function uniquely defined by the equations

\[
\begin{align*}
\dot{\zeta}(t, \tau, z) &= A(t)\zeta(t, \tau, z) \\
\zeta(\tau, \tau, z) &= z
\end{align*}
\quad \forall t, \tau \in [0, \infty), \forall z \in \mathbb{R}^n \tag{6.9}
\]

Let $e_i, i \in \{1, 2, \ldots, n\}$ denote the $i$th unit vector in $\mathbb{R}^n$. The state transition matrix \(^{2}\) of $A(t)$ is the $n \times n$ matrix

\[
\Phi(t, \tau) \triangleq \begin{bmatrix} \zeta(t, \tau, e_1) & \zeta(t, \tau, e_2) & \ldots & \zeta(t, \tau, e_n) \end{bmatrix}
\tag{6.10}
\]

The definition implies that

\[
\Phi(\tau, \tau) = \begin{bmatrix} e_1 & e_2 & \ldots & e_n \end{bmatrix} = I_{n \times n}
\]

and also that

\[
\begin{align*}
\dot{\Phi}(t, \tau) &= \begin{bmatrix} \dot{\zeta}(t, \tau, e_1) & \dot{\zeta}(t, \tau, e_2) & \ldots & \dot{\zeta}(t, \tau, e_n) \end{bmatrix} \\
&= \begin{bmatrix} A(t)\zeta(t, \tau, e_1) & A(t)\zeta(t, \tau, e_2) & \ldots & A(t)\zeta(t, \tau, e_n) \end{bmatrix} \\
&= A(t)\Phi(t, \tau)
\end{align*}
\]

In other words, the definition of $\Phi(t, \tau)$ in (6.10) implies that

\[
\begin{align*}
\dot{\Phi}(t, \tau) &= A(t)\Phi(t, \tau) \\
\Phi(\tau, \tau) &= I_{n \times n}
\end{align*}
\quad \forall t, \tau \in [0, \infty) \tag{6.11}
\]

But because of uniqueness of solutions to linear differential equations, there can be only one matrix possessing these two properties. Thus (6.11) also implies (6.10). Because of this, (6.11) an be taken as an alterative defninition for the state transition matrix of $A(t)$. It is entirely equivalent to the original definition of $\Phi(t, \tau)$ given by (6.10).

The importance of the state transition matrix stems from the fact that

\[
\zeta(t, \tau, z) = \Phi(t, \tau)z, \quad \forall t, \tau \in [0, \infty), \forall z \in \mathbb{R}^n \tag{6.12}
\]

To verify that this is so, fix $\tau$ and $z$ and set $\phi(t) \triangleq \Phi(t, \tau)z$; then because of (6.11), $\phi$ must satisfy (6.8) and $\phi(\tau) = z$. Since these equations are satisfied by just one continuous function, namely $\zeta(t, \tau, z)$, it must be that (6.12) is true.

Let us note that if the dynamical system (6.8) is in state $x_1$ at time $t_1$ and state $x_2$ at time $t_2$ then the “transition” from $x_1$ to $x_2$ is described by the equation

\[x_2 = \Phi(t_2, t_1)x_1\]

\(^{1}\)Here as always, “." denotes differentiation with respect to $t$.

\(^{2}\)In most mathematics texts, the state transition matrix of $A(t)$ is called the fundamental matrix of $A(t)$. The name “state transition matrix” is commonly used in many fields. We’ve adopted it here because state transition is a more descriptive modifier than fundamental.
Hence the name “state transition matrix.”

Since \( \Phi(t, \tau) \) is a continuous function of \( t \), (6.11) is equivalent to the integral equation

\[
\Phi(t, \tau) = I + \int_{\tau}^{t} A(\mu) \Phi(\mu, \tau) d\mu \quad \forall t, \tau \in [0, \infty)
\]  (6.13)

This equation can be solved by Picard Iteration. In particular the sequence of matrices \( \Theta_0(t, \tau) \), \( \Theta_1(t, \tau) \), \( \Theta_2(t, \tau) \), \ldots, defined for all \( t, \tau \in [0, \infty) \) by

\[
\Theta_0(t, \tau) \triangleq I \quad \text{and} \quad \Theta_{i+1}(t, \tau) = I + \int_{\tau}^{t} A(\mu) \Theta_i(\mu, \tau) d\mu, \quad i \in \{0, 1, \ldots\},
\]  (6.14)

converges uniformly on every finite interval \([0, t_1]\) and in the limit

\[
\Phi(t, \tau) = \lim_{i \to \infty} \Theta_i(t, \tau) \quad \forall t, \tau \in [0, \infty) \]  (6.15)

6.2.2 The Matrix Exponential

It is especially useful at this point to consider briefly what the \( \Theta_i(t, \tau) \) look like in the special case when \( A(t) \) is a constant matrix \( B \). If this is so, then because of (6.14)

\[
\Theta_1(t, \tau) = I + \int_{\tau}^{t} B d\mu = I + (t - \tau)B
\]

By similar reasoning

\[
\Theta_2(t, \tau) = I + \int_{\tau}^{t} (I + \mu B) d\mu = I + (t - \tau)B + \frac{1}{2}((t - \tau)B)^2
\]

Continuing this process, one eventually arrives at the formula

\[
\Theta_i(t, \tau) = \sum_{j=0}^{i} \frac{1}{j!}((t - \tau)B)^j
\]

In view of (6.15), the series

\[
\sum_{j=0}^{i} \frac{1}{j!}((t - \tau)B)^j
\]

converges to \( \Phi(t, \tau) \) as \( i \to \infty \) for every finite value of \( t \) and \( \tau \) in \([0, \infty)\). In other words

\[
\Phi(t, \tau) = e^{(t-\tau)B}, \quad t, \tau \in [0, \infty)
\]  (6.16)

where for any \( n \times n \) matrix \( M \), the symbol \( e^M \) stands for the the \( n \times n \) matrix exponential

\[
e^M \triangleq \sum_{j=0}^{\infty} \frac{1}{j!}M^j
\]  (6.17)

Note that this series must converge because, as we already know, the series

\[
\sum_{j=0}^{\infty} \frac{1}{j!}((t - \tau)B)^j
\]
converges to \( \Phi(t, \tau) \) for any constant matrix \( B \), and any finite values of \( t \) and \( \tau \) in \([0, \infty)\) - to complete the proof just take \( B = M \) and \( t - \tau = 1 \).

From the fact that \( \Phi(t, \tau) \) satisfies (6.11) we arrive at the formulas

\[
\begin{align*}
\frac{d}{dt} \{ e^{tB} \} & = B e^{tB} \\
\left. e^{tB} \right|_{t=0} & = I_{n \times n}
\end{align*}
\] \quad \forall t \in [0, \infty)
\]

In these ways, the matrix exponential behaves much like a scalar exponential.

A little latter we will explain how to explicitly calculate \( e^{(t-\tau)B} \). For the present, we conclude this excursion into the study of the case when \( A(t) \) is a constant, with a few words of caution: The expression for \( \Phi(t, \tau) \) in (6.16) holds only if \( A(t) \) is a constant matrix \( B \). If \( A(t) \) is nonconstant, it is not even true that

\[
\Phi(t, \tau) = e^{\int_{\tau}^{t} A(\mu)d\mu}
\]

except in the very special case when \( A(t) \left( \int_{\tau}^{t} A(\mu)d\mu \right) = \left( \int_{\tau}^{t} A(\mu)d\mu \right) A(t), \ \forall t, \tau \geq 0. \)

### 6.2.3 State Transition Matrix Properties

We now return to the general case when \( A(t) \) is any \( n \times n \) piecewise continuous matrix. Our aim in this section is to derive several basis properties of \( A \)'s state transition matrix \( \Phi(t, \tau) \).

#### Exponential Growth for Bounded \( A(t) \)

To begin let us note that (6.13) enables us to write for \( t \geq \tau \geq 0 \)

\[
||\Phi(t, \tau)|| = \left\| I + \int_{\tau}^{t} A(\mu)\Phi(\mu, \tau)d\mu \right\| \leq ||I|| + \int_{\tau}^{t} ||A(\mu)||\Phi(\mu, \tau)||d\mu
\]

Thus

\[
||\Phi(t, \tau)|| \leq ||I|| + \int_{\tau}^{t} ||A(\mu)||\Phi(\mu, \tau)||d\mu
\]

By the Bellman-Gronwall Lemma,

\[
||\Phi(t, \tau)|| \leq ||I||e^{\int_{\tau}^{t} ||A(\mu)||d\mu}
\] \quad (6.19)

Suppose \( A(t) \) is bounded in norm on \([0, \infty)\); i.e. suppose

\[
\lambda \Delta \sup_{t \in [0, \infty)} ||A(t)|| < \infty
\]

Then (6.19) simplifies to

\[
||\Phi(t, \tau)|| \leq ||I||e^{(t-\tau)\lambda}
\]

Similar reason leads to the inequality

\[
||\zeta(t, \tau, z)|| \leq ||z||e^{(t-\tau)\lambda}
\]

where \( \zeta(t, \tau, z) \) is the solution to (6.8) passing through state \( z \) at \( t = \tau \). These inequalities show that if \( A(t) \) is bounded on \([0, \infty)\), then the norms of \( \Phi(t, \tau) \) and all solutions to (6.8) can grow at most exponentially fast as \( t \to \infty \).
Composition Rule and Consequences

As noted in the last chapter, the composition rule for a dynamical system \( \dot{x} = f(x, t) \) is
\[
\zeta(t, \tau, z) = \zeta(t, \mu, \zeta(\mu, \tau, z)), \quad \forall z \in \mathbb{R}^n, \quad \forall t, \tau, \mu \in [0, \infty)
\]
where, as usual, \( \zeta(t, \tau, z) \) denotes the the solution to \( \dot{x} = f(x, t) \) which passes through \( z \) at \( t = \tau \). For the linear differential equation \( \dot{x} = A(t)x \), we know because of (6.12) that
\[
\zeta(t, \tau, z) = \Phi(t, \tau)z, \quad \zeta(t, \mu, \zeta(\mu, \tau, z)) = \Phi(t, \mu)\zeta(\mu, \tau, z), \quad \text{and} \quad \zeta(\mu, \tau, z) = \Phi(\mu, \tau)z
\]
It thus follows that
\[
\Phi(t, \tau)z = \Phi(t, \mu)\Phi(\mu, \tau)z, \quad \forall z \in \mathbb{R}^n, \quad \forall t, \tau, \mu \in [0, \infty) \tag{6.20}
\]
From this it follows that
\[
\Phi(t, \tau) = \Phi(t, \mu)\Phi(\mu, \tau), \quad \forall t, \tau, \mu \in [0, \infty) \tag{6.21}
\]
This is the composition rule for state transition matrices\(^3\).

Since the composition rule holds for all values of \( t, \tau, \mu \in [0, \infty) \), there is nothing to prevent us from setting \( \tau = t \). This results in \( I = \Phi(t, \mu)\Phi(\mu, t) \) and thus the remarkable fact that
\[
\Phi^{-1}(t, \mu) = \Phi(\mu, t), \quad \forall t, \mu \in [0, \infty) \tag{6.22}
\]
In other words, the inverse of \( \Phi(t, \mu) \) is obtainable by simply interchanging its arguments.

Note that we have incidently proved that \( \Phi(t, \mu) \) is nonsingular for all finite values of \( t \) and \( \mu \). A little thought reveals that this is in essence a restatement of the fact that trajectories starting at the same time but in different states, cannot cross in finite time.

Properties of Matrix Exponentials

Let us again briefly return to the case when \( A(t) \) is a constant matrix \( B \). Recall that when this is so \( \Phi(t, \tau) \) is the matrix exponential
\[
\Phi(t, \tau) = e^{(t-\tau)B}
\]
In this case (6.21) and (6.22) simplify to
\[
e^{(t-\tau)B} = e^{(t-\mu)B}e^{(\mu-\tau)B}, \quad \forall t, \tau, \mu \in [0, \infty)
\]
and
\[
(e^{(t-\mu)B})^{-1} = e^{-(t-\mu)B}, \quad \forall t, \mu \in [0, \infty)
\]
respectively. The latter formula proves that for any \( n \times n \) matrix \( M \)
\[
(e^M)^{-1} = e^{-M} \tag{6.23}
\]
whereas the former shows that
\[
e^{(t_1+t_2)M} = e^{t_1M}e^{t_2M}, \quad \forall t_1, t_2 \in \mathbb{R} \tag{6.24}
\]
To this extent, matrix exponentials are just like scalar exponentials. Note however that if \( F \) and \( G \) are two \( n \times n \) matrices, then in general
\[
e^{(F+G)} \neq e^Fe^G
\]except in the special case when \( F \) and \( G \) commute.

\(^3\)In arriving at (6.21) from (6.20) we’ve used the algebraic fact that if \( C \) and \( D \) are two \( m \times n \) matrices such that \( Cy = Dy, \forall y \in \mathbb{R}^m \) then \( C = D \). This in turn can be proved by noting that the hypothesis \( Cy = Dy, \forall y \in \mathbb{R}^m \) implies the matrix equation \([Cy_1 \quad Cy_2 \quad \cdots \quad Cy_n] = [Dy_1 \quad Dy_2 \quad \cdots \quad Dy_n] \) for all choices of the \( y_i \in \mathbb{R}^n \). Picking \( y_i \) to be the \( i \)th unit vector in \( \mathbb{R}^n \) then yields the desired result.
In the sequel we shall derive an interesting formula for the determinant of \( \Phi(t, \tau) \) in terms of the integral of the trace\(^4\) of \( A(t) \). To do this we need first to develop a formula for the derivative with respect to \( t \) of the determinant of \( \Phi(t, \tau) \). In the sequel \( \phi_{ij} \) denotes the \( ij \)th entry of \( \Phi \).

To begin we note that since \( \det \Phi \) is a function of its \( n^2 \) entries \( \phi_{ij} \),

\[
\frac{d}{dt} \{ \det \Phi(t, \tau) \} = \sum_{i=1}^{n} \sum_{j=1}^{n} \left( \frac{\partial \{ \det \Phi \} }{\partial \phi_{ij}} \right) \dot{\phi}_{ij} \quad (6.25)
\]

Next recall that \( \det \Phi \) can be written as

\[
\det \Phi = \sum_{i=1}^{n} \tilde{\phi}_{ij} \phi_{ij}
\]

where \( \tilde{\phi} \) is the cofactor of \( \phi_{ij} \). Since \( \tilde{\phi}_{ij} \) does not depend on \( \phi_{ij} \) it must be true that

\[
\frac{\partial \det \Phi}{\partial \phi_{ij}} = \tilde{\phi}_{ij}
\]

From this and (6.25) it follows that

\[
\frac{d}{dt} \{ \det \Phi(t, \tau) \} = \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{\phi}_{ij} \dot{\phi}_{ij} \quad (6.26)
\]

Let \( C \) denote the cofactor matrix \( C = [ \tilde{\phi}_{ij} ]_{n \times n} \). A simple calculation shows that

\[
\sum_{i=1}^{n} \tilde{\phi}_{ij} \dot{\phi}_{ij}
\]

is \( j \)th diagonal element of the product \( C' \dot{\Phi} \). Therefore

\[
\text{trace} \ C' \dot{\Phi} = \sum_{j=1}^{n} \sum_{i=1}^{n} \tilde{\phi}_{ij} \dot{\phi}_{ij}
\]

From this and (6.26) it follows that

\[
\frac{d}{dt} \{ \det \Phi \} = \text{trace} \ C' \dot{\Phi}
\]

Using the relation \( \dot{\Phi} = A \Phi \) we therefore have

\[
\frac{d}{dt} \{ \det \Phi \} = \text{trace} \ \{ C' A \Phi \}
\]

Thus\(^5\)

\[
\frac{d}{dt} \{ \det \Phi \} = \text{trace} \ \{ A \Phi C' \} \quad (6.27)
\]

To proceed, recall that the inversion formula for \( \Phi \) is

\[
\Phi^{-1} = \frac{1}{\det \Phi} C'
\]

\(^4\)The trace of a square matrix is the sum of its diagonal elements; i.e. \( \text{trace} \ [q_{ij}]_{n \times n} = \sum_{i=1}^{n} q_{ii} \).

\(^5\)Here we are using the identity \( \text{trace} \ \{ MN \} = \text{trace} \ \{ NM \} \) which holds for any pair of \( n \times n \) matrices \( M \) and \( N \).
Hence $\Phi' = (\det \Phi) I$. Substitution into (6.27) thus yields the one-dimensional linear differential equation
\[
\frac{d}{dt} \{ \det \Phi(t, \tau) \} = \text{trace} \{ A(t) \} \{ \det \Phi(t, \tau) \}
\]
Therefore
\[
\det \Phi(t, \tau) = \{ \det \Phi(\tau, \tau) \} e^{\int_{\tau}^{t} \text{trace} A(\mu) d\mu}
\]
Since $\Phi(\tau, \tau) = I$, we arrive at the expression
\[
\det \Phi(t, \tau) = e^{\int_{\tau}^{t} \text{trace} A(\mu) d\mu}
\]
which is sometimes called the *Abel, Jacobi, Liouville Formula*.

### 6.3 Forcéd Linear Differential Equations

We now turn to the forced linear differential equation
\[
\dot{x} + A(t)x + b(t)
\]  
(6.28)

Our main objective is to develop a formula which expresses this equation’s solution in terms of $b$ and the value of $x$ at some specified time $t_0$. In doing this, we will make use of the state transition matrix of $A(t)$ defined in the last section.

#### 6.3.1 Variation of Constants Formula

Let us begin by supposing that $x$ satisfies (6.28) and that $\Phi(t, \tau)$ is the state transition matrix of $A(t)$. Then as noted previously
\[
\begin{align*}
\dot{\Phi}(t, \tau) & = A(t)\Phi(t, \tau) \\
\Phi(\tau, \tau) & = I_{n \times n} \\
\end{align*}
\]  
(6.29)

What we want to do next is to make a change of variable which replaces $x$ by a new state vector $y$ whose differential equation is easier to solve than (6.28). With the benefit of considerable hindsight we shall define $y$ to be
\[
y(t) \triangleq \Phi(t_0, t)x(t)
\]  
(6.30)

where $t_0$ is a fixed number in $[0, \infty)$ The definition implies that
\[
y(t_0) = x(t_0)
\]  
(6.31)

because of (6.29) and that
\[
x(t) = \Phi(t, t_0)y(t)
\]  
(6.32)

because of (6.22). Differentiating both sides of (6.32) we get
\[
\dot{x} = \dot{\Phi}(t, t_0)y + \Phi(t, t_0)\dot{y}
\]
From (6.29) it thus follows that
\[
\dot{x} = A(t)\Phi(t, t_0)y + \Phi(t, t_0)\dot{y}
\]  
(6.33)
Meanwhile from (6.28) and (6.32) we have that
\[ \dot{x} = A(t)\Phi(t, t_0)y + b(t) \]
Substituting this into (6.33) to eliminate \( \dot{x} \) we get
\[ A(t)\Phi(t, t_0)y + b(t) = A(t)\Phi(t, t_0)y + \Phi(t, t_0)\dot{y} \]
After cancellation and rearrangement of terms this simplifies to
\[ \Phi(t, t_0)\dot{y} = b(t) \]
In view of (6.22) we can therefore write
\[ \dot{y} = \Phi(t_0, t)b(t) \]
Therefore
\[ y(t) = y(t_0) + \int_{t_0}^{t} \Phi(t_0, \mu)b(\mu)d\mu, \quad t \in [0, \infty) \]
Multiplying both sides of this equation by \( \Phi(t, t_0) \) gives
\[ \Phi(t, t_0)y(t) = \Phi(t, t_0)y(t_0) + \int_{t_0}^{t} \Phi(t, t_0)\Phi(t_0, \mu)b(\mu)d\mu, \quad t \in [0, \infty) \]
Using (6.32), (6.31) and the composition rule \( \Phi(t, t_0)\Phi(t_0, \mu) = \Phi(t, \mu) \) we arrive at the equation
\[
\boxed{x(t) = \Phi(t, t_0)x(t_0) + \int_{t_0}^{t} \Phi(t, \mu)b(\mu)d\mu, \quad t \in [0, \infty) \quad (6.34)}
\]
which is known as the variation of constants formula.

In the special case when \( A(t) \) is a constant matrix \( B \), \( \Phi(t, \tau) \) is the matrix exponential
\[ \Phi(t, \tau) = e^{(t-\tau)B} \]
Thus in the case, the variation of constants formula becomes
\[
\boxed{x(t) = e^{(t-t_0)B}x(t_0) + \int_{t_0}^{t} e^{(t-\mu)B}b(\mu)d\mu, \quad t \in [0, \infty) \quad (6.35)}
\]
It is interesting to note that the integral in this expression is of the form
\[ \int W(t - \mu)h(\mu)d\mu \]
Such integral forms arise in many contexts. They are called convolution integrals.

### 6.4 Periodic Homegeneous Differential Equations

In the light of the discussions in the preceding sections it is clear that the state transition matrix of \( A(t) \) plays a central role in characterizing solutions to the differential equation \( \dot{x} = A(t)x + b(t) \). There remains the problem of determining just what \( \Phi(t, \tau) \) is for a given \( A(t) \). Unfortunately for general \( A \)’s it is more
or less impossible to derive explicit analytical expressions for $\Phi(t, \tau)$ in terms of (say) elementary functions. Even differential equations as simple looking as the famous Mathieu Equation \{Here $\epsilon$ is a positive number.\}

$$\ddot{y} + (1 + 2\epsilon \cos 2t)y = 0$$ \hspace{1cm} (6.36)

defy closed form solutions. In fact, constant $A$'s are really the only general types of matrices for which closed-form expressions for state transition matrices can actually be worked out.

Between general $A$'s and the class of constant $A$'s lies the intermediate class of $A$'s which are periodic; i.e., $A$'s satisfying

$$A(t + T) = A(t), \quad t \geq 0$$ \hspace{1cm} (6.37)

where $T$ is a fixed nonnegative number. While it is still not possible to give a complete closed-form description for the state transition matrix of such an $A$, it is possible to go further than in the general case when $A$ is an arbitrary nonconstant matrix. The aim of this section is to explain just how much further.

### 6.4.1 Coordinate Transformations

Consider the differential equation

$$\dot{x} = A(t)x$$ \hspace{1cm} (6.38)

where $A(t)$ is a continuous $n \times n$ matrix. For the moment, suppose $A$ is otherwise arbitrary; i.e., not necessarily periodic. Let $Q(t)$ be a $n \times n$ matrix which is nonsingular and continuously differentiable on $[0, \infty)$. Suppose we define

$$z \triangleq Q(t)x$$ \hspace{1cm} (6.39)

Our aim is to derive a linear homogeneous differential equation for $z$. For this, let us note that

$$\dot{z} = \dot{Q}(t)x + Q(t)\dot{x} = \dot{Q}(t)x + Q(t)A(t)x$$

In view of (6.39), it must be true that

$$\dot{z} = (\dot{Q}(t)Q^{-1}(t) + Q(t)A(t)Q^{-1}(t))z$$ \hspace{1cm} (6.40)

Thus the change of variables

$$x \mapsto Q(t)x$$

induces a corresponding change in $A(t)$, namely

$$A(t) \mapsto \dot{Q}(t)Q^{-1}(t) + Q(t)A(t)Q^{-1}(t)$$

{Note, by the way, that if $Q$ were constant, $\dot{Q}$ would be zero and the preceding would be a similarity transformation.}

Now if we’re lucky, $\dot{Q}(t)Q^{-1}(t) + Q(t)A(t)Q^{-1}(t)$ might turn to be a constant matrix $B$ in which case (6.40) would be the constant coefficient differential equation

$$\dot{z} = Bz$$ \hspace{1cm} (6.41)

Actually, in this instance we can control our own luck and even more: we can always find a matrix $Q(t)$ which is nonsingular on $[0, \infty)$ such that

$$\dot{Q}(t)Q^{-1}(t) + Q(t)A(t)Q^{-1}(t) = B$$ \hspace{1cm} (6.42)
In fact, we can find such a $Q(t)$ for any given $n \times n$ matrix $B$. All we need to do is to solve the linear matrix differential equation
\begin{equation}
\dot{Q} + QA(t) = BQ \tag{6.43}
\end{equation}
with the initial value
\begin{equation}
Q(0) = I \tag{6.44}
\end{equation}
For if (6.43) holds and $Q(t)$ is nonsingular on $[0, \infty)$, then (6.42) clearly must hold as well. Therefore to justify the claim, all we need do is show that the solution to (6.43) and (6.44) is nonsingular on $[0, \infty)$. What we will actually show, is something a little stronger, namely that the solution to (6.43) and (6.44) is
\begin{equation}
Q(t) = e^{tB} \Phi^{-1}(t, 0) \tag{6.45}
\end{equation}
where $\Phi(t, \tau)$ is the state transition matrix of $A(t)$.

To justify (6.45), define
\begin{equation}
P(t) \triangleq e^{tB} \Phi^{-1}(t, 0)
\end{equation}
Then
\begin{equation}
P(0) = I \tag{6.46}
\end{equation}
and
\begin{equation}
P(t)\Phi(t, 0) = e^{tB}
\end{equation}
Thus
\begin{equation}
\dot{P}\Phi + P\dot{\Phi} = Be^{tB}
\end{equation}
Since $\dot{\Phi} = A\Phi$, it must be true that
\begin{equation}
\dot{P}\Phi + PA\Phi = Be^{tB}
\end{equation}
But $\Phi(t, 0)$ is nonsingular for all finite $t$ so
\begin{equation}
\dot{P} + P A = Be^{tB}\Phi(t, 0)^{-1} = BP
\end{equation}
Since this is the same differential equation as (6.43) and since $Q(0)$ and $P(0)$ are both equal the identity matrix, by uniqueness of solutions to linear differential equations, $Q$ and $P$ must be equal for all $t$. Thus (6.45) is true and so $Q(t)$ is nonsingular on $[0, \infty)$.

6.4.2 Lyapunov Transformations

The preceding shows that we can transform any time varying homogeneous linear differential equation into an essentially arbitrary time invariant one by suitable change of variables. In view of the arbitrariness of $B$, such a transformation may of course not be very illuminating. One class of transformations which leads to useful results, consists of those $Q$’s which preserve “stability.” Such $Q$’s are called “Lyapunov” transformations. More precisely, $Q$ is a Lyapunov Transformation if
\begin{enumerate}
\item $Q$ is continuously differentiable on $[0, \infty)$
\item $Q$ and $\dot{Q}$ are bounded on $[0, \infty)$
\item there exists a positive number $\mu$ such that $|\det Q(t)| \geq \mu$ for $t \in [0, \infty)$
\end{enumerate}

By defining a state vector consisting of the columns of $Q$ stacked one atop the next, one can rewrite this equation as a standard linear vector differential equation; existence and uniqueness theorems thus extend effortlessly to linear matrix differential equations such as this.
The last condition stipulates that the determinant of $Q$ must be bounded away from zero as $t \to \infty$. {Note, for example, that in the scalar case $Q \triangleq e^{-t}$ fails to have this property.} Together conditions 1 and 3 insure that $Q^{-1}(t)$ exists and is bounded on $[0, \infty)$.

Let $Q$ be a Lyapunov Transformation and suppose that $x$ and $z$ satisfy (6.38) and (6.40) respectively. It can then be shown that if all solutions to (6.38) are bounded on $[0, \infty)$ then so are all solutions to (6.40) - and conversely if all solutions to (6.40) are bounded on $[0, \infty)$ then so are all solutions to (6.38). It can also be shown that all solutions to (6.38) must tend to zero as $t \to \infty$ if and only if all solutions to (6.40) must tend to zero as $t \to \infty$. In other words, if $Q$ is a Lyapunov Transformation, then the stability properties of the differential equations in (6.38) and (6.40) are equivalent.

The definition of a Lyapunov Transformation $Q$ does not require $Q$ to satisfy (6.42). If $A$ is such that it is in fact possible to find a Lyapunov Transformation $Q$ for which (6.42) holds with $B$ a constant matrix, then is (6.38) is said to be reducible to a constant linear differential equation. A reducible differential equations is thus a differential equation whose stability properties can be decided {in theory at least} by examining the stability properties of a related constant coefficient differential {i.e., (6.40)}. In the next chapter we will explain how to determine the stability properties of the later.

Of course if one were to follow this program, one would still have to have methods for determining which linear differential equations are reducible. In general this is a difficult matter. There is however, one type of linear differential equation which is necessarily reducible. In particular we will now show that any linear differential equation with a periodic $A$ is reducible.

### 6.4.3 Floquet’s Theorem

We now invoke the assumption that $A(t)$ is a continuous matrix for which

$$A(t + T) = A(t), \quad t \in [0, \infty)$$

(6.47)

for some number $T \geq 0$. What we want is for $B$ to be defined so that

$$e^{BT} = \Phi(T, 0)$$

(6.48)

Since explaining just how to do this would take us beyond the scope of this course, we will limit are discussion of this point to the following remark.

**Remark 3** The problem of finding solution a solution $X_{n \times n}$ to the transcendental matrix equation

$$e^X = R_{n \times n}$$

(6.49)

is discussed on pages 239 - 241 of [1]. It turns out that all that’s required for such a solution to exist is that $R$ be nonsingular. However (6.49) may not have a real-valued solution even if $R$ is real. Consider for example the scalar case in which $R = -1$. In general solutions to (6.49) are not unique. On the other hand, for given $R$ all solutions have certain properties in common and these are precisely the properties needed to characterize the the stability of $\dot{z} = Bz$.

What we want to do now is to show that the matrix $Q$ defined by (6.43) and (6.44), or equivalently by (6.45), is periodic. For ease of reference we rewrite (6.45) here:

$$Q(t) = e^{tB} \Phi^{-1}(t, 0)$$

(6.50)
To begin, let us note that
\[ Q(t + T) = e^{(t + T)B} \Phi^{-1}(t + T, 0) \]

By the composition rule
\[ \Phi(t + T, 0) = \Phi(t + T, T) \Phi(T, 0) \quad \text{and} \quad e^{(t + T)B} = e^{tB} e^{TB} \]

Thus
\[ Q(t + T) = e^{tB} e^{TB} \Phi^{-1}(T, 0) \Phi^{-1}(t + T, T) \]

so by (6.48)
\[ Q(t + T) = e^{tB} \Phi^{-1}(t + T) \quad (6.51) \]

Meanwhile,
\[ \dot{\Phi}(t, 0) = A(t) \Phi(t, 0) \]

and
\[ \dot{\Phi}(t + T, T) = A(t + T) \Phi(t + T, T) \]

In view of (6.47)
\[ \dot{\Phi}(t + T, T) = A(t) \Phi(t + T, T) \]

Thus \( \Phi(t, 0) \) and \( \Phi(t + T, T) \) both satisfy differential equation. Since \( \Phi(t, 0) \) and \( \Phi(t + T, T) \) also both equal the identity matrix at \( t = 0 \), by uniqueness of solutions to linear differential equations it must be that \( \Phi(t, 0) = \Phi(t + T, T) \) for all \( t \geq 0 \). From this, (6.51) and (6.50) it follows that
\[ Q(t + T) = e^{tB} \Phi^{-1}(t, 0) = Q(t), \quad t \in [0, \infty) \]

Therefore \( Q(t) \) is periodic.

The definition of \( Q \) in (6.50) and the assumed continuity of \( A \) imply that \( Q \) is continuously differentiable. This and the fact that \( Q \) is periodic, further imply that \( Q \) and \( \dot{Q} \) are bounded on \([0, \infty)\). Moreover, since \( e^{tB} \) and \( \Phi(t, 0) \) are nonsingular for finite \( t \), and since \( Q \) is periodic, it must be that the determinant of \( Q \) is bounded away from zero for all \( t \). In other words, \( Q \) is a Lyapunov Transformation and \( \dot{x} = A(t)x \) is a reducible differential equation.

It is possible to carry things a bit further. Note from (6.50) that
\[ \Phi(t, 0) = Q^{-1}(t)e^{tB} \]

Moreover since
\[ \Phi(0, \tau) = \Phi^{-1}(\tau, 0) \]

it must be true that
\[ \Phi(0, \tau) = e^{-\tau B} Q(\tau) \]

Thus
\[ \Phi(t, \tau) = \Phi(t, 0) \Phi(0, \tau) = Q^{-1}(t)e^{tB} e^{-\tau B} Q(\tau) \]

We have proved the following theorem.

**Theorem 6 (Floquet)** It is possible to write the state transition matrix \( \Phi(t, \tau) \) of any continuous matrix \( A(t) \) which satisfies \( A(t + T) = A(t), \quad t \geq 0 \) as
\[ \Phi(t, \tau) = Q^{-1}(t)e^{(t - \tau)B} Q(\tau), \quad t, \tau \in [0, \infty) \]

where \( B \) is a constant matrix and \( Q(t) \) is a nonsingular, continuously differentiable matrix satisfying \( Q(t + T) = Q(t), \quad t \geq 0 \).
CHAPTER 6. LINEAR DIFFERENTIAL EQUATIONS
Chapter 7

Matrix Similarity

Two $n \times n$ matrices $A$ and $B$ with elements in a field $\mathbb{K}$ are similar if there exists a nonsingular $n \times n$ matrix $T$ with elements in $\mathbb{K}$ such that

$$TAT^{-1} = B \quad (7.1)$$

Matrix similarity is an equivalence relation on the set of all $n \times n$ matrices with elements in $\mathbb{K}$. The aim of this chapter is to explain the significance and implications of matrix similarity.

7.1 Motivation

Consider the differential equation

$$\dot{x} = Ax \quad (7.2)$$

where $A$ is a constant $n \times n$ matrix with elements in $\mathbb{K}$. Suppose that $T$ and $B$ are any $\mathbb{K}$-matrices such that (7.1) holds. If we define the new state vector

$$z \triangleq Tx \quad (7.3)$$

then from (7.2) we see that

$$\dot{z} = T\dot{x} = TAx = TAT^{-1}z$$

Therefore because of (7.1)

$$\dot{z} = Bz \quad (7.4)$$

Thus the similarity transformation

$$A \mapsto TAT^{-1} = B$$

describes what happens to $A$ when one makes the change of state variables

$$x \mapsto Tx = z \quad (7.5)$$

We claim that

$$e^{tB} = Te^{tA}T^{-1} \quad (7.6)$$

and thus that

$$e^{tA} \mapsto Te^{tA}T^{-1} = e^{tB}$$
describes what happens to $e^{tA}$ under the change of state variables given by (7.5). To understand why (7.6) is so, first observe that the matrix

$$\Theta(t) \triangleq Te^{tA}T^{-1}$$

satisfies

$$\dot{\Theta} = T Ae^{tA}T^{-1} = (TAT^{-1})(Te^{tA}T^{-1}) = (TAT^{-1})\Theta$$

and thus

$$\dot{\Theta} = B\Theta$$

Since $e^{tB}$ satisfies the same differential equation and since $\Theta(t)$ and $e^{tB}$ both equal the identity matrix at $t = 0$, by uniqueness it must be true that $e^{tB} = \Theta(t)$, $t \geq 0$. In other words, (7.6) is true.

Equation (7.6) suggests a procedure for calculating $e^{tA}$: Find matrices $T$ and $B$ such that (7.1) holds and in addition, such that $e^{tB}$ can be evaluated by inspection. Let us digress briefly to discuss one class of matrices whose matrix exponentials are very easy to compute.

### 7.1.1 The Matrix Exponential of a Diagonal Matrix

Let us consider an arbitrary diagonal matrix

$$D \triangleq \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & d_n \end{bmatrix},$$

whose diagonal entries are elements in $\mathbb{K}$. For such a matrix, the definition of a matrix exponential

$$e^D \triangleq \sum_{i=0}^{\infty} \frac{1}{i!} D^i$$

readily reduces to

$$e^D \triangleq \begin{bmatrix} \sum_{i=0}^{\infty} \frac{d_i^i}{i!} & 0 & \cdots & 0 \\ 0 & \sum_{i=0}^{\infty} \frac{d_i^i}{i!} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{i=0}^{\infty} \frac{d_i^i}{i!} \end{bmatrix} = \begin{bmatrix} e^{d_1} & 0 & \cdots & 0 \\ 0 & e^{d_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{d_n} \end{bmatrix}$$

In other words the matrix exponential $e^D$ of a diagonal matrix $D$ is the diagonal matrix

$$e^D = \begin{bmatrix} e^{d_1} & 0 & \cdots & 0 \\ 0 & e^{d_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{d_n} \end{bmatrix}$$

### 7.1.2 The State Transition Matrix of a Diagonalizable Matrix

Suppose that we can somehow find a nonsingular $n \times n$ matrix $T$ such that

$$TAT^{-1} = \Lambda$$

(7.7)
7.2 Eigenvalues

where Λ is a diagonal matrix; i.e.,

$$
\Lambda = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n
\end{bmatrix}
$$

(7.8)

Since tΛ is diagonal, it must therefore be true that

$$
e^{t\Lambda} = \begin{bmatrix}
e^{t\lambda_1} & 0 & \cdots & 0 \\
0 & e^{t\lambda_2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & e^{t\lambda_n}
\end{bmatrix}
$$

(7.9)

Moreover by exploiting (7.6) with $B \overset{\Delta}{=} \Lambda$ we can write

$$
e^{tA} = T^{-1}e^{t\Lambda}T, \ t \geq 0
$$

(7.10)

This suggests a procedure for calculating $e^{tA}$:

1. Find, if possible, a nonsingular matrix $T$ and a diagonal matrix $\Lambda$ as in (7.8) for which (7.7) holds.
2. Construct the matrix exponential $e^{t\Lambda}$ as in (7.9).
3. Use (7.10) to obtain $e^{tA}$.

For this to be a constructive method, one of course must know if step 1 can in fact be completed and if so, how one might go about doing it. These are issues in linear algebra. We now begin a discussion aimed at resolving both of them.

### 7.2 Eigenvalues

Let $A$ be a given $n \times n$ matrix over $\mathbb{K}$. Whether $A$ is real or complex, a complex number $\lambda \in \mathbb{C}$ is said to be an eigenvalue of $A$ is there exists a nonzero vector $x \in \mathbb{C}^n$ such that

$$Ax = \lambda x
$$

(7.11)

Any vector $x$ for which (7.11) holds is called an eigenvector of $A$ for eigenvalue $\lambda$. Note that the set of all $x$ for which (7.11) holds is closed under addition and scalar multiplication and is thus a subspace of $\mathbb{C}^n$; this subspace is uniquely determined by $A$ and $\lambda$ and is called the eigenspace of $\lambda$. It is easy to see that the kernel of $\lambda I - A$ is the eigenspace of $\lambda$. Clearly any nonzero vector in the eigenspace of $\lambda$ is an eigenvector for $\lambda$.

#### 7.2.1 Characteristic Polynomial

Suppose we wish to calculate a eigenvalue of $A$. In view of (7.11), what we need to do is to find a number $\lambda$ (the eigenvalue) and a nonzero vector $x$ such that

$$(\lambda I - A)x = 0
$$

(7.12)
where $I$ is the $n \times n$ identity matrix. Now for arbitrary but fixed $\lambda$, (7.12) will have a nonzero solution $x$ if and only if the kernel of $\lambda I - A$ is nonzero. For $\lambda I - A$ to have a nonzero kernel is equivalent to the requirement that $\lambda I - A$ have linearly dependent columns. Since $\lambda I - A$ is a square matrix, its columns will be linearly dependent just in case rank $(\lambda I - A) < n$ or equivalently $\lambda I - A$ is singular. In other words $\lambda$ will be an eigenvalue of $A$ if and only if
\begin{equation}
\det(\lambda I - A) = 0
\end{equation}

With the preceding in mind let us define
\begin{equation}
\alpha(s) \triangleq \det(sI - A)
\end{equation}
where $s$ is a variable. Note that $\alpha(s)$ must be a polynomial in $\mathbb{K}[s]$, the ring of polynomials in the variable $s$ with coefficients in $\mathbb{K}$. This is because $sI - A$ is a matrix whose elements are polynomials in $\mathbb{K}[s]$, because the determinant of any matrix is the sum of products of its elements, and because $\mathbb{K}[s]$ is closed under polynomial addition and multiplication. We call $\alpha(s)$ the characteristic polynomial of $A$.

Using basic properties of determinants it is not difficult to prove that $\alpha(s)$ is a polynomial of degree $n$ whose leading coefficient is one. Polynomials with the latter property are sometimes referred to as monic.

From the definition of $\alpha(s)$ and (7.13) it is clear that for $\lambda$ to be an eigenvalue of $A$, it is both necessary and sufficient that $\lambda$ be a root of $\alpha(s)$; i.e., $\lambda$ must be a solution to the characteristic equation
\begin{equation}
\alpha(\lambda) = 0
\end{equation}
This of course is just another way to write (7.13). From these observations we can draw the following conclusions:

1. The spectrum of $A$ {i.e., the set of eigenvalues of $A$} consists of precisely $n$ complex numbers, namely the $n$ roots of the characteristic polynomial of $A$.

2. If $A$ is a real matrix, then its characteristic polynomial $\alpha(s)$ is in $\mathbb{R}[s]$ and thus has real coefficients. Because of this, all of $\alpha(s)$’s nonreal roots must occur in complex conjugate pairs. Thus all nonreal eigenvalues of a real matrix $A$ must occur in conjugate pairs. To emphasize this, we sometimes say that $A$’s spectrum is symmetric.

Suppose $\lambda$ is an eigenvalue of $A$. Then $s - \lambda$ must be a factor of $\alpha(s)$. The largest value of $d$ for which $(s - \lambda)^d$ is a factor of $\alpha(s)$ is called the algebraic multiplicity of $\lambda$. Meanwhile the dimension of the eigenspace of $\lambda$ is called the geometric multiplicity of $\lambda$. The geometric multiplicity of $\lambda$ can never be larger that the algebraic multiplicity of $\lambda$.

7.2.2 Computation of Eigenvectors

Suppose $\lambda$ is an eigenvalue of $A$. To calculate an eigenvector $x$ for this eigenvalue, what we need to do is to find a nonzero solution to the linear equation
\begin{equation}
(\lambda I - A)x = 0
\end{equation}
Note that such a solution must exist because $(\lambda I - A)$ has linearly dependent columns. To calculate such an $x$ for small $n$, one might attempt to solve (7.16) by hand. For large $n$, one would turn something like Gauss Elimination, perhaps transforming $(\lambda I - A)$ into reduced echelon form and then solving for $x$ by inspection. Note that $x$ is never uniquely determined, so any attempt to come up with an explicit formula for $x$ is out of the question.
A matrix $A_{n \times n}$ over $\mathbb{K}$ is said to be semisimple if it has $n$ linearly independent eigenvectors. Suppose $A$ is semisimple with a linearly independent eigenvector set $\{q_1, q_2, \ldots, q_n\}$. Then $\{q_1, q_2, \ldots, q_n\}$ is a basis for $\mathbb{C}^n$. Therefore the matrix

$$Q \triangleq [q_1 \ q_2 \ \cdots \ q_n]$$

is nonsingular. Moreover, since

$$Aq_i = \lambda_i q_i, \quad i \in \{1, 2, \ldots, n\}$$

it must be true that

$$AQ = [Aq_1 \ Aq_2 \ \cdots \ Aq_n] = [\lambda_1 q_1 \ \lambda_2 q_2 \ \cdots \ \lambda_n q_n]$$

and thus that

$$AQ = Q \Lambda$$

where

$$\Lambda = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n
\end{bmatrix}$$

It follows that $A$ is similar to $\Lambda$; i.e.,

$$Q^{-1}AQ = \Lambda$$

We leave it to the reader to verify that the converse is also true. That is, if $B$ is a matrix for which there exists a matrix $P$ such that $P^{-1}BP$ is diagonal, then the columns of $P$ are eigenvectors of $B$ and consequently $B$ is a semisimple matrix. We summarize:

**Theorem 7** A matrix $A_{n \times n}$ is semisimple if and only if there exists a nonsingular matrix $Q$ such that $Q^{-1}AQ = \Lambda$ where $\Lambda$ is a diagonal matrix. If $Q$ and $\Lambda$ are such matrices, then the $n$ diagonal elements of $\Lambda$ are $A$’s eigenvalues and the $n$ columns of $Q$ are eigenvectors of $A$ which together form a linearly independent set.

**Example 44** Suppose

$$A = \begin{bmatrix}
0 & -2 & 0 \\
1 & -3 & 0 \\
0 & 0 & 0
\end{bmatrix}$$

Then

$$\alpha(s) = \det(sI - A) = \det \begin{bmatrix}
\lambda & 2 & 0 \\
-1 & s + 3 & 0 \\
0 & 0 & s
\end{bmatrix} = s(s^2 + 3s + 2) = s(s + 1)(s + 2)$$

Hence $A$’s three eigenvalues are 0, $-1$ and $-2$. To compute an eigenvectors $q_1$, $q_2$, $q_3$ for $\lambda_1 = 0$, $\lambda_2 = -3$, $\lambda_3 = -2$ we need to solve the equations

$$\begin{bmatrix}
0 & 2 & 0 \\
-1 & 3 & 0 \\
0 & 0 & 0
\end{bmatrix} q_1 = 0, \quad \begin{bmatrix}
-2 & 2 & 0 \\
-1 & 1 & 0 \\
0 & 0 & -2
\end{bmatrix} q_2 = 0, \quad \begin{bmatrix}
-1 & 2 & 0 \\
-1 & 2 & 0 \\
0 & 0 & -1
\end{bmatrix} q_3 = 0$$

One set of solutions is

$$q_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad q_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad q_3 = \begin{bmatrix} 2 \\ 1 \\ 0 \end{bmatrix}$$
Thus if we define \( Q \equiv [q_1 \ q_2 \ q_3] \), then

\[
Q^{-1}AQ = \begin{bmatrix}
0 & 0 & 0 \\
0 & -2 & 0 \\
0 & 0 & -1
\end{bmatrix}
\]

It is a sad but true fact that there is no simple way to decide if a given matrix is semisimple. On the other hand it turns out that “almost every” matrix is semisimple. Confused? Stay tuned . . .

### 7.2.4 Eigenvectors and Differential Equations

Let \( A_{n \times n} \) be a given constant matrix and consider the differential equation

\[
\dot{x} = Ax \tag{7.17}
\]

with initial state \( x(0) = d \). We want to describe the form of the solution to (7.17) for various types of \( d \)’s.

1. Suppose first that \( d \) is an eigenvector of \( A \) for eigenvalue \( \lambda \). Then

\[
x(t) \triangleq e^{t\lambda}d \tag{7.18}
\]

must be the solution to (7.17) because with \( x \) so defined \( x(0) = d \) and

\[
\dot{x}(t) = \lambda e^{t\lambda}d = e^{t\lambda}Ad = Ax(t)
\]

The form of \( x(t) \) in (7.18) clearly shows that if the initial state of (7.17) lies in \( \text{span} \{d\} \) then the value of the corresponding solution \( x(t) \) at each instant of time, is in \( \text{span} \{d\} \), since \( e^{t\lambda}d \) is a scalar multiple of \( d \). In other words, any trajectory of (7.17) starting in \( \text{span} \{d\} \) never leaves \( \text{span} \{d\} \).

**Example 45** If

\[
A = \begin{bmatrix}
0 & -2 & 0 \\
1 & -3 & 0 \\
0 & 0 & 0
\end{bmatrix}
\quad \text{and} \quad
d = \begin{bmatrix}
1 \\
1 \\
0
\end{bmatrix}
\]

then \( Ad = -2d \). Thus

\[
x(t) = \begin{bmatrix}
x_1(t) \\
x_2(t) \\
x_3(t)
\end{bmatrix} = \begin{bmatrix}
e^{-t} \\
e^{-t} \\
0
\end{bmatrix}
\]
2. Now suppose that $d$ is a vector in some subspace $E$ which in turn is spanned by some set of eigenvectors \{q_1, q_2, \ldots, q_m\}. In other words, suppose that
\[ d \in E = \mathcal{E}_1 + \mathcal{E}_2 + \cdots + \mathcal{E}_m \]
where for each $i$, $\mathcal{E}_i$ is spanned by $q_i$. This means there must be numbers $\mu_i \in \mathbb{C}$ such that
\[ d = \mu_1 q_1 + \mu_2 q_2 + \cdots + \mu_m q_m \]
as well as eigenvalues $\lambda_i, \ i \in \{1, 2, \ldots, m\}$ such that $A q_i = \lambda_i q_i, \ i \in \{1, 2, \ldots, m\}$. Using the same reasoning as before it is easy to establish that under these conditions, the solution to (7.17) is
\[ x(t) = \mu_1 e^{t \lambda_1} q_1 + \mu_2 e^{t \lambda_2} q_2 + \cdots + \mu_m e^{t \lambda_m} q_m \]
Since for each fixed value of $t$, $x(t)$ is a linear combination of the $q_i$, it must be true that for each $t \geq 0$, $x(t) \in E$. We can therefore conclude that any trajectory of (7.17) starting in $E$ remains in $E$ forever.

### 7.3 A - Invariant Subspaces

The subspace $E$ just discussed is an example of an “A - invariant subspace.” The aim of this section is to explain what an A - invariant subspace is and to discuss some of the implications of A - invariance.

Let $\mathcal{U}$ be a subspace of $\mathbb{C}^n \{\text{or perhaps } \mathbb{R}^n \text{ if } A \text{ is real}\}$ with the property that $Au \in \mathcal{U}$ for every $u \in \mathcal{U}$. We call $\mathcal{U}$ an $A$-invariant subspace. To denote the $A$-invariance of $\mathcal{U}$ we often write $A \mathcal{U} \subset \mathcal{U}$. The notation is appropriate if we take $A \mathcal{U}$ to mean the subspace $A \mathcal{U} \equiv \{Au : u \in \mathcal{U}\}$. Examples of $A$-invariant subspaces include the zero space, $\mathbb{C}^n$, eigenspaces and sums of eigenspaces. Later we will encounter other types of $A$-invariant subspaces.

Now suppose that the initial state $d$ of (7.17) is a vector in an $A$-invariant subspace $\mathcal{U}$. Since $\dot{x}(t) = Ax(t)$, it follows that $\dot{x}(0) = Ad \in \mathcal{U}$. Thus the solution to (7.17) which starts at $d$, is not only starting at a point in $\mathcal{U}$, but also has a derivative which is initially in $\mathcal{U}$. From this it is plausible \{and can easily be shown\}...
that the entire trajectory starting at $d$, remains in $U$ forever. In other words, trajectories of (7.17) starting in an $A$-invariant subspace never leave the subspace.

The fact that $U$ is an $A$-invariant subspace has implications concerning $A$’s algebraic structure under similarity transformations. Our aim is to explain what these implications are. For this let $V$ be any completion\(^1\) from $U$ to $\mathbb{K}^n$; i.e., any subspace such that $U \oplus V = \mathbb{K}^n$. Pick bases $\{u_1, u_2, \ldots, u_m\}$ and $\{v_1, v_2, \ldots, v_{n-m}\}$ for $U$ and $V$ respectively. Then $\{u_1, u_2, \ldots, u_m, v_1, v_2, \ldots, v_{n-m}\}$ is a basis for $\mathbb{K}^n$. It follows that the basis matrices

$$
U \triangleq \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix}_{n \times m} \quad \text{and} \quad U \triangleq \begin{bmatrix} v_1 & v_2 & \cdots & v_{n-m} \end{bmatrix}_{n \times (n-m)}
$$

have linearly independent columns and that the matrix

$$
T \triangleq \begin{bmatrix} U & V \end{bmatrix}_{n \times n}
$$

is nonsingular.

Note that since $A \mathbb{U} \subset U$, the span of the columns of $AU$ {namely $A\mathbb{U}$} must be a subspace of the span of the columns of $U$ {namely $U$}. Thus the necessary and sufficient condition is satisfied for there to exist a solution $B_{m \times m}$ to the linear matrix equation

$$
AU = UB
$$

(7.19)

Moreover, because $U$ has linearly independent columns, kernel $U = 0$ so $B$ is even unique. We will call $B$ a restriction of $A$ to $U$. On your homework, you are asked to prove that all possible restrictions of $A$ to $U$, resulting from all possible choices of bases for $U$, are similar.

In view of (7.19) we can write

$$
AU = \begin{bmatrix} U & V \end{bmatrix} \begin{bmatrix} B \\ 0 \end{bmatrix}
$$

or

$$
AU = T \begin{bmatrix} B \\ 0 \end{bmatrix}
$$

(7.20)

where the 0 in these partitioned matrices denotes the $(n - m) \times m$ zero matrix. What’s more, if we defined the partitioned matrix

$$
\begin{bmatrix} C \\ D \end{bmatrix} \triangleq T^{-1} AV,
$$

where $C$ and $D$ are $m \times (n - m)$ and $(n - m) \times (n - m)$ matrices respectively, then

$$
AV = T \begin{bmatrix} C \\ D \end{bmatrix}
$$

(7.21)

Combining (7.20) and (7.21) we thus obtain

$$
A[U \quad V] = T \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}
$$

or

$$
AT = T \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}
$$

---

\(^1\)Such a subspace can be constructed as follows. First pick bases $\{u_1, u_2, \ldots, u_m\}$ and $\{x_1, x_2, \ldots, x_n\}$ for $U$ and $\mathbb{K}^n$ respectively. Second, starting from the left, successively eliminate from the set $\{u_1, u_2, \ldots, u_m, x_1, x_2, \ldots, x_n\}$ vectors which are linearly dependent on their predecessors. What results is a basis for $\mathbb{K}^n$ of the form $\{u_1, u_2, \ldots, u_m, v_1, v_2, \ldots, v_{n-m}\}$ where each $v_i \in \{x_1, x_2, \ldots, x_n\}$. Defining $V$ to be the span of $\{v_1, v_2, \ldots, v_{n-m}\}$ yields a subspace with the required properties.
Therefore

\[ T^{-1}AT = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix} \]

The matrix on the right with square diagonal blocks \( B \) and \( D \) is called a \textit{block triangular matrix}. Thus we can conclude the following: \textit{If \( U \) is an \( A \)-invariant subspace, then \( A \) is similar to a block triangular matrix with one diagonal block being a restriction of \( A \) to \( U \).}

Why is this interesting? Well for one thing, the characteristic polynomial of a block triangular matrix is equal to the product of the characteristic polynomials of its diagonal blocks; i.e.,

\[ \det \left( sI - \begin{bmatrix} B & C \\ 0 & D \end{bmatrix} \right) = (\det(sI - B)) (\det(sI - D)) \]

Thus the spectrum of a block triangular matrix is the union of the spectra of its diagonal blocks. More on this a little later.

### 7.3.1 Direct Sum Decomposition of \( \mathbb{K}^n \)

Suppose that the completion \( V \) constructed above is also \( A \)-invariant. In other words, let us consider the situation in which

\[ \mathcal{U} \oplus \mathcal{V} = \mathbb{K}^n, \quad \mathcal{A}\mathcal{U} \subset \mathcal{U}, \quad \mathcal{A}\mathcal{V} \subset \mathcal{V} \quad (7.22) \]

Then it is reasonably clear and can easily be verified using the same reasoning as above that with \( T, B, C, D \) as just defined, \( C \) will turn out to be the zero matrix. In other word, under the conditions of (7.22),

\[ T^{-1}AT = \begin{bmatrix} B & 0 \\ 0 & D \end{bmatrix} \]

Thus \( A \) is similar to a \textit{block diagonal matrix} whose diagonal blocks \( B \) and \( D \) are restrictions of \( A \) to \( \mathcal{U} \) and \( \mathcal{V} \) respectively.

These observations generalize. Suppose, for example, that \( X_1, X_2, \ldots, X_m \) are subspaces of \( \mathbb{K}^n \) satisfying

\[ X_1 \oplus X_2 \oplus \cdots \oplus X_m = \mathbb{K}^n \quad \text{and} \quad \mathcal{A}X_i \subset X_i, \quad i \in \{1, 2, \ldots, m\} \]

Then by forming the nonsingular matrix

\[ T \triangleq \begin{bmatrix} x_1^{(1)} & x_2^{(1)} & \cdots & x_{n_1}^{(1)} & x_1^{(2)} & x_2^{(2)} & \cdots & x_{n_2}^{(2)} & \cdots & x_1^{(m)} & x_2^{(m)} & \cdots & x_{n_m}^{(m)} \end{bmatrix} \]

where \( \{x_1^{(i)}, x_2^{(i)}, \ldots, x_{n_i}^{(i)}\} \) is a basis for \( X_i \), one can transform \( A \) into a block diagonal matrix as

\[ T^{-1}AT = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_m \end{bmatrix} \]

where \( A_i \) is a \( n_i \times n_i \) restriction of \( A \) to \( X_i \). Note that the diagonalization of a matrix into scalar diagonal blocks is a special case of this in which the \( X_i \) are independent eigenspaces and the \( A_i \) are \( 1 \times 1 \) matrices, namely the eigenvalues of \( A \). We summerize: \textit{If \( \mathbb{K}^n \) can be decomposed into a direct sum of \( m \) \( A \)-invariant subspaces \( X_i \), then \( A \) can be transformed by means of a similarity transformation into a block diagonal matrix whose diagonal blocks are restrictions of \( A \) to the \( X_i \).}
7.4 Similarity Invariants

There are certain properties of $A$ which are shared by all matrices similar to $A$. For example, if $A$ is $n \times n$, then so is every matrix similar to $A$. We say that the size of $A$ is \textit{invariant} under similarity transformations.

Let us note that if $T$ is any nonsingular matrix, then

$$
\det(sI - A) = \det(TT^{-1}(sI - A)) = \det(T) \det(T^{-1}) \det(sI - A) = \det(T(sI - A)T^{-1}) = \det(sI - TAT^{-1})
$$

Thus $A$ and $TAT^{-1}$ have one and the same characteristic polynomial. In other words, \textit{the characteristic polynomial of any square matrix is invariant under similarity transformations}. From this it follows that \textit{the spectrum of any square matrix is invariant under similarity transformations}.

Note for example that if $A$ is similar to a block triangular matrix with diagonal blocks $A_1$, $A_2$, \ldots, $A_m$, then the characteristic polynomial \{spectrum\} of $A$ must equal the product \{union\} of the characteristic polynomials \{spectra\} of the $A_i$. It is largely for these reasons that the direct sum decompositions discussed in the last section are of interest.

7.5 The Cayley-Hamilton Theorem

The following theorem is one of the most important results in linear algebra.

\textbf{Theorem 8 (Cayley-Hamilton)} \textit{Let $A$ be a given $n \times n$ matrix with characteristic polynomial $\alpha(s)$. Then}

$$
\alpha(A) = 0_{n \times n}
$$

The theorem says that if $\alpha(s)$ is the characteristic polynomial of $A$, then the matrix polynomial $\alpha(A)$ equals the zero matrix. Although somewhat confusing, the assertion “$A$ satisfies its own characteristic equation” is often taken as a statement of the Cayley-Hamilton Theorem.

\textbf{Example 46} As noted earlier, the characteristic polynomial of the matrix

$$
A = \begin{bmatrix}
0 & -2 & 0 \\
1 & -3 & 0 \\
0 & 0 & 0
\end{bmatrix}
$$

is

$$
\alpha(s) = s^3 + 3s^2 + 2s
$$

According to the Cayley-Hamilton

$$
A^3 + 3A^2 + 2A + 0I = 0
$$

To verify that this is so we first compute

$$
A^2 = \begin{bmatrix}
-2 & 6 & 0 \\
-3 & 7 & 0 \\
0 & 0 & 0
\end{bmatrix}
$$

and

$$
A^3 = \begin{bmatrix}
6 & -14 & 0 \\
7 & -15 & 0 \\
0 & 0 & 0
\end{bmatrix}
$$
7.6 Minimal Polynomial

So far we have shown that any semisimple matrix $A_{n \times n}$ can be diagonalized by means of a similarity transformation. Unfortunately, not every matrix is semisimple as the following example shows.

Example 47 The eigenvalues of the matrix

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

are both 0. Thus for

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

to be an eigenvector, it must be a nonzero solution to the equations

$$\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

But all such solutions must clearly be of the form

$$x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}$$

It is therefore impossible to find two eigenvectors which are not scalar multiples of each other. In other words $A$ does not have two linearly independent eigenvectors so it is not semisimple.

Some readers may think that the $A$ matrix used in this example is pathological, and that in practice one typically does not encounter matrices of this type. While it is true that the matrix has some very special features {e.g., its eigenvalues are both the same}, this matrix is nevertheless commonly encountered in practical modeling problems. For example, the cascade interconnection of two Simulink integrators shown in the following diagram, admits a state space model of the form

$$\begin{bmatrix} \dot{y}_1 \\ \dot{y}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u$$

In order to understand more clearly just why some matrices are semisimple while others are not, we need some new concepts. In the sequel we shall explain what they are.
7.6.1 The Minimal Polynomial of a Vector

Let $A_{n \times n}$ be a given $K$ matrix and let $x$ be any nonzero vector in $K^n$. Let $m$ be the least integer such that the set of vectors $\{x, Ax, \ldots, A^{m-1}x\}$ is a linearly independent set. Clearly $1 \leq m \leq n$. Since $\{x, Ax, \ldots, A^{m-1}x\}$ is a linearly independent set and $\{x, Ax, \ldots, A^m x\}$ is not, $A^m x$ must be a linear combination of the vectors in $\{x, Ax, \ldots, A^{m-1}x\}$. In other words, there must be exactly one set of numbers $a_i$ such that

$$A^m x = -a_1 x - a_2 Ax - \cdots - a_m A^{m-1} x$$

Thus

$$(A^m + a_m A^{m-1} + \cdots + a_2 A + a_1 I)x = 0 \quad (7.23)$$

Therefore if we define the polynomial

$$\alpha_x(s) \triangleq s^m + a_m s^{m-1} + \cdots + a_2 + a_1$$

then (7.23) can be rewritten as

$$\alpha_x(A)x = 0 \quad (7.24)$$

We call $\alpha_x(s)$ the minimum polynomial of $x$ {relative to $A$}. Note that $\alpha_x(s)$ is uniquely determined by $A$ and $x$. {For consistency we define the minimal polynomial of the zero vector to be the polynomial 1.}

Let us note that if $I_x$ is the {nonempty} set of all polynomials $\beta(s) \in K[s]$ such that $\beta(A)x = 0$, then $I_x$ is closed under polynomial addition and multiplication by polynomials in $K[s]$; i.e.,

$$\beta(A)x = 0, \quad \delta(A)x = 0 \quad \Rightarrow \quad (\beta(A) + \delta(A))x = 0$$

$$\beta(A)x = 0, \quad \pi(s) \in K[s] \quad \Rightarrow \quad \pi(A)\beta(A)x = 0$$

Thus $I_x$ is an ideal in $K[s]$. It follows from Proposition 2 that $I_x$ contains a unique monic polynomial $\gamma(s)$ of least degree which divides {without remainder} every polynomial in $I_x$. Since $\alpha_x(s)$ is monic, is in $I_x$, and by construction is of least degree it must be true that $\gamma(s) = \alpha_x(s)$. In other words, if $\beta(s)$ is any polynomial such that $\beta(A)x = 0$ then $\beta(s)$ must be divisible by $\alpha_x(s)$.

7.6.2 The Minimal Polynomial of a Finite Set of Vectors

Now suppose that $X \triangleq \{x_1, x_2, \ldots, x_m\}$ is a finite set of vectors in $K^n$. If we define $I_X$ to be the set of all polynomials $\beta(s)$ such that

$$\beta(A)x_i = 0, \quad i \in \{1, 2, \ldots, m\}$$

then $I_X$ is also an ideal in $K[s]$. {The reader may wish to verify this using the same reasoning as above.} Again from Proposition 2 it follows that there is a unique monic polynomial $\alpha_X(s)$ of least degree which divides every polynomial in $I_X$. We call $\alpha_X(s)$ the minimal polynomial of $X$ relative to $A$. It is not difficult to prove that

$$\alpha_X(s) = \text{lcm}\{\alpha_{x_1}(s), \alpha_{x_2}(s), \ldots, \alpha_{x_m}(s)\}$$

where $\alpha_{x_i}(s)$ is the minimal polynomial of $x_i$.

7.6.3 The Minimal Polynomial of a Subspace

Now suppose that $U$ is a subspace of $K^n$ and let $I_U$ be the set of all polynomials $\beta(s) \in K[s]$ such that

$$\beta(A)u = 0, \quad \forall u \in U$$
7.6. MINIMAL POLYNOMIAL

$I_U$ is also an ideal and its unique, monic generating polynomial, $\alpha_U(s)$, is called the minimal polynomial of $U$ relative to $A$. Suppose $U$ is not the zero subspace and let $\{u_1, u_2, \ldots, u_m\}$ be a basis for $U$. It is easy to verify that $\beta(s) \in I_U$ if and only if

$$\beta(A)u_i = 0, \ i \in \{1, 2, \ldots, m\}$$

From this it follows that the minimal polynomial of $U$ is the same as the minimal polynomial of any basis for $U$. Hence

$$\alpha_U = \text{lcm}\{\alpha_{u_1}(s), \alpha_{u_2}(s), \ldots, \alpha_{u_m}(s)\}$$

7.6.4 The Minimal Polynomial of $A$

In the special case when $U = \mathbb{K}^n$, $\alpha_U$ is written without a subscript as $\alpha(s)$ and is called the minimal polynomial of $A$. Thus the minimal polynomial of $A$ is the unique monic polynomial of least degree such that

$$\alpha(A)x = 0, \ \forall x \in \mathbb{K}^n$$

This is equivalent to saying that $\alpha(s)$ is the unique monic polynomial of least degree such that

$$\alpha(A) = 0$$

Let us recall that if $\beta(s)$ is the characteristic polynomial of $A$ then $\beta(A) = 0$ because of the Cayley-Hamilton Theorem. Since $\alpha(s)$ divides every polynomial $\delta(s)$ such that $\delta(A) = 0$ it must be true that the minimal polynomial of $A$ divides the characteristic polynomial of $A$. An useful consequence of this fact is that if $A$ had $n$ distinct eigenvalues, the $A$'s minimal and characteristic polynomials must be one and the same.

It can be shown without difficulty that if $Y_1$ and $Y_2$ are any nonempty subsets or subspaces of $\mathbb{K}$ and if $Y_1 \subset Y_2$, then $\alpha_{Y_1}(s)$ divides $\alpha_{Y_2}(s)$. Thus the minimal polynomial of any subspace of $\mathbb{K}^n$ divides the minimal polynomial of $A$.

Example 48 If

$$A = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix} \quad \text{and} \quad x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

then

$$\text{minimum polynomial of } x = s - 2 \quad (A - 2I)x = 0$$

$$\text{minimum polynomial of } A = (s - 2)^2 \quad (A - 2I)^2 = 0$$

$$\text{characteristic polynomial of } A = (s - 2)^3 \quad (A - 2I)^3 = 0$$

7.6.5 A Similarity Invariant

Let $A$ and $B$ be similar $n \times n$ matrices over $\mathbb{K}$. Then there must be a nonsingular matrix $T$ such that

$$B = TAT^{-1}$$
Note that
\[ TA^0T^{-1} = I = B^0 \]
\[ TAT^{-1} = B \]
\[ TA^2T^{-1} = TAA^{-1} = TAT^{-1} = B^2 \]
\[ \vdots \]
\[ TAT^{-1} = B^i \]
\[ \vdots \]

More generally, for any polynomial \( \beta(s) \in \mathbb{K}[s] \),
\[ T\beta(A)T^{-1} = \beta(B) \]

Thus if \( \beta \) is the minimal polynomial of \( B \), then \( T\beta(A)T^{-1} = 0 \). Since \( T \) is nonsingular, \( \beta(A) = 0 \). Therefore \( \beta(s) \) must be divisable by the minimal polynomial of \( A \). By repeating the same argument with \( B \) and \( A \) interchanged, we can establish that the minimal polynomial of \( A \) must be divisable by \( \beta(s) \). Since both polynomials divide each other and are monic, they must be one and the same. We have proved that the minimal polynomial of \( A \) equals the minimal polynomial of any matrix similar to \( A \). In other words, the minimum polynomial of \( A \) is invariant under similarity transformations.

### 7.7 A Useful formula

The rational matrix \((sI - A)^{-1}\), called the resolvent of \( A \), plays a key role in the determination of \( e^{tA} \) by means of Laplace Transform Techniques. While \((sI - A)^{-1}\) can be expanded in the formal power series
\[ (sI - A)^{-1} = \sum_{i=1}^{\infty} A^{i-1}s^{-i} \]
which converges for all values of \(|s|\) sufficiently large, it is sometimes useful to express this matrix in more explicit terms. To do this, let
\[ \alpha(s) \triangleq s^m + a_n s^{m-1} + \cdots + a_2 s + a_1 \]
be any monic polynomial such that \( \alpha(A) = 0 \). Thus for example, \( \alpha(s) \) might be either the minimal or characteristic polynomial of \( A \). Next, successively define polynomials \( \alpha_m(s) \), \( \alpha_{m-1}(s) \), \ldots, \( \alpha_1(s) \), \( \alpha_0(s) \) so that
\[ \alpha_m(s) \triangleq 1 \]
\[ \alpha_{i-1}(s) \triangleq s\alpha_i(s) + a_i, \quad i \in \{m, m-1, \ldots, 1\} \]

{Note that \( \alpha_0(s) = \alpha(s) \).} Starting with the easily verifiable identity
\[ (sI - A) \sum_{i=1}^{m} \pi_i(s)A^{i-1} = s\pi_1(s) - \pi_m(s)A^m + \sum_{i=2}^{m} \{s\pi_i(s) - \pi_{i-1}(s)\} A^{i-1}, \]
which holds for any set of \( m \) polynomials \( \pi_1(s) \), \( \pi_2(s) \), \ldots, \( \pi_m(s) \), one can show that
\[ (sI - A)^{-1} = \frac{1}{\alpha(s)} \sum_{i=1}^{m} \alpha_i(s)A^{i-1} \quad (7.25) \]
7.7.1 Application

Recall that the matrix exponential $e^{tA}$ uniquely solves the matrix differential equation/initial condition

$$\frac{d}{dt}e^{tA} = Ae^{tA}, \quad e^{tA}\big|_{t=0} = I$$

Thus if $X(s)$ is the Laplace Transform$^2$ of $e^{tA}$, then

$$sX(s) - I = AX(s)$$

Therefore

$$(sI - A)X(s) = I$$

so

$$X(s) = (sI - A)^{-1}$$

Hence using (7.25) we can write

$$X(s) = \frac{1}{\alpha(s)} \sum_{i=1}^{m} \alpha_i(s)A^{i-1}$$

Therefore

$$e^{tA} = \sum_{i=1}^{m} g_i(t)A^{i-1}$$

where $g_i(t)$ is the inverse Laplace Transform of the rational function $\frac{\alpha_i(s)}{\alpha(s)}$. Of course to complete the calculation of $e^{tA}$ one would have to actually carry out the evaluations of the $g_i(t)$. This, in turn, is not really any easier that the eigenvalue approach we’re pursuing here.

7.8 Cyclic Subspaces

Let $A_{n \times n}$ be a given matrix with elements in $\mathbb{K}$. A nonzero subspace $\mathcal{U} \subset \mathbb{K}^n$ is $A$-cyclic if there exists a vector $g \in \mathcal{U}$, called a generator, such that

$$\mathcal{U} = \text{span}\{g, Ag, \ldots, A^{n-1}g\}$$

Note that each vector $x \in \mathbb{K}^n$ generates a unique cyclic subspace, namely

$$\text{span}\{x, Ax, \ldots, A^{n-1}x\}$$

Cyclic subspaces are $A$-invariant subspaces. To understand why this is so, suppose $u \in \mathcal{U}$ where $\mathcal{U} = \text{span}\{g, Ag, \ldots, A^{n-1}g\}$. Then there must be numbers $\mu_i \in \mathbb{K}$ such that

$$u = \mu_1g + \mu_2Ag + \cdots + \mu_nA^{n-1}g$$

Therefore $Au = \mu_1Ag + \mu_2A^2g + \cdots + \mu_nA^ng$. But by the Cayley-Hamilton Theorem, there must be numbers $a_i \in \mathbb{K}$ such that

$$A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \cdots + a_2A + a_1I = 0$$

Thus

$$Au = -a_1g + (\mu_1 - a_2)Ag + (\mu_2 - a_3)A^2g + \cdots + (\mu_{n-1} - a_n)A^{n-1}g$$

so $Au \in \mathcal{U}$. This establishes the $A$-invariance of $\mathcal{U}$.

$^2$Review this material when you study Laplace Transforms.
It is not difficult to prove that if $\mathcal{U}$ is cyclic with generator $g$, and if $m$ is the largest integer such that $\{g, Ag, \ldots, A^{m-1}g\}$ is an independent set, then $\{g, Ag, \ldots, A^{m-1}g\}$ is a basis for $\mathcal{U}$. Using this, it can then be shown that the minimal polynomial of $\mathcal{U}$ is the same as the minimal polynomial of $g$.

If the whole space $\mathbb{K}^n$ is cyclic, $A$ is said to be a cyclic matrix. Thus $A$ is cyclic just in case there is a vector $x \in \mathbb{K}^n$ such that $\{x, Ax, \ldots, A^{n-1}x\}$ is a basis for $\mathbb{K}^n$. A little thought reveals that cyclic matrices are precisely those matrices whose minimal and characteristic polynomials are one and the same. Previously we noted that matrices with all eigenvalues distinct have the same minimal and characteristic polynomials. It follows that matrices with distinct eigenvalues are cyclic.

### 7.8.1 Companion Forms

Let $A_{n \times n}$ be a cyclic matrix with characteristic polynomial

$$\alpha(s) = s^n + a_n s^{n-1} + \cdots + a_2 s + a_1$$

Let $g$ be any vector which generates $\mathbb{K}^n$. Then $\{g, Ag, \ldots, A^{n-1}g\}$ is a basis for $\mathbb{K}^n$ and thus

$$Q \triangleq [g \ Ag \ \ldots \ A^{n-1}g]$$

is a nonsingular matrix. Define the row vector

$$c \triangleq [0 \ 0 \ \cdots \ 0 \ 1]_{1 \times n} Q^{-1}$$

Then

$$cQ = [0 \ 0 \ \cdots \ 0 \ 1]_{1 \times n}$$

so

$$cg = 0, \ cAg = 0, \ldots, cA^{n-2}g = 0 \quad (7.26)$$

and

$$cA^{n-1}g = 1 \quad (7.27)$$

Define the matrix

$$T \triangleq \begin{bmatrix} c \\ cA \\ \vdots \\ cA^{n-1} \end{bmatrix}_{n \times n}$$

(7.28)

and note that

$$Tg = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}_{n \times 1}$$

We claim that $T$ is nonsingular. Suppose it were not; then it would be possible to find numbers $b_i$ such that

$$cA^{n-1} = \sum_{i=1}^{n-1} b_i cA^{i-1}$$

This would imply that

$$cA^{n-1}g = \sum_{i=1}^{n-1} b_i cA^{i-1}g = 0$$
In view of the definition of $T$ in (7.28)

$$TA = \begin{bmatrix} cA \\ cA^2 \\ \vdots \\ cA^n \end{bmatrix} \quad (7.29)$$

and

$$cA^i = e'_{i+1}T, \quad i \in \{1, 2, \ldots, n-1\} \quad (7.30)$$

where $e_{i+1}$ is the $ith$ unit vector in $\mathbb{K}^n$. Meanwhile, by the Cayley-Hamilton Theorem

$$cA^n = -a_1c - a_2cA - \cdots - a_n cA^{n-1}$$

so

$$cA^n = [-a_1 - a_2 \cdots - a_n]T \quad (7.31)$$

because of the definition of $T$ in (7.28). From (7.29) - (7.31) we thus obtain the expression

$$TA = \begin{bmatrix} e'_2T \\ e'_3T \\ \vdots \\ e'_nT \\ [-a_1 - a_2 \cdots - a_n]T \end{bmatrix}$$

Factoring out $T$ on the right we get

$$TA = A_CT \quad (7.32)$$

where

$$A_C = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_n \end{bmatrix}_{n \times n}$$

Therefore

$$A_C \triangleq TAT^{-1} \quad (7.33)$$

$A_C$ is called a companion form. Note that the bottom row of $A_C$ consists of the negative of the row vector of coefficients of the characteristic polynomial of $A$. In other words, $A_C$ is uniquely determined by $A$’s characteristic polynomial and consequently by $A$ itself. It thus makes sense to call $A_C$ the companion (canonical) form of $A$. For the equivalence relation similarity defined on the class of $n \times n$ cyclic matrices, companion forms are thus canonical forms and the characteristic polynomial is a complete invariant. The following theorem says essentially the same thing.

**Theorem 9** Each $n \times n$ cyclic matrix $A$ with elements in $\mathbb{K}$ is similar to exactly one companion form with elements in $\mathbb{K}$ and two $n \times n$ cyclic matrices are similar if and only if they have the same characteristic polynomial.

**Remark 4** It is worth noting that any monic polynomial $\beta(s)$ of any degree, uniquely determines a companion matrix $B_C$ whose bottom row is the negative of the row vector of $\beta(s)$’s coefficients. This provides an especially easy way to construct a matrix $B_C$ whose characteristic polynomial is a given polynomial $\beta(s)$. Note in particular, that the roots of $\beta(s)$ need not be calculated in order to carry out this construction.
Theorem 9 states that any cyclic matrix $A$ is similar to exactly one companion form $A_C$ and moreover that $A_C$ is uniquely determined by the characteristic polynomial of $A$. Unfortunately not every matrix is cyclic.

**Example 49** For the matrix

$$A = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

and any vector $x \in \mathbb{K}^2$, $Ax = 3x$ which is a scalar multiple of $x$. Thus there is no vector in $\mathbb{K}^2$ which generates $\mathbb{K}^2$ and so $A$ cannot be cyclic. Another reason $A$ fails to be cyclic is because $A$’s minimal polynomial $s - 3$ is not equal to its characteristic polynomial $(s - 3)^2$.

To treat in detail the generalization of the preceding ideas to matrices which are not cyclic, is beyond the scope of this course. What we shall do instead is to state without proof the main technical result applicable to the noncyclic case. The reader interested in more detail should consult a good text on matrix algebra.

**Theorem 10 (Rational Canonical Decomposition)** For fixed $n > 0$, let $A_{n \times n}$ be a given matrix with elements in $\mathbb{K}$. There exists a positive integer $m$ and $m$ nonzero cyclic subspaces $X_i$, $i \in \{1, 2, \ldots, m\}$, with minimal polynomials $\alpha_i(s)$, $i \in \{1, 2, \ldots, m\}$ respectively, such that $\alpha_1(s)$ is the minimum polynomial of $A$, $\alpha_i(s)$ divides $\alpha_{i-1}(s)$ for $i \in \{2, 3, \ldots, m\}$, and

$$\mathbb{K}^n = X_1 \oplus X_2 \oplus \cdots \oplus X_m$$

Moreover if $\bar{X}_1$, $\bar{X}_2$, $\ldots$, $\bar{X}_m$ is any set of subspaces with the preceding properties, then $\bar{m} = m$ and the minimal polynomial of $\bar{X}_i$ equals $\alpha_i(s)$, $i \in \{1, 2, \ldots, m\}$.

A constructive procedure for finding the component subspaces $X_i$ and the $\alpha_i(s)$ which satisfy the conditions of the preceding theorem can be found in [1], Chapter 7, Section 4.

The first implication of the Rational Canonical Decomposition Theorem is that for any given $n \times n$ matrix $A$, one can calculate a unique set of polynomials $\{\alpha_1(s), \alpha_2(s), \ldots, \alpha_m(s)\}$, which is ordered so that each $\alpha_i(s)$ divides its predecessor. The $\alpha_i(s)$ are called the invariant factors of $A$.

The second implication of the theorem is that $A$ is similar to a block diagonal matrix

$$A^* = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_m \end{bmatrix}_{n \times n}$$

whose diagonal blocks are companion matrices, $\alpha_i(s)$ being the characteristic polynomial of $A_i$. $A^*$ is called the rational canonical form of $A^*$.

The block diagonal structure of $A^*$ is a direct consequence of the fact that $\mathbb{K}^n$ is the direct sum of the $A$-invariant subspaces $X_i$, $i \in \{1, 2, \ldots, m\}$. Each $A_i$ is a cyclic matrix with characteristic polynomial $\alpha_i(s)$ because $X_i$ is a cyclic subspaces with minimal polynomial $\alpha_i(s)$. Each $A_i$ is actually a companion form because of Theorem 9.

It can be shown that the invariant factors of $A$ are similarity invariants. Since each $n \times n$ matrix $A$ is similar to exactly one rational canonical form and since each canonical form is uniquely determined by $A$’s invariant factors, we have a criterion for deciding when two matrices are similar.
Theorem 11 Two \( n \times n \) matrices with elements in \( \mathbb{K} \) are similar if and only if they have the same invariant factors.

The block diagonal structure of \( A^* \) implies that the characteristic polynomial of \( A^* \) (and hence \( A \)) is equal to the product of the invariant factors of \( A \). That is

\[
\alpha(s) = \alpha_1(s)\alpha_2(s) \cdots \alpha_m(s)
\]

where \( \alpha(s) \) is \( A \)'s characteristic polynomial. Since \( \alpha_1(s) \) is the minimum polynomial of \( A \), it follows that the minimum polynomial of \( A \) divides the characteristic polynomial of \( A \), which is something we also concluded earlier by appealing to the Cayley-Hamilton Theorem. Since \( \alpha_1(A) = 0 \), it must be that \( \alpha(A) = 0 \) which is a restatement of the Cayley-Hamilton Theorem. In other words, the Cayley-Hamilton Theorem is a consequence of the Rational Canonical Decomposition Theorem. Clear?

Example 50 The matrix

\[
A = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
0 & -2 & -3 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & -2
\end{bmatrix}
\]

is in rational canonical form and its invariant factors are \( \alpha_1(s) = s^3 + 3s^2 + 2s = s(s + 1)(s + 2) \) and \( \alpha_2(s) = s^2 + 2s = s(s + 2) \). \( A \)'s characteristic polynomial and minimal polynomial are \( s^3(s + 1)(s + 2)^2 \) and \( s(s + 1)(s + 2) \) respectively, and its spectrum is \( \{0, 0, -1, -2, -2\} \). If \( B \) is any matrix similar to \( A \) then \( B \) must have the same invariant factors; conversely, if \( B \)'s invariant factors are \( \alpha_1(s) \) and \( \alpha_2(s) \), then \( B \) must be similar to \( A \).

7.9 Coprime Decompositions

While the rational canonical form is useful in certain applications, there is another canonical form which is even more useful - especially in the characterization of \( e^{tA} \). To develop this canonical form, we need a several new ideas which we now explain.

Again let \( A \) be a given \( n \times n \) matrix with elements in \( \mathbb{K} \). Let \( \alpha(s) \) be either the minimal or characteristic polynomial of \( A \). Suppose that \( \alpha(s) \) can be written as

\[
\alpha(s) = \alpha_1(s)\alpha_2(s)
\]

where \( \alpha_1(s) \) and \( \alpha_2(s) \) are monic polynomials which are at the same time coprime; i.e.,

\[
\gcd\{\alpha_1(s), \alpha_2(s)\} = 1
\]

Suppose we define

\[
\mathcal{X}_i \triangleq \ker \alpha_i(A), \quad i \in \{1, 2\}
\]

Thus \( \mathcal{X}_i \) is the set of all \( x \in \mathbb{K}^n \) such that \( \alpha_i(A)x = 0 \).

Claim 1: \( A\mathcal{X}_i \subset \mathcal{X}_i, \quad i \in \{1, 2\} \)

To show that this is so, fix \( i \in \{1, 2\} \) and \( x \in \mathcal{X}_i \). Then \( \alpha_i(A)x = 0 \) so \( A\alpha_i(A)x = 0 \). Since \( \alpha_i(A) \) and \( A \) commute, we can write \( \alpha_i(A)Ax = 0 \) or equivalently \( Ax \in \ker \alpha_i(A) \). Therefore \( Ax \in \mathcal{X}_i \). Since this must
Therefore, for every $x \in X_i$, claim 1 must be true. 

Note that this proof has not made use of any special properties of the $\alpha_i(s)$. We have therefore incidentally proved that for any polynomial $\beta(s)$, kernel $\beta(A)$ is an $A$-invariant subspace.

### Claim 2: $X_1 + X_2 = \mathbb{K}^n$

To establish this claim we need to make use of (7.35). In particular, the coprimeness assumption insures the existence of polynomials $\beta_1(s)$ and $\beta_2(s)$ such that

$$\beta_1(s)\alpha_1(s) + \beta_2(s)\alpha_2(s) = 1 \tag{7.37}$$

This is a consequence of Proposition 3 which appears at the very beginning of these notes. Observe that (7.37) implies the matrix identity

$$\beta_1(A)\alpha_1(A) + \beta_2(A)\alpha_2(A) = I_{n \times n} \tag{7.38}$$

If this surprises you, check it out for (say) $\alpha_1(s) = s^2 + s + 1$, $\alpha_2(s) = s^2 + 2$, $\beta_1(s) = -\frac{1}{2}(s+1)$, $\beta_2(s) = s+2$, and any $A$ of any size you like.

To proceed, let $x$ be any vector in $\mathbb{K}^n$. Then from (7.38)

$$x = \beta_1(A)\alpha_1(A)x + \beta_2(A)\alpha_2(A)x \tag{7.39}$$

Thus if we define

$$x_1 = \beta_2(A)\alpha_2(A)x \quad \text{and} \quad x_2 = \beta_1(A)\alpha_1(A)x \tag{7.40}$$

Then

$$x = x_1 + x_2 \tag{7.41}$$

To prove claim 2 it is enough to show that $x_i \in X_i, \ i \in \{1, 2\}$. Since $\alpha_1(s)\alpha_2(s)$ is the minimal or characteristic polynomial of $A$

$$\alpha_1(A)\alpha_2(A) = 0 \tag{7.42}$$

From this and definitions of the $x_i$ in (7.40) it follows that

$$\begin{align*}
\alpha_1(A)x_1 &= \alpha_1(A)\beta_2(A)\alpha_2(A)x = \beta_2(A)\alpha_1(A)\alpha_2(A)x = 0 \\
\alpha_2(A)x_2 &= \alpha_2(A)\beta_1(A)\alpha_1(A)x = \beta_1(A)\alpha_1(A)\alpha_2(A)x = 0
\end{align*}$$

Therefore $x_i \in \text{kernel } \alpha_i(A) = X_i, \ i \in \{1, 2\}$. 

### Claim 3: $X_1 \oplus X_2 = \mathbb{K}^n$

In view of claim 2, to establish claim 3 we need only show that $X_1 \cap X_2 = 0$. For this, pick $x \in X_1 \cap X_2$ and let $x_1$ and $x_2$ be defined as in (7.40). Since $x \in X_1 \cap X_2$, it must be true that $x \in X_2$ so $\alpha_2(A)x = 0$. From this and (7.40) it follows that $x_1 = 0$. By similar reasoning $x_2 = 0$. Therefore by (7.41), $x = 0$. Thus $x \in X_1 \cap X_2$ implies $x = 0$ so $X_1 \cap X_2 = 0$. Therefore claim 3 is true.

Putting claims 1 and 3 together we see that if $\alpha_1(s)\alpha_2(s)$ is a coprime factorization of either the minimal or characteristic polynomial of $A$, and if $X_i \triangleq \text{kernel } \alpha_i(A), i \in \{1, 2\}$, then $X_1$ and $X_2$ are independent $A$-invariant subspaces whose direct sum equals $\mathbb{K}^n$. In view of the discussion in section 7.3.1, we can therefore conclude that $A$ must be similar to a block diagonal matrix of the form

$$
\begin{bmatrix}
A_1 & 0 \\
0 & A_2
\end{bmatrix}
$$

where $A_i$ is a restriction of $A$ to $X_i$. The $T$ matrix in the similarity transformation $A \mapsto T^{-1}AT$ which achieves this is of the form $T = [X_1 \ X_2]$ where for each $i$, $X_i$ is a basis matrix for $X_i$. In addition, it can
be shown that if \( \alpha_1(s)\alpha_2(s) \) is the characteristic \{minimal\} polynomial of \( A \), then for each \( i \), \( \alpha_i(s) \) is the characteristic \{minimal\} polynomial of \( A_i \). These findings readily generalize to factorizations of the minimal or characteristic of \( A \) consisting of any number of pairwise coprime factors:

**Theorem 12 (Coprime Decomposition)** Let \( A \) be any matrix in \( \mathbb{K}^{n \times n} \) and let \( \alpha(s) \) denote either its characteristic or minimal polynomial. If \( \alpha_1(s), \alpha_2(s), \ldots, \alpha_m(s) \) are monic, polynomials satisfying

\[
\gcd\{\alpha_i(s), \alpha_j(s)\} = 1, \forall \ i \neq j
\]

and

\[
\alpha(s) = \alpha_1(s)\alpha_2(s)\cdots\alpha_m(s),
\]

and if \( X_i \) is a basis matrix for kernel \( \alpha_i(A) \), then the matrix

\[
T \equiv [X_1 \ X_2 \ \cdots \ X_m]
\]

is nonsingular and

\[
T^{-1}AT = \begin{bmatrix}
A_1 & 0 & \cdots & 0 \\
0 & A_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & A_m
\end{bmatrix}
\]

Moreover, if \( \alpha(s) \) is the characteristic \{minimal\} polynomial of \( A \), then for \( i \in \{1, 2, \ldots, m\} \), \( \alpha_i(s) \) is the characteristic \{minimal\} polynomial of \( A_i \).

Suppose \( A \) has \( n \) distinct eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_n \). Then \( A \)'s characteristic polynomial can be written as a product of the \( n \) pairwise coprime polynomials \( \alpha_i(s) \equiv s - \lambda_i, \ i \in \{1, 2, \ldots, n\} \). In this case, the kernel \( \alpha_i(A) \) are eigenspaces and each \( A_i = \lambda_i \). We can thus state the following.

**Corollary 1** If \( A_{n \times n} \) has \( n \) distinct eigenvalues, then \( A \) is semisimple and thus diagonalizable by a similarity transformation.

**Remark 5** At this point we know that any matrix \( A_{n \times n} \) with \( n \) distinct eigenvalues is both semisimple and cyclic. Unfortunately the converse implications are not true; i.e., there are semisimple matrices and cyclic matrices which do not have distinct eigenvalues. For example, the matrix

\[
A = \begin{bmatrix}
3 & 1 \\
0 & 3
\end{bmatrix}
\]

is cyclic but does not have distinct eigenvalues. And the matrix

\[
A = \begin{bmatrix}
3 & 0 \\
0 & 3
\end{bmatrix}
\]

is semisimple but does not have distinct eigenvalues. Here’s a tough question: Is there a matrix which does not have distinct eigenvalues which is both semisimple and cyclic? More generally, is the set of \( n \times n \) matrices which have distinct eigenvalues the same as the set of \( n \times n \) matrices which are both semisimple and cyclic?
7.9.1 The Jordan Normal Form

By making use of the Rational Canonical and Coprime Decomposition Theorems as they apply to any given matrix \( A_{n \times n} \), it is possible to construct an \( n \times n \) matrix \( T \) which transforms \( A \) via similarity into a block diagonal matrix whose diagonal blocks have the smallest sizes possible. The type of matrix we’re talking about is call a “Jordan Normal Form.” The aim of this section is to explain what the Jordan Normal Form is and to very briefly outline the ideas upon which the construction of the corresponding transforming matrix \( T \) is based. In what follows, \( A \) is any given \( n \times n \) matrix with elements in \( \mathbb{K} \) and \( \alpha(s) \) is its characteristic polynomial.

As a first step, let us use the fact that any polynomial with coefficients in either \( \mathbb{R} \) or \( \mathbb{C} \) can be written as a product of powers of prime polynomials in \( \mathbb{C}[s] \). For our purposes this means that \( \alpha(s) \) can be written as
\[
\alpha(s) = (s - \lambda_1)^{d_1}(s - \lambda_2)^{d_2} \cdots (s - \lambda_q)^{d_q}
\]
where the \( \lambda_i \) are distinct roots of \( \alpha(s) \) over \( \mathbb{C} \), \{i.e., \( \lambda_i \neq \lambda_j \) if \( i \neq j \} \) and \( d_i \) is \( \lambda_i \)'s algebraic multiplicity. Meanwhile the geometric multiplicity of \( \lambda_i \) is the dimension of the eigenspace of \( \lambda_i \). Recall from elementary algebra that such a factorization is always possible and is unique up to a reordering of the distinct roots of \( \alpha(s) \). The \( \lambda_i \) are, of course, the distinct eigenvalues of \( A \) and \( d_1 + d_2 + \cdots + d_q = n \).

As a second step, let us use the Coprime Decomposition Theorem to find a nonsingular matrix \( R \) such that
\[
R^{-1}AR = \begin{bmatrix}
A_1 & 0 & \cdots & 0 \\
0 & A_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & A_q
\end{bmatrix}
\]  
(7.43)

where for each \( i \in \{1, 2, \ldots, q\} \), \( A_i \) is a square matrix with characteristic polynomial \((s - \lambda_i)^{d_i}\).

As a third step, for each \( i \in \{1, 2, \ldots, q\} \) let us use the Rational Canonical Decomposition Theorem to find a \( d_i \times d_i \) matrix \( P_i \) such that
\[
P_i^{-1}A_iP_i = \begin{bmatrix}
B_{i1} & 0 & \cdots & 0 \\
0 & B_{i2} & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & B_{im_i}
\end{bmatrix}
\]  
(7.44)

where the matrix on the right is the rational canonical form of \( A_i \). \{We’ve not really explained exactly how to do this, but this is not so important at this point.\} Recall that each diagonal block in a rational canonical form is a cyclic matrix whose characteristic polynomial divides that of its predecessor. Since \((s - \lambda_i)^{d_i}\) is the characteristic polynomial of \( A_i \), this means that the characteristic polynomial of the \( B_{ij} \) must be of the form
\[
\beta_{ij}(s) = (s - \lambda_i)^{n_{ij}}
\]
where the \( n_{ij} \) are positive integers satisfying
\[
n_{i1} \geq n_{i2} \geq \cdots \geq n_{im_i}
\]
and
\[
d_i = n_{i1} + n_{i2} + \cdots + n_{im_i}
\]

The forth step is to construct matrices \( S_{ij} \) for which
\[
S_{ij}^{-1}B_{ij}S_{ij} = J_{ij}
\]  
(7.45)
The matrix $J_{ij}$ is called the *Jordan Block* of $(s - \lambda_i)^{n_{ij}}$. The existence of $S_{ij}$ is a consequence of three things:

1. $B_{ij}$ and $J_{ij}$ both have the same characteristic polynomial, namely $(s - \lambda_i)^{n_{ij}}$.

2. $B_{ij}$ and $J_{ij}$ are both cyclic. \{Cyclicity of $J_{ij}$ can be established by showing that with $x$ the $n_{ij}$th unit vector in $\mathbb{K}^{n_{ij}}$, the matrix $[x \ J_{ij}x \ \cdots \ J_{ij}^{n_{ij}-1}x]$ is nonsingular; this in turn can be verified by checking that this matrix is upper triangular with one's on the main diagonal.\}

3. All cyclic matrices with the same characteristic polynomial are similar \{c.f., Theorem 9\}.

Now suppose we define

$$T \triangleq R \begin{bmatrix} P_1S_1 & 0 & \cdots & 0 \\ 0 & P_2S_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_qS_q \end{bmatrix}$$

where

$$S_i = \begin{bmatrix} S_{i1} & 0 & \cdots & 0 \\ 0 & S_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & S_{im_i} \end{bmatrix}$$

Then using (7.43)-(7.45) we arrive at the similarity transformation

$$T^{-1}AT = J$$

where $J$ is the block diagonal matrix

$$J \triangleq \begin{bmatrix} J_{11} & J_{12} & \cdots & \cdots \\ J_{12} & J_{22} & \cdots & \cdots \\ \cdots & \cdots & \ddots & \cdots \\ \cdots & \cdots & \cdots & J_{mm} \end{bmatrix}$$

where

$$J_{ij} = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_i & \ddots & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \cdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix}_{n_{ij} \times n_{ij}} \quad (7.46)$$
J is called the Jordan Normal Form of A. J’s diagonal blocks are each of the form (7.46) and there is one such block for each polynomial \((s - \lambda_i)^{n_{ij}}, \quad i \in \{1, 2, \ldots, q\}, \quad j \in \{1, 2, \ldots, m_i\}\). The \((s - \lambda_i)^{n_{ij}}\) are called the elementary divisors of A. They are uniquely determined by A and in turn uniquely determine A’s Jordan Normal Form up to the ordering of its diagonal blocks. Given all this, it is not very difficult understand why two \(n \times n\) matrices are similar if and only if they both have the same set of elementary divisors.

Note that in the special case when all of A’s elementary divisors are of degree one, the Jordan Normal Form of A is nothing more than a diagonal matrix whose diagonal entries are A’s eigenvalues. Thus semisimple matrices are precisely those matrices whose elementary divisors all have degree one. Unfortunately the problem of explicitly computing the elementary divisors of a matrix is a bit of an exercise. There is no simple way to conclude that a given square matrix is semisimple.

As with invariant factors and rational canonical forms, our purpose has here has been only to make the reader aware of the existence of elementary divisors and the Jordan Normal Form. Those individual wishing to learn more about this topic should consult a good matrix algebra text or reference such as [1], Chapter 7.

**Example 51** The matrix

\[
A = \begin{bmatrix}
-2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 3 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & -2 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -2 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & -3 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & -3
\end{bmatrix}
\]

is in Jordan Normal Form and its elementary divisors are \((s + 2)^3\), \((s - 3)\), \((s + 2)^2\), and \((s + 3)^2\).

---

### 7.10 The Structure of the Matrix Exponential

For sure the main application of the Jordan Normal Form of a matrix A is in characterizing A’s state transition matrix. The aim of this section is to explain this characterization, and to discuss some of its implications.

Let us begin by assuming that A is a given \(n \times n\) matrix, that J is its Jordan Normal Form and that T is a nonsingular matrix such that

\[A = TJT^{-1}\]

Then as we’ve already explained at the beginning of the chapter

\[e^{tA} = Te^{tJ}T^{-1}\]

Now since J is a block diagonal matrix of the form

\[
J = \begin{bmatrix}
J_1 & 0 & \cdots & 0 \\
0 & J_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & J_m
\end{bmatrix}
\]
we can write
\[ e^{tJ} = \begin{bmatrix} e^{tJ_1} & 0 & \cdots & 0 \\ 0 & e^{tJ_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{tJ_m} \end{bmatrix} \]

Thus to characterize the kinds of time functions which can appear in \( e^{tA} \), what we really need to do it to characterize the kinds of time functions which appear in matrix exponentials of the form \( e^{tJ_i} \) where \( J_i \) is a typical Jordon Block. Whatever these time functions are, they are all going to appear in \( e^{tA} \) in linear combination with similar time functions from \( J_i \)'s other Jordon Blocks. This is because for any matrix \( B \), each entry of the matrix \( TBT^{-1} \) is a linear combination of the entries of the matrix \( B \).

Now it is quite clear that if
\[ J_i = [\lambda_i]_{1 \times 1} \]
then \( e^{tJ_i} \) is simply the scalar exponential
\[ e^{tJ_i} = e^{t\lambda_i} \]

Consider next the case when
\[ J_i = \begin{bmatrix} \lambda_i & 1 \\ 0 & \lambda_i \end{bmatrix}_{2 \times 2} \]
Then
\[ e^{tJ_i} = \begin{bmatrix} e^{t\lambda_i} & te^{t\lambda_i} \\ 0 & e^{t\lambda_i} \end{bmatrix} \]
This can easily be verified by checking to see that with \( e^{tJ_i} \) so defined,
\[ \frac{d}{dt} e^{tJ_i} = J_i e^{tJ_i}, \quad \text{and} \quad e^{tJ_i}\big|_{t=0} = I \]

Similarly, for
\[ J_i = \begin{bmatrix} \lambda_i & 1 & 0 \\ 0 & \lambda_i & 1 \\ 0 & 0 & \lambda_i \end{bmatrix} \]
one finds that
\[ e^{tJ_i} = \begin{bmatrix} e^{t\lambda_i} & te^{t\lambda_i} & \frac{t^2}{2} e^{t\lambda_i} \\ 0 & e^{t\lambda_i} & \frac{t}{2} e^{t\lambda_i} \\ 0 & 0 & e^{t\lambda_i} \end{bmatrix} \]

More generally, if
\[ J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_i & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix}_{m_i \times m_i} \]
then

\[ e^{t\lambda_i} = \begin{bmatrix}
    e^{t\lambda_i} & te^{t\lambda_i} & \cdots & \frac{t^{(m_i-2)}e^{t\lambda_i}}{(m_i-2)!} & \frac{t^{(m_i-1)}e^{t\lambda_i}}{(m_i-1)!} \\
    0 & e^{t\lambda_i} & \cdots & \frac{t^{(m_i-3)}e^{t\lambda_i}}{(m_i-3)!} & \frac{t^{(m_i-2)}e^{t\lambda_i}}{(m_i-2)!} \\
    0 & 0 & e^{t\lambda_i} & \cdots & \frac{t^{(m_i-4)}e^{t\lambda_i}}{(m_i-4)!} & \frac{t^{(m_i-3)}e^{t\lambda_i}}{(m_i-3)!} \\
    \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
    0 & 0 & 0 & \cdots & e^{t\lambda_i} & te^{t\lambda_i} \\
    0 & 0 & 0 & \cdots & 0 & e^{t\lambda_i}
\end{bmatrix}_{m_i \times m_i} \]

Thus we see that the Jordan Block of the elementary divisor \((s - \lambda_i)^{m_i}\) contains within it, linear combinations \{(actually just scalar multiples)} of the time functions

\[ e^{t\lambda_i}, te^{t\lambda_i}, \ldots, t^{(m_i-1)}e^{t\lambda_i} \]

We’ve arrived at the following characterization of the state transition matrix of \(A\).

**Theorem 13** Let \(A\) be an \(n \times n\) matrix with elementary divisors

\((s - \lambda_1)^{m_1}, (s - \lambda_2)^{m_2}, \ldots, (s - \lambda_q)^{m_q}\)

Then each of the entries of \(e^{tA}\) is a linear combination of the time functions

\[ e^{t\lambda_i}, te^{t\lambda_i}, \ldots, t^{(m_i-1)}e^{t\lambda_i}, \quad i \in \{1, 2, \ldots, q\} \]

Thus the qualitative behavior of \(e^{tA}\) and consequently all solutions to \(\dot{x} = Ax\), is completely characterized by \(A\)’s eigenvalues and the sizes of their Jordan blocks.

### 7.10.1 Real Matrices

Suppose \(A\) is real, that \(\lambda\) is one of its eigenvalues and that \(m\) is the size of its Jordan block. Then as we’ve just noted, linear combinations of the time functions

\[ e^{\lambda t}, te^{\lambda t}, \ldots, t^{m-1}e^{\lambda t} \]

must appear in \(e^{tA}\). Suppose \(\lambda\) is not real; i.e.,

\[ \lambda = a + j\omega \]

where \(a\) and \(\omega\) are real numbers and \(\omega \neq 0\). Then

\[ \lambda^* = a - jb \]

must also be an eigenvalue of \(A\) with multiplicity \(m\). Because \(A\) is real, it can be shown that the linear combinations of the time functions

\[ e^{\lambda t}, te^{\lambda t}, \ldots, t^{m-1}e^{\lambda t}, e^{\lambda^* t}, te^{\lambda^* t}, \ldots, t^{m-1}e^{\lambda^* t} \]

which appear in \(e^{tA}\) can also be written as linear combinations of time functions of the forms

\[ e^{at} \cos \omega t, e^{at} \sin \omega t, te^{at} \cos \omega t, te^{at} \sin \omega t, \ldots, t^{m-1}e^{at} \cos \omega t, t^{m-1}e^{at} \sin \omega t \]
Let $A$ be either real or complex valued. On the basis of Theorem 13 we can draw the following conclusions.

1. If all of $A$’s eigenvalues have negative real parts, then $\|e^{tA}\| \to 0$ exponentially fast, as $t \to \infty$. Such $A$ are said to be **asymptotic stability matrices**.

2. If at least one eigenvalue of $A$
   (a) has a positive real part $a$, then $\|e^{tA}\|$ must grow without bound, as fast as $e^{at}$, as $t \to \infty$.
   (b) has a zero real part and a Jordan block of size $m > 1$, then $\|e^{tA}\|$ must grow without bound as fast $t^{m-1}$, as $t \to \infty$.

In either case, such $A$ are said to be **unstable matrices**.

3. If all of $A$’s eigenvalues have nonpositive real parts, and those with zero real parts all have Jordan blocks of size 1, then $\|e^{tA}\|$ will remain finite as $t \to \infty$ but will not approach 0. Such $A$ are said to be **stable matrices**.

The key observation here is summarized as follows:

**Theorem 14** A $n \times n$ matrix $A$ is asymptotically stable; i.e.,
$$\lim_{t \to \infty} e^{tA} = 0$$
if and only if all of $A$’s eigenvalues have negative real parts.

### 7.11 Linear Recursion Equations

Just about everything we’ve discussed in the preceding section for the linear differential equation
$$\dot{x} = Ax,$$
extends an a natural way to a linear recursion equation of the form
$$x(t+1) = Ax(t), \quad t = 0, 1, 2, \ldots$$
For example, for the recursion equation we can write
$$x(t) = A^{t-\tau}x(\tau), \quad \forall t, \tau \geq 0$$
For this reason $A^{t-\tau}$ is called the {discrete-time} **state transition matrix** of $A$. Note that if $J$ is $A$’s Jordan normal form and $A = TJT^{-1}$, then
$$A^t = TJ^tT^{-1},$$
so the asymptotic behavior of $A^t$ is characterized by $A$’s eigenvalues, just as in the continuous time case discussed above. We leave it to the reader to verify the following:

---

3Note that for any finite integer $i \geq 0$, and any positive number $\mu$, $t^ie^{-\mu t} \to 0$ as $t \to \infty$. 
1. If all of $A$’s eigenvalues have magnitude less than 1, then $\|A^t\| \to 0$ exponentially fast, as $t \to \infty$. Such $A$ are said to be discrete-time asymptotic stability matrices.

2. If at least one eigenvalue of $A$
   
   (a) has magnitude $a$ greater than 1, then
   
   $\|A^t\|$
   
   must grow without bound, as fast as $|a|^t$, as $t \to \infty$.
   
   (b) has magnitude 1 and a Jordan block of size $m > 1$, then
   
   $\|A^t\|$
   
   must grow without bound as fast $t^{m-1}$, as $t \to \infty$.

In either case, such $A$ are said to be discrete-time unstable matrices.

3. If all of $A$’s eigenvalues have magnitudes no greater than 1, and those with magnitude 1 each have Jordan blocks of size 1, then

   $\|A^t\|$

   will remain finite as $t \to \infty$ but will not approach 0. Such $A$ are said to be discrete-time stable matrices.

The discrete-time version of Theorem 14 is as follows:

**Theorem 15** An $n \times n$ matrix $A$ is discrete-time asymptotically stable; i.e.,

$$\lim_{t \to \infty} A^t = 0$$

if and only if all of $A$’s eigenvalues have magnitudes less than 1.

**NOTE:** Throughout this section $t$ is a discrete variable taking values in the set of natural numbers \{0, 1, 2, 3, \ldots\}!

Later in these notes we will develop alternative methods for testing for discrete and continuous-time stability of a matrix $A$, which do not require one to compute the eigenvalues of $A$.

---

Note that for any finite integer $i \geq 0$, and any real number $\mu$ with magnitude less than 1, $t^i \mu^t \to 0$ as $t \to \infty$.  

---
Chapter 8

Inner Product Spaces

In Chapter 4 we briefly discussed the concepts of a normed vectorspace. We saw that by exploiting the idea of a norm, one could extend from the set of real numbers to normed vector spaces, a number of concepts from real analysis or calculus such as continuity, convergence, openness, boundness, etc. The concept of a normed space is a special case of a more general concept known as a “metric space.” On the other hand, normed spaces are generalizations of “inner product spaces.” The aim of this chapter is to briefly discuss some of the basic properties of inner product spaces.

8.1 Definition

Let \( V \) be a (possibly infinite dimensional) vector space over \( \mathbb{K} \). A function \((\cdot, \cdot)\) mapping domain \( V \times V \) into codomain \( \mathbb{K} \) is said to be an inner product on \( V \) if for all \( x, y, z \in \mathcal{X} \) and all \( \mu \in \mathbb{K} \)

1. \((x, y) = (y, x)^*\)
2. \((\mu x, y) = \mu^* (x, y)\)
3. \((x + y, z) = (x, z) + (y, z)\)

Here * denotes complex conjugation; i.e., for \( a, b \in \mathbb{R} \) and \( j = \sqrt{-1} \), \((a + jb)^* = a - jb\).

Let us note that requirements 1 to 3 have the following consequences:

2'. \((x, \mu y) = \mu(x, y)\)
3'. \((x, y + z) = (x, y) + (x, z)\)

What’s more, whether \( \mathbb{K} \) is \( \mathbb{R} \) or \( \mathbb{C} \), it follows from 1. that for every vector \( x \), the inner product \((x, x)\) is a real number. For our purposes, we shall impose the additional requirement that \((x, x)\) be positive definite; i.e.,

---

1 A metric space is a real or complex vector space \( \mathcal{X} \) on which is defined for every pair of vectors \( x \) and \( y \), a real, scalar-valued “distance function” or “metric” \( d(x, y) \) which is a measure of the “distance” between \( x \) and \( y \). For such a function to be a metric, it (i) must be nonnegative-valued, (ii) must equal 0 just in case \( x = y \), (iii) must satisfy \( d(x, y) = d(y, x) \) for all \( x, y \in \mathcal{X} \) and (iv) must satisfy the triangle inequality \( d(x, y) \leq d(x, z) + d(z, y) \) for every \( x, y, z \in \mathcal{X} \). If \( \mathbb{K} = \mathbb{R} \), then \( \mathcal{X} \) is called a unitary space with Hermitian metric. If \( \mathbb{K} = \mathbb{C} \), then \( \mathcal{X} \) is called a Euclidean space with Euclidean metric. The metric on a normed space is simply \( d(x, y) = ||x - y|| \).
4. \((x, x) > 0\) for all \(x \neq 0\)
5. \((x, x) = 0\) if \(x = 0\)

In view of properties 4 and 5 it makes sense to define the norm of \(x\) by

\[
\|x\| \overset{\Delta}{=} \sqrt{(x, x)}
\]

It can be shown that properties 1 - 5 make \(V\) into a normed vector space.

**Example 52** Suppose \(V = \mathbb{Q}^n\) and \((x, y)\) is defined so that

\[
(x, y) = (x^*)' y
\]

where as usual, prime denotes transpose. It is easy to verify that properties 1 - 5 hold and that \(V\) is a unitary space. Note that if

\[
x = \begin{bmatrix}
x_1 \\
x_2 \\
\vdots \\
x_n
\end{bmatrix}
\]

then

\[
\|x\|^2 = (x^*)' x = \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} |x_i|^2
\]

where \(|x_i|\) denotes the magnitude of \(x_i\).

**Example 53** Suppose \(V = \mathbb{R}^n\) and \((x, y)\) is defined so that

\[
(x, y) = x'y
\]

It is quite easy to see that properties 1 - 5 hold. Thus in this case \(V\) is a Euclidean space\(^2\).

### 8.1.1 Triangle Inequality

Note that in \(\mathbb{R}^2\), \(\|x - y\|\) can be thought of as the distance between the directed line segment \(x\) and the directed line segment \(y\) as shown in Figure 8.1.

![Figure 8.1](image)

---

\(^2\) The reader should not confuse the inner product of \(x\) and \(y\), namely the scalar \(x'y\), with the outer product of \(x\) and \(y\), namely the \(n \times n\) matrix \(xy'\). Outer products play an important role in quantum mechanics. Note for example that \(xy'zw' = (y'z)zw'\). Manipulations such as this arise in connection with Dirac notation.
If \( x, y \) and \( z \) are arbitrary vectors in \( \mathbb{R}^2 \), then we can make use of well known properties of triangles to conclude that the triangle inequality

\[
||x - y|| + ||y - z|| \geq ||x - z||
\]

is valid for any three vectors in any inner product space of any dimension \( \text{cf, Figure 8.2} \). The inequality can be derived directly from the defining properties 1 – 4 above.

![Figure 8.2:](image)

8.2 Orthogonality

By far the most important feature of an inner product space which distinguishes it from a more general normed space, is the concept of orthogonality. In the sequel we define orthogonality and discuss some of its consequences.

8.2.1 Orthogonal Vectors

Two vectors \( x \) and \( y \) in an inner product space are said to be orthogonal if their inner product equals zero. In other words

\[
\text{\( x \) and \( y \) orthogonal} \iff (x,y) = 0
\]

Note also that the identity

\[
||x + y||^2 = ||x||^2 + ||y||^2 + (x,y) + (y,x)
\]

implies that

\[
||x + y||^2 = ||x||^2 + ||y||^2
\]

whenever \( x \) and \( y \) are orthogonal. This equation, is valid for orthogonal vectors in any inner product space - it can be viewed as a generalization of Pythagoras’s Theorem for right triangles.

8.2.2 Orthonormal Sets

A set vectors \( Y \triangleq \{y_1, y_2, \ldots, y_n\} \) whose elements all satisfy \( ||y_i|| = 1 \), is said to be normalized. A set of vectors \( X \triangleq \{x_1, x_2, \ldots, x_m\} \) whose elements are pairwise orthogonal \( \text{i.e.,} \ (x_i, x_j) = 0, \ \forall i \neq j \) is called
an orthogonal set. It is possible to normalize and such orthogonal set, without changing its span, by simply replacing each \( x_i \) by the normalized vector

\[
\tilde{x}_i = \frac{1}{\|x_i\|} x_i
\]

Note that each such \( \tilde{x}_i \) satisfies

\[
\|\tilde{x}_i\| = \sqrt{\langle \tilde{x}_i, \tilde{x}_i \rangle} = \sqrt{\left( \frac{1}{\|x_i\|} x_i, \frac{1}{\|x_i\|} x_i \right)} = \sqrt{\frac{1}{\|x_i\|^2} \langle x_i, x_i \rangle} = 1
\]

so \( \tilde{X} \overset{\Delta}{=} \{ \tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m \} \) is both an orthogonal and a normalized set. Such sets are called orthonormal. \( X \) and \( \tilde{X} \) both span the same subspace because each \( \tilde{x}_i \) is just a nonzero scalar multiply of \( x_i \).

8.3 Gram’s Criterion

Let \( \{x_1, x_2, \ldots, x_m\} \) be a finite set of vectors in a possibly infinite dimensional inner product space \( X \). Our aim is to develop a simple test for deciding whether or not \( \{x_1, x_2, \ldots\} \) is a linearly independent set.

Suppose that \( \{x_1, x_2, \ldots\} \) is linearly dependent. Then there must be numbers \( a_i \), not all zero, such that

\[
a_1 x_1 + a_2 x_2 + \cdots + a_m x_m = 0 \quad (8.1)
\]

Clearly

\[
\langle x_i, \{a_1 x_1 + a_2 x_2 + \cdots + a_m x_m\} \rangle = 0, \quad i \in \{1, 2, \ldots, m\}
\]

Therefore

\[
a_1 \langle x_i, x_1 \rangle + a_2 \langle x_i, x_2 \rangle + \cdots + a_m \langle x_i, x_m \rangle = 0 \quad i \in \{1, 2, \ldots, m\} \quad (8.2)
\]

In matrix form, this is equivalent to

\[
G a = 0 \quad (8.3)
\]

where

\[
G \overset{\Delta}{=} [ (x_i, x_j) ]_{m \times m} \quad (8.4)
\]

and

\[
a = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}
\]

But \( a \neq 0 \) by hypothesis, so \( G \) must have linearly dependent columns. Since \( G \) is a square matrix, \( G \) must therefore be singular. In other words, linear dependence of the set \( \{x_1, x_2, \ldots, x_m\} \) implies that \( G \) must be singular.

Conversely, suppose that the matrix \( G \) defined by (8.4) is singular. Then (8.3) must have a nonzero solution \( a \) so there must be numbers \( a_i \), not all zero, for which (8.2) holds. Multiplying the ith equation in (8.2) by \( a_i^* \) we obtain

\[
a_i^* a_1 \langle x_i, x_1 \rangle + a_i^* a_2 \langle x_i, x_2 \rangle + \cdots + a_i^* a_m \langle x_i, x_m \rangle = 0
\]

Therefore

\[
(a_i x_i, \{a_1 x_1 + a_2 x_2 + \cdots + a_m x_m\}) = 0, \quad i \in \{1, 2, \ldots, m\}
\]

Summing these \( m \) equations thus yields

\[
\|a_1 x_1 + a_2 x_2 + \cdots a_m x_m\|^2 = 0
\]
or equivalently

\[ a_1x_1 + a_2x_2 + \cdots + a_mx_m = 0 \]

Since the \( a_i \) are not all zero, \( \{x_1, x_2, \ldots, x_m\} \) must be a dependent set. In other words, if \( G \) is singular, \( \{x_1, x_2, \ldots, x_m\} \) must be a dependent set.

The matrix \( G \) defined by (8.4) is called the **Gramian** of \( \{x_1, x_2, \ldots, x_m\} \). We’ve proved the following.

**Theorem 16** A set of vectors \( \{x_1, x_2, \ldots, x_m\} \) in an inner product space is linearly independent if and only if its Gramian is a nonsingular matrix.

**Example 54** Note that with the inner product \( (x, y) \triangleq x'y \), the Gramian of the columns of the real matrix \( A_{m \times n} \) is \( A'A \). Thus the columns of \( A \) are linearly independent if and only if the \( n \times n \) matrix \( A'A \) is nonsingular.

**Example 55** Let \( f_1(t), f_2(t), \ldots, f_n(t) \) be continuous, real scalar-valued functions of \( t \) defined on the interval \([0, 1]\). The \( f_i \) can be viewed as vectors in the infinite dimensional vector space of continuous, real scalar functions defined on \([0, 1]\). We can make this vector space into a Euclidean space by introducing the inner product

\[
(f, g) \triangleq \int_0^1 f(t)g(t)dt
\]

The Gramian of \( \{f_1, f_2, \ldots, f_n\} \) is therefore the \{constant\} matrix

\[
G = \begin{bmatrix}
\int_0^1 f_1(t)f_1(t)dt & \int_0^1 f_1(t)f_2(t)dt & \cdots & \int_0^1 f_1(t)f_n(t)dt \\
\int_0^1 f_2(t)f_1(t)dt & \int_0^1 f_2(t)f_2(t)dt & \cdots & \int_0^1 f_2(t)f_n(t)dt \\
\vdots & \vdots & \ddots & \vdots \\
\int_0^1 f_n(t)f_1(t)dt & \int_0^1 f_n(t)f_2(t)dt & \cdots & \int_0^1 f_n(t)f_n(t)dt
\end{bmatrix}
\]

The set \( \{f_1, f_2, \ldots, f_n\} \) is thus linearly independent if and only if \( \det G \neq 0 \).

### 8.4 Cauchy-Schwartz Inequality

Suppose that \( x \) and \( y \) are vectors in an inner product space. Observe that the inequality

\[
\|\|y\|^2x - (y, x)y\|^2 \geq 0
\]

expands out to

\[
\|y\|^4\|x\|^2 - \|y\|^2((x, (y, x)y) + ((y, x)y, x)) + \|y\|^2\|x\|^2 \geq 0
\]

But since \((x, (y, x)y) = \|y\|^2 = ((y, x)y, x)\), (8.5) can be rewritten as

\[
\|y\|^2(\|y\|^2\|x\|^2 - \|y, x\|^2) \geq 0
\]

Whether \( \|y\| = 0 \) or not, (8.6) implies that

\[
\|y\|^2\|x\|^2 \geq \|y, x\|^2
\]

which when rewritten as

\[
\|y\|\|x\| \geq \|y, x\|
\]

is called the **Cauchy-Schwartz Inequality**.

---

³As always with an inner product space, \( d(f, g) \triangleq \|f - g\| \) and \( \|f\| \triangleq \sqrt{(f, f)} \).
8.5 Orthogonalization of a Set of Vectors

Let \( \{x_1, x_2, \ldots, x_m\} \) be a set of vectors in an inner product space and write \( \mathcal{S} \triangleq \text{span} \{x_1, x_2, \ldots, x_m\} \). Our aim is to use to construct from the \( x_i \), another set \( \{y_1, y_2, \ldots, y_m\} \) which also spans \( \mathcal{S} \) and which, in addition, is orthogonal. The procedure for doing this is as follows.

1. Set \( y_1 = x_1 \)

2. Set \( y_2 = x_2 - \frac{(y_1, x_2)}{(y_1, y_1)} y_1 \)

i. Set \( y_i = x_i - \frac{(y_1, x_i)}{(y_1, y_1)} y_1 - \frac{(y_2, x_i)}{(y_2, y_2)} y_2 - \cdots - \frac{(y_{i-1}, x_i)}{(y_{i-1}, y_{i-1})} y_{i-1} \)

Note that \( (y_1, y_2) = (y_1, x_2) - \frac{(y_1, x_2)}{(y_1, y_1)} (y_1, y_1) = 0 \)

so \( y_1 \) and \( y_2 \) are orthogonal. Similarly, for \( j < i \)

\[
(y_j, y_i) = \frac{(y_1, x_i)}{(y_1, y_1)} (y_j, y_1) - \frac{(y_2, x_i)}{(y_2, y_2)} (y_j, y_2) - \cdots - \frac{(y_{i-1}, x_i)}{(y_{i-1}, y_{i-1})} (y_j, y_{i-1})
\]

Since \( y_j \) is orthogonal to all of the \( y_k \) for \( k < i \) and \( k \neq j \), it follows that

\[
(y_j, y_i) = (y_j, x_i) - \frac{(y_j, x_i)}{(y_j, y_j)} (y_j, y_j) = 0
\]

Therefore the set \( \{y_1, y_2, \ldots, y_m\} \) is orthogonal.

Let us note finally that

\[\text{span} \{y_1\} = \text{span} \{x_1\}\]
\[\text{span} \{y_1, y_2\} = \text{span} \{x_1, x_2\}\]
\[\text{span} \{y_1, y_2, \ldots, y_i\} = \text{span} \{x_1, x_2, \ldots, x_i\}\]

Thus

\[\text{span} \{y_1, y_2, \ldots, y_m\} = \text{span} \{x_1, x_2, \ldots, x_m\}\]

**Example 56** Let \( \mathcal{X} \) be the vector space of all piecewise-continuous, real-valued scalar functions defined on the closed interval \([-1, 1]\). The inner product

\[
(f, g) \triangleq \int_{-1}^{1} f(t)g(t)dt, \quad \forall f, g \in \mathcal{X}
\]
makes $\mathcal{X}$ a Euclidean space. Using Gram’s test, one can then verify that $\{1, t, t^2, \ldots, t^m\}$ is a linearly independent set of vectors in this space. This set can be orthogonalized as follows:

Set

$$y_1 \triangleq 1$$

$$y_2 \triangleq t - \left(\int_{-1}^{1} \tau d\tau \right) y_1 = t$$

$$y_3 \triangleq t^2 - \left(\int_{-1}^{1} \tau^2 d\tau \right) y_1 - \left(\int_{-1}^{1} \tau^3 d\tau \right) y_2$$

etc. It is easy to verify that the $y_i$ coincide up to scale factor with the well known Legendre polynomials

$$\frac{1}{2^{(i-1)}(i-1)!} \frac{d^{(i-1)}}{dt^{(i-1)}} \{t^2 - 1\}^{(i-1)}, \quad i \in \{1, 2, \ldots\}$$

By changing the definition of the inner product on $\mathcal{X}$, we can develop other sets of orthogonal functions. For example, by orthogonalizing the set $\{1, t, \ldots\}$ using the inner product

$$(f, g) \triangleq \int_{-1}^{1} \frac{1}{\sqrt{1 - t^2}} f(t)g(t)dt, \quad f, g \in \mathcal{X}$$

one obtains the Chebyshev polynomials

$$\frac{1}{2^{(i-1)}} \cos(i \arccos(t)), \quad i \in \{1, 2, \ldots\}$$

### 8.6 Orthogonal Projections

Let $\{x_1, x_2, \ldots, x_m\}$ be a finite set of orthogonal vectors in a (possibly infinite dimensional) inner product space $\mathcal{X}$. Let $\mathcal{S} \triangleq \text{span} \{x_1, x_2, \ldots, x_m\}$. Note that $\mathcal{S}$ is finite dimensional because it is spanned by a finite set of vectors. Let $x$ be any given vector in $\mathcal{X}$. We claim that there are unique vectors $s$ and $y$ such that

1. $x = s + y$
2. $s \in \mathcal{S}$
3. $y$ is orthogonal to every vector in $\mathcal{S}$; i.e., $(y, x_i) = 0$, $i \in \{1, 2, \ldots, m\}$.

We call $s$ the orthogonal projection of $x$ on $\mathcal{S}$.

To construct $s$, we first form the projection of $x$ on the span of each $x_i$:

$$(x, x_i) x_i / (x_i, x_i) x_i$$

The definition of $s$ is then simply

$$s \triangleq \sum_{i=1}^{m} \frac{(x, x_i)}{(x_i, x_i)} x_i$$

Note that property 2 above must automatically be true because $s$ is a linear combination of the $x_i$. 

Set \( y \triangleq x - s \) thereby establishing property 1. Note that for each \( j \in \{1, 2, \ldots, m\} \),

\[
(y, x_j) = (x, x_j) - (s, x_j) = (x, x_j) - \left( \sum_{i=1}^{m} \frac{(x, x_i)}{(x_i, x_i)} x_i, x_j \right)
\]

But \( \{x_1, x_2, \ldots, x_m\} \) is an orthogonal set so \((x_i, x_j) = 0\) for \( i \neq j \). Therefore

\[
(y, x_j) = (x, x_j) - \frac{(x, x_j)}{(x_j, x_j)} (x_j, x_j) = 0
\]

Thus \( y \) is orthogonal to each vector in a basis for \( S \). This implies that \( y \) is orthogonal to each vector in \( S \). In other words, property 3 above is true.

**Example 57** In \( \mathbb{R}^3 \) suppose that \( S \triangleq \text{span} \{x_1, x_2\} \) where

\[
x_1 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix} \quad \text{and} \quad x_2 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}
\]

Then \( x_1 \) and \( x_2 \) are orthogonal since \( x_1^\top x_2 = 0 \). If

\[
x \triangleq \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}
\]

then

\[
s = \left( x_1^\top x_1 \right) x_1 + \left( x_2^\top x_2 \right) x_2 = \frac{1}{2} x_1 + \frac{1}{2} x_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \\ 0 \end{bmatrix}
\]

and

\[
y = x - s = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ 0 \end{bmatrix}
\]

### 8.7 Bessel’s Inequality

Now suppose \( \{x_1, x_2, \ldots, x_m\} \) is an orthonormal set in some inner product space \( X \) - and let \( x \) be any vector in \( X \). Then the orthogonal projection of \( x \) on the subspace \( S \triangleq \text{span} \{x_1, x_2, \ldots, x_m\} \) is

\[
s = \sum_{i=1}^{m} a_i x_i
\]

where

\[
a_i \triangleq (x, x_i)
\]
We can write
\[ ||x||^2 = ||s + x - s||^2 = ||s||^2 + ||x - s||^2 + (s, x - s) + (x - s, s) \]
But \( s \) and \( x - s \) are orthogonal so
\[ ||x||^2 = ||s||^2 + ||x - s||^2 \]
Therefore
\[ ||x||^2 \geq ||s||^2 \]
Moreover, since \( \{x_1, x_2, \ldots, x_m\} \) is an orthonormal set
\[ ||s||^2 = \sum_{i=1}^{m} |a_i|^2 \]
Thus
\[ ||x||^2 \geq \sum_{i=1}^{m} |a_i|^2 \] (8.7)
which is called Bessel’s inequality. Note that (8.7) holds with an equal sign whenever \( x \in \text{span} \{x_1, x_2, \ldots, x_m\} \).

8.8 Least Squares

Suppose that \( \mathcal{X} \) is an infinite dimensional inner product space and that \( x_1, x_2, \ldots \) is an infinite sequence of orthonormal vectors. Let \( x \) be any fixed vector in \( \mathcal{X} \) with finite norm \( ||x|| \). Define \( a_i \equiv (x, x_i), \ i \geq 1 \). Since (8.7) holds for every \( m \geq 1 \), the sequence of numbers
\[ \sum_{i=1}^{m} |a_i|^2, \ \ m \in \{1, 2, \ldots\} \]
is bounded from above by $||x||$. From this it follows that the infinite series

$$\sum_{i=1}^{\infty} |a_i|^2$$

converges, and moreover that

$$\sum_{i=1}^{\infty} |a_i|^2 \leq ||x||^2$$

It turns out that for any fixed $m > 0$, the minimum of the norm square $||x - \sum_{i=1}^{m} b_i x_i||^2$ is attained when $b_i = a_i, i \in \{1, 2, \ldots, m\}$. {Verify this !} Thus the projection of $x$ on $S_m \overset{\Delta}{=} \text{span} \{x_1, x_2, \ldots, x_m\}$, namely

$$s_m \overset{\Delta}{=} a_1 x_1 + a_2 x_2 + \cdots + a_m x_m,$$

can be thought of as that vector in $S_m$ which “best” approximates $x$ in the “least squares” sense. The corresponding least square error is

$$e_m^2 = ||x - \sum_{i=1}^{m} a_i x_i||^2 = (x, x) + \left( \sum_{i=1}^{m} a_i x_i, \sum_{i=1}^{m} a_i x_i \right) - \left( x, \sum_{i=1}^{m} a_i x_i \right) - \left( \sum_{i=1}^{m} a_i x_i, x \right) = ||x||^2 + \sum_{i=1}^{m} |a_i|^2 - 2 \sum_{i=1}^{m} |a_i|^2$$

or

$$e_m^2 = ||x||^2 - \sum_{i=1}^{m} |a_i|^2$$

In the event that

$$\lim_{i \to \infty} e_m^2 = 0,$$

the infinite series

$$\sum_{i=1}^{\infty} a_i x_i$$

is said to converge in the mean to the vector $x$; if this is so,

$$||x||^2 = \sum_{i=1}^{\infty} |a_i|^2$$

If for each $x \in \mathcal{X}$, the corresponding series

$$\sum_{i=1}^{\infty} a_i x_i \quad \{a_i \overset{\Delta}{=} (x, x_i)\}$$
converges, then the sequence of vectors \( x_1, x_2, \ldots \) is called complete and \( \mathcal{X} \) is a complete vector space. Complete inner product spaces such as this are usually called Hilbert spaces \(^4\). Thus in a Hilbert space, the sequence of least squares approximations \( s_1, s_2, \ldots \) of any given vector \( x \), converges to \( x \) at \( m \to \infty \).

### Example 58

Let \( \mathcal{X} \) denote the \( \mathbb{C} \) vector space of all complex-valued, scalar functions \( f \) that are defined and piecewise continuous on the closed interval \( [0, 2\pi] \). Include in the definition of \( \mathcal{X} \), the inner product

\[
(f, g) \triangleq \int_0^{2\pi} f(t)g^*(t)dt
\]

where \( * \) denotes complex conjugate. Consider the infinite sequence of functions

\[
\frac{1}{\sqrt{2\pi}}e^{jkt}, \quad k = 0, \pm 1, \pm 2, \ldots
\]

where \( j \triangleq \sqrt{-1} \). Note that these functions form an orthonormal set since

\[
(e^{ikt}, e^{i\ell t}) = \frac{1}{2\pi} \int_0^{2\pi} e^{j(i-\ell)t}dt = \begin{cases} 0 & \text{for } i \neq \ell \\ 1 & \text{for } i = \ell \end{cases}
\]

In can be shown that this set of functions is also complete. Thus if \( f \) is any given function in \( \mathcal{X} \), then it is possible to express \( f \) as an ‘infinite linear combination’ of the preceding exponentials. In other words,

\[
f(t) = \sum_{k=-\infty}^{k=\infty} a_k e^{jkt}
\]

in the sense that

\[
\lim_{m \to \infty} ||f - \sum_{k=-m}^{k=m} a_k e^{jkt}|| = 0
\]

where

\[
a_k \triangleq (f, e^{jkt}) = \frac{1}{\sqrt{2\pi}} \int_0^{2\pi} f(t)e^{-jkt}dt, \quad k = 0, \pm 1, \pm 2, \ldots
\]

The preceding expression for \( f \) is the **Fourier series** expansion of \( f \) and the \( a_k \) are called the **Fourier coefficients**.

---

\(^4\)Recall that a complete normed spaced is called a Banach space.
CHAPTER 8. INNER PRODUCT SPACES
Chapter 9

Normal Matrices

For the remainder of these notes we shall assume that \( \mathcal{X} \) is either the unitary space \( \mathbb{C}^n \) with inner product \( (x^*)'y \) or the Euclidean space \( \mathbb{R}^n \) with inner product \( x'y \). Since in either case \( \mathcal{X} \) has a basis, we can orthogonalize and normalize such a basis to obtain an orthonormal basis

Suppose that \( \{x_1, x_2, \ldots, x_n\} \) and \( \{y_1, y_2, \ldots, y_n\} \) are two orthonormal bases for \( \mathcal{X} \). Then there must be a nonsingular matrix \( T \) such that

\[
Tx_i = y_i, \quad i \in \{1, 2, \ldots, n\}
\]

These \( n \) equations are equivalent to the single matrix equation

\[
TX = Y
\]

where \( X \triangleq [x_1 \ x_2 \ \cdots \ x_n] \) and \( Y \triangleq [y_1 \ y_2 \ \cdots \ y_n] \). Thus

\[
T = YX^{-1}
\]

In the sequel it will be shown that \( T, X, X^{-1} \) and \( Y \) are all examples of “orthogonal matrices” if \( \mathcal{X} = \mathbb{R}^n \) or “unitary matrices” if \( \mathcal{X} = \mathbb{C}^n \). Orthogonal and unitary matrices as well as “Hermitian” and symmetric matrices are examples of “normal matrices.” The aim of this chapter is to define these matrices different matrix types and to discuss their properties.

9.1 Algebraic Groups

A group \( \mathfrak{G} \triangleq \{G, \circ\} \) is an algebraic system consisting of a set of elements \( G \) together with a single operation \( \circ \) between pairs of elements of \( G \) under which the set is closed, and in addition, certain axioms hold. In particular, for all \( g_1, g_2 \in G \), \( \circ \) must be defined so that \( g_1 \circ g_2 \in G \) and also so that

i. for all \( g_i \in G \), \( g_1 \circ (g_2 \circ g_3) = (g_1 \circ g_2) \circ g_3 \) \{i.e., \( \circ \) is associative\}.

ii. there is an element \( e \in G \), called an identity or unit such that \( e \circ g = g \circ e = g \) for all \( g \in G \).

iii. for each \( g \in G \), there is an inverse in \( G \), written \( g^{-1} \) such that \( g \circ g^{-1} = e \).

If, in addition, the group operation \( \circ \) is commutative \{i.e., \( g_1 \circ g_2 = g_2 \circ g_1 \), \( \forall g_i \in G \)\} then \( \mathfrak{G} \) is called either a commutative or Abelian group. \{What is purple and commutes?\}\(^1\) It is common practice to write \( g \in \mathfrak{G} \)

\(^1\) An Abelian grape.
whenever one means \( g \in G \). There is a quite extensive theory of groups. Algebraic groups arise in many applications including solid state and atomic physics, the study of nonlinear dynamics, and communications. There are more advanced course devoted solely various topics in group theory.

Example 59

1. The set of all \( n \times n \) nonsingular, real-valued matrices together with matrix multiplication is a
   \{noncommutative\} group called the general linear group. For this group, \( I_{n \times n} \) is the identity.

2. The set of all \( n \times n \) matrices with exactly one 1 in each row and in each column and 0’s elsewhere, together with matrix multiplication is the \( n \)-dimensional permutation group. \( I_{n \times n} \) is also the identity for this group.

3. The set of all \( n \times n \) nonsingular, diagonal, real-valued matrices together with matrix multiplication is an Abelian group. Again for this group, \( I_{n \times n} \) is the identity.

4. The set of all \( n \times m \), real-valued matrices together with matrix addition is an Abelian group. For this group, \( 0_{n \times m} \) is the identity. The inverse of a matrix \( M \) in this group \{with respect to the group operation of addition\} is simply \(-M\).

9.1.1 Orthogonal Group

A real-valued, nonsingular matrix is said to be orthogonal if its inverse is equal to its transpose. In other words

\[ A \text{ is orthogonal } \iff A' = A^{-1} \]

Let \( O^n \) denote the set of all \( n \times n \) orthogonal matrices. We claim that \( O^n \) is a multiplicative group.

i. Matrix multiplication is associative. Moreover \( O^n \) is closed under multiplication: in particular, if \( A \) and \( B \) are in \( O^n \), then

\[ (AB)' = B'A' = B^{-1}A^{-1} = (AB)^{-1} \]

ii. \( I_{n \times n} \) is in \( O^n \) and serves as the group’s identity.

iii. If \( A \in O^n \) and \( B = A^{-1} \), then

\[ B' = (A^{-1})' = (A')' = A = B^{-1} \]

so \( B' = B^{-1} \). Thus \( B \) is orthogonal and therefore in \( O^n \). Thus \( A^{-1} \) serves as the group inverse of \( A \).

Thus \( O^n \) is a group.

Let us note that a given real matrix \( A_{n \times n} \) is orthogonal if and only if

\[ A'A = I_{n \times n} \quad \text{(9.4)} \]

Writing \( A \) as \( A = [a_1 \ a_2 \ \cdots \ a_n] \), we therefore see that

\[ A'A = \begin{bmatrix} a'_1 a_1 & a'_1 a_2 & \cdots & a'_1 a_n \\ \vdots & \vdots & \vdots & \vdots \\ a'_n a_1 & a'_n a_2 & \cdots & a'_n a_n \end{bmatrix} \]
is the Gramian of the columns of $A$. I clearly follows that (9.4) holds just in case $\{a_1, a_2, \ldots, a_n\}$ is an orthonormal set. Therefore a real-valued square matrix is orthogonal if and only if its columns form an orthonormal set.

Returning to (9.2) above, we see that if $X = \mathbb{R}^n$, the matrices $X$ and $Y$ are orthogonal. Since $O^n$ is a group, it follows from (9.3) that $T$ is orthogonal as well.

9.1.2 Unitary Group

A complex-valued, nonsingular matrix $A$ is called a unitary matrix if the inverse of $A$ equals the transpose of $A^*$, $A^*$ being the matrix whose elements are the complex conjugates of the corresponding elements of $A$; i.e., $A^* = [a^*_{ij}]$ where $A = \begin{bmatrix} a_{ij} \end{bmatrix}$. $A^*$ is called the complex conjugate of $A$. Thus for $A$ to be unitary means

\[ A^{-1} = (A^*)' \]

It is easy to verify that the set of all $n \times n$ unitary matrices together with ordinary matrix multiplication is a group. This group is called the unitary group.

Now suppose that $\mathcal{X} \triangleq \mathbb{C}^n$ equipped with the inner product $(x, y) \overset{\Delta}{=} (x^*)'y$. Using a Gramian, the reader should verify that a square matrix $A$ is unitary if and only if its columns form an orthonormal set with respect to this inner product. In (9.2) above, $X$ and $Y$ are thus unitary matrices as is $T$ in (9.3) because the set of $n \times n$ unitary matrices from a multiplicative group.

9.2 Adjoint Matrix

Again let $\mathcal{X}$ denote either $\mathbb{R}^n$ or $\mathbb{C}^n$ and define $(x, y) \overset{\Delta}{=} x'y$ if $\mathcal{X} = \mathbb{R}^n$ or $(x, y) \overset{\Delta}{=} (x^*)'y$ if $\mathcal{X} = \mathbb{C}^n$. In either case, the adjoint of $A$, written $A^a$, is that unique matrix such that

\[ (Ax, y) = (x, A^a y) \quad (9.5) \]

It is easy to verify that

\[ A^a = \begin{cases} A' & \text{if } \mathcal{X} = \mathbb{R}^n \\ (A^*)' & \text{if } \mathcal{X} = \mathbb{C}^n \end{cases} \quad (9.6) \]

Observe from (9.6) that a real-valued {resp. complex-valued} matrix is orthogonal {resp. unitary} if and only if it equal to the inverse of its adjoint.

Note from (9.5) that orthogonal and unitary matrices are isometric in the sense that they preserve metric or length. In other words $A^a = A^{-1}$ implies that

\[ \|Ax\| = \sqrt{(Ax, Ax)} = \sqrt{(x, A^a Ax)} = \sqrt{(x, x)} = \|x\| \]

and conversely. Said differently, for all $x$ the length of $Ax$ equals the length of $x$ just in case $A$ is an orthogonal {or unitary} matrix.

9.2.1 Orthogonal Complement

With $\mathcal{X}$ as above, let $S$ be a subspace of $\mathcal{X}$. The orthogonal complement of $S$, written $\mathcal{X}^\perp$, is the set of all vectors in $\mathcal{X}$ which are orthogonal to all vectors in $S$. Thus

\[ S^\perp \overset{\Delta}{=} \{ x : (x, s) = 0, \forall s \in S \} \]
It is easy to verify that $S^\perp$ is a subspace of $\mathcal{X}$. Since all vectors in $S^\perp$ are orthogonal to all vectors in $S$, it must be that $S^\perp$ and $S$ are independent subspaces. What’s more, if $x$ is any given vector in $\mathcal{X}$, we know how to construct vectors $s \in S$ and $y \in S^\perp$ such that $x = s + y$. This implies that $\mathcal{X} \subset S + S^\perp$ and therefore that $\mathcal{X} = S + S^\perp$. But since $S$ and $S^\perp$ are independent subspaces, this sum must be direct:

$$\mathcal{X} = S \oplus S^\perp$$  \hspace{1cm} (9.7)

### 9.2.2 Properties of Adjoint Matrices

Again let $\mathcal{X}$ denote either $\mathbb{R}^n$ or $\mathbb{C}^n$ and define $(x, y) \triangleq x'y$ if $\mathcal{X} = \mathbb{R}^n$ or $(x, y) \triangleq (x^*)'y$ if $\mathcal{X} = \mathbb{C}^n$. Let $A$ be a fixed matrix in $\mathbb{R}^{n \times n}$ if $\mathcal{X} = \mathbb{R}^n$ or in $\mathbb{C}^{n \times n}$ if $\mathcal{X} = \mathbb{C}^n$. Suppose that $S$ is an $A$-invariant subspace and let $S^\perp$ denote the orthogonal complement of $S$. What we want to prove is that $S^\perp$ is $A^\circ$-invariant. In other words, we want to show that

$$AS \subset S \Rightarrow A^\circ S^\perp \subset S^\perp$$  \hspace{1cm} (9.8)

Following the standard game plan for establishing such an implication, let $y$ be any vector in $S^\perp$. Then $y$ is orthogonal to each vector in $S$; i.e.,

$$(y, s) = 0, \ \forall s \in S$$

But since $S$ is $A$-invariant, $As \in S$ for all $s \in S$. Therefore

$$(y, As) = 0, \ \forall s \in S$$

But $(y, As) = (A^\circ y, s)$ so

$$(A^\circ y, s) = 0 \ \forall s \in S$$

Thus $A^\circ y$ is orthogonal to every vector $s \in S$ which means that $A^\circ y \in S^\perp$. We have shown that for each vector $y \in S^\perp$, $A^\circ y \in S^\perp$. Therefore (9.8) is true. Using similar reasoning, it can be easily shown that the reverse implication in (9.8) is also true.

What we want to show next, is that if $A$ and $A^\circ$ have a common eigenvector, then the corresponding eigenvalues must be complex conjugates. In other words, we want to prove that

$$\begin{cases} Ax = \lambda x \\ A^\circ x = \mu x \end{cases} \text{ and } x \neq 0 \Rightarrow \lambda^* = \mu$$  \hspace{1cm} (9.9)

To prove that this is so, let us note that $Ax = \lambda x$ and $A^\circ x = \mu x$ imply

$$(x, Ax) = (x, \lambda x) = \lambda^* ||x||^2$$ \quad \text{and} \quad $$\quad (A^\circ x, x) = (\mu x, x) = \mu ||x||^2$$

respectively. But $(A^\circ x, x) = (x, Ax)$ and $||x||^2 \neq 0$ so $\lambda^* = \mu$ as claimed.

### 9.3 Normal Matrices

So far in our discussion of metric spaces we’ve encountered two special types of matrices, namely orthogonal and unitary. These matrices are examples of a more general class of matrices, called ‘normal matrices’ which have especially important properties worth understanding. The aim of this section is to explain what these properties are and why normal matrices possess them. We begin with the definition of a normal matrix.

A square matrix $A$ defined over either $\mathbb{R}$ or $\mathbb{C}$ is said to be normal if it commutes with its adjoint. Thus $A$ is normal just in case

$$AA^\circ = A^\circ A$$
Since both orthogonal and unitary matrices equal the inverses of their respective adjoints, it follows that both are normal matrices. There are two other types of normal matrices of particular importance in engineering and the sciences:

1. **Symmetric matrices**: A matrix \( A \) with elements in \( \mathbb{R} \) is symmetric if it equals its adjoint; i.e.,

\[
A \in \mathbb{R}^{n \times n} \text{ is symmetric } \iff A = A^a
\]

Note that for such matrices, \( A^a = A' \), so \( A \) is symmetric just in case it equals its transpose.

2. **Hermitian matrices**: A matrix \( A \) with elements in \( \mathbb{C} \) is Hermitian if it equals its adjoint; i.e.,

\[
A \in \mathbb{C}^{n \times n} \text{ is Hermitian } \iff A = A^a
\]

Note that for such matrices, \( A^a = (A^*)' \), so \( A \) is Hermitian just in case it equals its conjugate transpose.

### 9.3.1 Diagonalization of a Normal Matrix

All normal matrices \{e.g., unitary, orthogonal, symmetric, Hermitian\} have a special property. In particular, each can be diagonalized by means of a similarity transformation determined by a unitary matrix. We now explain why this is so.

We begin by proving the following:

**Claim:** Any two \( n \times n \) matrices which commute have a common eigenvector.

To prove this, let \( A_{n \times n} \) and \( B_{n \times n} \) be any two square matrices satisfying

\[
AB = BA \tag{9.10}
\]

Suppose that \( x \) is an eigenvector of \( A \); i.e., \( x \neq 0 \) and

\[
Ax = \lambda_0 x \tag{9.11}
\]

for some \{possibly complex\} number \( \lambda_0 \). It follows from (9.10) and (9.11) that

\[
AB^k x = \lambda_0 B^k x, \quad k = 0, 1, 2, \ldots \tag{9.12}
\]

so for each \( k \geq 0 \), \( B^k x \) is an eigenvector of \( A \) \{provided \( B^k \neq 0 \). Indeed, any nonzero vectors in the \( B \)-cyclic subspace

\[
\text{span } \{ x, Bx, \ldots, B^{n-1} x \}
\]

is an eigenvector of \( A \). Let \( \alpha(s) \) be the minimal polynomial of \( x \) relative to \( B \). Therefore

\[
\alpha(B)x = 0 \tag{9.13}
\]

Let \( \lambda_1 \) be a root of \( \alpha(s) \). Then \( \alpha(s) \) can be written as

\[
\alpha(s) = (s - \lambda_1)\beta(s) \tag{9.14}
\]

for some polynomial \( \beta(s) \). What’s more, \( \beta(A)x \neq 0 \) because \( \alpha(s) \) is the minimal polynomial of \( x \) and \( \deg \beta(s) < \deg \alpha(s) \). Set \( y \overset{\Delta}{=} \beta(A)x \). Thus \( y \) is nonzero; moreover \( y \) is in the span of the set of vectors \( \{ x, Bx, \ldots, B^{n-1} x \} \). Therefore \( y \) is an eigenvector of \( A \). Furthermore, from (9.13) and (9.14) and the definition of \( y \) it follows that

\[
(B - \lambda_1 I) y = (B - \lambda_1 I) \beta(B)x = \alpha(B)x = 0
\]
so \( y \) is an eigenvector of \( B \). Thus the preceding claim is true. \( \blacksquare \)

Now suppose that \( A_{n\times n} \) is a normal matrix. This means that \( A \) must commute with \( A^a \). But we've just proved that any two matrices which commute must have a common eigenvector. Therefore there must be a nonzero vector \( x_1 \) and \{possibly complex\} numbers \( \lambda_1 \) and \( \mu_1 \) such that

\[
Ax_1 = \lambda_1 x_1 \quad \text{and} \quad A^a x_1 = \mu_1 x_1
\]  
\( (9.15) \)

Because of this, the subspace \( S_1 \) spanned by \( x_1 \) is both \( A \)-invariant and \( A^a \)-invariant. But the \( A^a \)-invariance of \( S_1 \) together with \( (9.8) \) imply that \( S_1^\perp \) must be \( A \)-invariant as well. In summary

\[
AS_1 \subset S_1, \quad AS_1^\perp \subset S_1^\perp, \quad \text{and} \quad S_1 \oplus S_1^\perp = \mathcal{X}
\]

The direct sum was noted before in \( (9.7) \).

Let \( u_1 \) be a normalized version of \( x_1 \) and let \( \{u_2, u_3, \ldots, u_n\} \) be an orthonormal basis for \( S_1^\perp \). Since \( S_1 \oplus S_1^\perp = \mathcal{X} \), \( \{u_1, u_2, u_3, \ldots, u_n\} \) must be an orthonormal basis for \( \mathcal{X} \). Therefore

\[
U_1 \triangleq \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix}
\]

must be a unitary matrix. Moreover since \( AS_1 \subset S_1 \), and \( AS_1^\perp \subset S_1^\perp \), it must be true that

\[
AU_1 = U_1 \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}
\]  
\( (9.16) \)

where \( \lambda_1 \) is the eigenvalue of \( A \) associated with eigenvector \( x_1 \) and \( A_1 \) is the unique solution to the linear equation

\[
A \begin{bmatrix} u_2 & u_3 & \cdots & u_n \end{bmatrix} = \begin{bmatrix} u_2 & u_3 & \cdots & u_n \end{bmatrix} A_1
\]  
\( (9.17) \)

Clearly \( (9.16) \) can be written as

\[
U_1^{-1} AU_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}
\]

We claim that \( A_1 \) is a normal matrix. To prove this, let us first note that

\[
\begin{bmatrix} \lambda_1^a & 0 \\ 0 & A_1^a \end{bmatrix} = \left( U_1^{-1} AU_1 \right)^a = U_1 A^a (U_1^{-1})^a = U_1^{-1} A^a U_1
\]

Thus

\[
\begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix} \begin{bmatrix} \lambda_1^a & 0 \\ 0 & A_1^a \end{bmatrix} = \left( U_1^{-1} AU_1 \right) \left( U_1^{-1} A^a U_1 \right) = U_1^{-1} AA^a U_1 = U_1^{-1} A^a U_1 = \left( U_1^{-1} A^a U_1 \right) \left( U_1^{-1} AU_1 \right)
\]

Clearly \( A_1 \lambda_1^a = A_1^a A_1 \) so \( A_1 \) is normal. To summarize, we've just shown that for any normal matrix \( A_{n\times n} \), there must be a \( n \times n \) unitary matrix \( U_1 \), an eigenvalue \( \lambda_1 \) of \( A \) and a \( (n-1) \times (n-1) \) normal matrix \( A_1 \) such that

\[
U_1^{-1} AU_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}
\]

Because \( A_1 \) is normal, it is possible to repeat this process. In particular there must be a \( (n-1) \times (n-1) \) unitary matrix \( V_2 \), an eigenvalue \( \lambda_2 \) of \( A_1 \) and a \( (n-2) \times (n-2) \) normal matrix \( A_2 \) such that

\[
V_2^{-1} A_1 V_2 = \begin{bmatrix} \lambda_2 & 0 \\ 0 & A_2 \end{bmatrix}
\]
It follows that if we define
\[ W_2 \triangleq \begin{bmatrix} 1 & 0 \\ 0 & V_2 \end{bmatrix} \in \mathbb{C}^{n \times n} \]
then \( W_2 \) must be unitary and
\[
W_2^{-1} \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix} W_2 = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & A_2 \end{bmatrix}
\]
Thus with the unitary matrix \( U_2 \triangleq U_1 W_2 \) we can write
\[
U_2^{-1} A U_2 = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & A_2 \end{bmatrix}
\]
Clearly this process can be continued for a total of \( n \) steps. Thus we have proved that for any normal matrix \( A \) there must be a unitary matrix \( U \) which diagonalizes \( U^{-1} A U \).

The converse is also true. In particular, suppose that \( A \) is now any \( n \times n \) matrix and that \( U \) is a unitary matrix which diagonalizes \( U^{-1} A U \). The because \( U^{-1} A U \) is diagonal
\[
(U^{-1} A U)^a = (U^{-1} A U)(U^{-1} A U)^a
\]
But
\[
U^{-1} A^a U = U^{-1} A^a U U^{-1} A U = (U^{-1} A U)^a (U^{-1} A U)
\]
From this and (9.18) it follows that
\[
U^{-1} A^a A = (U^{-1} A U)(U^{-1} A U)^a = U^{-1} A U U^{-1} A^a U = U^{-1} A A^a U
\]
This implies that \( A^a A = AA^a \) so \( A \) is normal. We have proved the following important theorem.

**Theorem 17** A matrix \( A \) with elements in either \( \mathbb{R} \) or \( \mathbb{C} \) is normal if and only if there exists a unitary matrix \( U \) which diagonalizes \( U^{-1} A U \).

### 9.4 Hermitian, Unitary, Orthogonal and Symmetric Matrices

The aim of this section is to discuss the implications of Theorem 17 for Hermitian, unitary, orthogonal and symmetric matrices. We begin with Hermitian matrices.

#### 9.4.1 Hermitian Matrices

Suppose that \( A \) is a Hermitian matrix. Then \( A \) is normal so by Theorem (17) there must be a unitary matrix \( U \) such that \( U^{-1} A U = \Lambda \) where
\[
\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n \end{bmatrix},
\]
the \( \lambda_i \) being the eigenvalues of \( A \). But
\[
\Lambda^* = (\Lambda^*)^t = \Lambda^a = (U^{-1} A U)^a = (U^{-1} A^a U) = U^a A (U^{-1})^a = U^{-1} A U = \Lambda
\]
This can be possible only if \( \Lambda \) is a real matrix. We’ve proved the “only if” part of the following corollary.
**Corollary 2** A matrix $A$ is Hermitian if and only if there exists a unitary matrix $U$ which transforms $A$ into a real diagonal matrix

$$U^{-1}AU$$

The preceding implies, among other things, that whether $A$ is real or not, if it is Hermitian its spectrum \{i.e., set of eigenvalues\} is real. The reader should try to prove the “if” part of Corollary 2.

### 9.4.2 Unitary Matrices

Suppose now that $A$ is a unitary matrix. Then $A$ is normal so by Theorem (17) there must be a unitary matrix $U$ such that $U^{-1}AU = \Lambda$ where

$$\Lambda = \begin{bmatrix}
\lambda_1 & 0 & 0 & \cdots & 0 \\
0 & \lambda_2 & 0 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & \lambda_n
\end{bmatrix},$$

the $\lambda_i$ being the eigenvalues of $A$. But

$$(\Lambda)^\alpha = (U^{-1}AU)^\alpha(U^{-1}AU) = (U^\alpha A^\alpha (U^{-1})^\alpha)(U^{-1}AU) = U^\alpha A^\alpha AU = U^\alpha U = I$$

Thus the the $i$th diagonal element of $(\Lambda)^\alpha$, namely $|\lambda_i|^2$ must equal one. We’ve proved the “only if” part of the following corollary.

**Corollary 3** A matrix $A$ is unitary if and only if there exists a unitary matrix $U$ which transforms $A$ into a diagonal matrix

$$U^{-1}AU$$

whose diagonal elements all have magnitude one.

The preceding implies that whether $A$ is real or not, if it is unitary its eigenvalues all lie on the unit circle in the complex plane. The reader should try to prove the “if” part of Corollary 3.

### 9.4.3 Orthogonal Matrices

Orthogonal matrices are simply real unitary matrices. Thus we have at once

**Corollary 4** A real matrix $A$ is orthogonal if and only if there exists a unitary matrix $U$ which transforms $A$ into a diagonal matrix

$$U^{-1}AU$$

whose diagonal elements all have magnitude one.

### 9.4.4 Symmetric Matrices

Now suppose that $A$ is real and symmetric. Then $A$ is clearly Hermitian so Corollary (2) is applicable and $A$ can be diagonalized buy a unitary matrix $U$. There is of course no reason to believe that $U$ must be real \{and therefore orthogonal\}; in fact it need not be. Nevertheless $A$ can be diagonalized by a real unitary matrix,
but we need to prove that this is so. Let us begin by noting that \( A \) must have real eigenvalues because of the emphasized statement just below Corollary 2. Thus if \( \lambda_1 \) is one of \( A \)'s eigenvalues, then \( \lambda_1 I - A \) must be real. This means that we can find a real eigenvector \( x_1 \) \{any vector in the kernel of \( \lambda_1 I - A \) will do\} such that

\[
Ax_1 = \lambda_1 x_1
\]

Thus if \( S_1 \) is the span of \( x_1 \), then \( AS_1 \subset S_1 \). Moreover \( A^t S_1^\perp \subset S_1^\perp \) because of (9.7). But \( A' = A^t = A \), so \( AS_1^\perp \subset S_1^\perp \). In summary, we have the situation that

\[
AS_1 \subset S_1, \quad AS_1^\perp \subset S_1^\perp, \quad \text{and} \quad S_1 \oplus S_1^\perp = \mathbb{R}^n
\]

except in this case \( \mathcal{S} = \mathbb{R}^n \).

Let \( y_1 \) be a normalized version of \( x_1 \) and let \( \{y_2, y_3, \ldots, y_n\} \) be an orthonormal basis for \( S_1^\perp \). Since \( S_1 \oplus S_1^\perp = \mathbb{R}^n \), \( \{y_1, y_2, y_3, \ldots, y_n\} \) must be an orthonormal basis for \( \mathbb{R}^n \). Therefore

\[
T_1 \triangleq \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}
\]

must be an orthogonal matrix. Moreover since \( AS_1 \subset S_1 \), and \( AS_1^\perp \subset S_1^\perp \), it must be true that

\[
AT_1 = T_1 \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix} \tag{9.19}
\]

where \( A_1 \) is the unique solution to the linear equation

\[
A \begin{bmatrix} y_2 & y_3 \cdots & y_n \end{bmatrix} = \begin{bmatrix} y_2 & y_3 \cdots & y_n \end{bmatrix} A_1 \tag{9.20}
\]

Clearly (9.19) can be written as

\[
T_1^{-1} AT_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}
\]

We leave it to the reader to prove that \( A_1 \) must be symmetric. To summarize, for any real symmetric matrix \( A_{n \times n} \), there must be a \( n \times n \) orthogonal matrix \( T_1 \), a real eigenvalue \( \lambda_1 \) of \( A \) and a \((n - 1) \times (n - 1)\) real symmetric matrix \( A_1 \) such that

\[
T_1^{-1} AT_1 = \begin{bmatrix} \lambda_1 & 0 \\ 0 & A_1 \end{bmatrix}
\]

By continuing this process, much like before when we discussed Hermitian matrices, we can eventually transform \( A \) into a diagonal matrix by means of a sequence of orthogonal similarity transformations. Since the product of orthogonal matrices is orthogonal, we see that any symmetric matrix is orthogonally similar to a real diagonal matrix. The reader should try to prove the converse. We summarize.

**Corollary 5** A real matrix \( A \) is symmetric if and only if there exists an orthogonal matrix \( T \) such that \( T^{-1} AT \) is a real diagonal matrix.

The diagonal entries of \( T^{-1} AT \) are of course the eigenvalues of \( A \).

**Example 60** Let

\[
A \triangleq \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}
\]

Then \( A \)'s characteristic polynomial is

\[
\det [sI - A] = \det \begin{bmatrix} s - 1 & -2 \\ -2 & s - 1 \end{bmatrix} = (s - 1)^2 - 4 = (s - 3)(s + 1)
\]
Therefore the eigenvalues of $A$ are 3 and $-1$. To compute an eigenvector for 3, we need to find a nonzero vector in the kernel of the matrix

$$3I - A = \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}$$

The vector

$$x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

will do. Normalizing we get

$$y_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

By a similar process we obtain for the eigenvalue $-1$ the normalized eigenvector

$$y_2 = \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$$

Note that $y_1$ and $y_2$ are automatically orthogonal as predicted by the theory. Since $y_1$ and $y_2$ are both normalized, the matrix

$$T = \begin{bmatrix} y_1 & y_2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$$

is orthogonal. Finally we see that

$$T^{-1}AT = \begin{bmatrix} 3 & 0 \\ 0 & -1 \end{bmatrix}$$

9.5 Real Quadratic Forms

By a real quadratic form is meant an algebraic expression of the type

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j \quad (9.21)$$

where the $x_i$ are real scalar variables and the $a_{ij}$ are fixed real numbers. A quadratic form can be thought of as a special type of real bilinear form

$$\sum_{i=1}^{n} \sum_{j=1}^{m} b_{ij} x_i y_j \quad (9.22)$$

in which $m = n$, $b_{ij} = a_{ij}$, and $y_i = x_i$. The term bilinear stems from the fact that (9.22) is linear in the $x_i$ {or the $y_j$} if the $y_j$ {or the $x_i$} are held fixed. The reason (9.21) is called a quadratic form is because it is a sum of pairwise products of the $x_i$. Quadratic forms find many applications in engineering and the sciences.

It is especially useful to exploit matrix notation when dealing with quadratic forms. By introducing the vectors

$$x \triangleq \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad y \triangleq \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

and the matrices

$$A \triangleq [a_{ij}]_{n \times n}, \quad B \triangleq [b_{ij}]_{n \times m},$$

(9.22) and (9.21) can be written as

$$x'By \quad \text{and} \quad x'Ax$$

respectively.
9.5. REAL QUADRATIC FORMS

9.5.1 Symmetric Quadratic Forms

A quadratic form $x'Ax$ is said to be symmetric if $A$ is a symmetric matrix. It turns out that any quadratic form $x'Bx$ can be rewritten as a symmetric quadratic form. To understand how to do this, let us first note that

$$x'Bx = (x'Bx)'$$

because $x'Bx$ is a scalar. Since $(x'Bx)' = x'B'x$ it follows that

$$x'Bx = x'B'x$$

Therefore

$$x'Bx = \frac{1}{2}(x'Bx + x'B'x) = x'\left(\frac{1}{2}(B + B')\right)x$$

Hence if we define

$$A \triangleq \frac{1}{2}(B + B')$$

then $A$ is symmetric and

$$x'Bx = x'Ax, \quad \forall x$$

Thus, without loss of generality, we may as well assume that every quadratic form is symmetric.

9.5.2 Change of Variables: Congruent Matrices

Suppose $x'Ax$ is a real, symmetric quadratic form. Let $T$ be a nonsingular matrix and define the new vector of variables

$$y \triangleq T^{-1}x$$

Then $x = Ty$ so

$$x'Ax = y'T'ATy$$

Note that $T'AT$ is also a symmetric matrix. We see that corresponding to the change of variables $x \mapsto y$, there is the coefficient matrix transformation $A \mapsto T'AT$.

More precisely, two real $n \times n$ symmetric matrices $A$ and $B$ are said to be congruent if there exists a real nonsingular matrix $T$ such that $B = T'AT$ - and the mapping $A \mapsto T'AT$ is called a congruence transformation. It is a simple matter to verify that matrix congruence is an equivalence relation on the set of all real-valued $n \times n$ matrices. Thus, for example, if $A$ is congruent to $B$ and $B$ is congruent to $C$ then $A$ is congruent to $C$.

9.5.3 Reduction to Principal Axes

There are various ways to simplify the structure of a quadratic form by changing variables. Suppose for example that we are given the real symmetric quadratic form

$$x'Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij}x_ix_j$$

and we wish to define a new set of variables $y_1, y_2, \ldots, y_n$ by a formula of the form

$$y = T^{-1}x$$

(9.23)

in such a way that
i. $y$ has the same length as $x$ and

ii. the new expression for $x' Ax$ in terms of $y_i$ variables involves no cross terms; i.e.

$$x' Ax = \sum_{i=1}^{n} b_i y_i^2$$  \hspace{1cm} (9.24)

To satisfy requirement i. it is clearly necessary and sufficient that $T$ be an orthogonal matrix. To satisfy requirement ii., $T' AT$ must be a diagonal matrix whose $i$th diagonal entry is $b_i$. In other words, $T$ must be an orthogonal matrix which diagonalizes $T' AT$. Now we’ve already explained how to construct an orthogonal matrix $T$ which diagonalizes $T^{-1} AT$ - and with $T$ so defined, the diagonal elements of $T^{-1} AT$ are necessarily the eigenvalues of $A$. Moreover, since $T$ is orthogonal $T^{-1} = T'$ so $T$ diagonalizes $T' AT$ as required. In other words, with $T$ so defined, (9.24) holds and $||y|| = ||x||$ for all pairs $\{x, y\}$ satisfying (9.23). The $y_i$ are sometimes called the principal axes of the quadratic form.

### 9.5.4 Sum of Squares

Again consider the symmetric quadratic form

$$x' Ax = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$

What we aim to do now is to define a new set of variables $z_1, z_2, \ldots, z_n$ by a formula of the form

$$z = T^{-1} x$$  \hspace{1cm} (9.25)

in such a way that the new expression for $x' Ax$ in terms of $z_i$ variables involves no cross terms and in addition, the coefficients of the $z_i^2$ are numbers in the finite set $\{1, -1, 0\}$. In other words, we will define the $z_i$ so that

$$x' Ax = \sum_{i=1}^{n} c_i z_i^2$$  \hspace{1cm} (9.26)

where each $c_i$ is either 1, $-1$ or 0. To accomplish this $T$ must be nonsingular, but not necessarily orthogonal, and $T' AT$ must be a diagonal matrix with diagonal elements in the set $\{1, -1, 0\}$. The construction of $T$ is as follows.

First pick an orthogonal matrix $M$ which diagonalizes $M^{-1} AM$. Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ denote the diagonal elements of $M^{-1} AM$. The $\lambda_i$ are of course the eigenvalues of $A$. Let $n_1$ denote the number of eigenvalues that are positive and $n_2$ the number which are negative. For simplicity, suppose that $M$ has been defined so that the first $n_1$ $\lambda_i$ are positive, the next $n_2$ $\lambda_i$ negative, and the last $n - n_1 - n_2$ $\lambda_i$ are all zero. Then

$$M' AM = \begin{bmatrix} A_1 & 0 & 0 \\ 0 & A_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

where

$$A_1 = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n_1} \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} \lambda_{n_1+1} & 0 & \cdots & 0 \\ 0 & \lambda_{n_1+2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_{n_1+n_2} \end{bmatrix}$$
It follows that if we define
\[ N_1 \Delta= \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{\lambda_{n_1}} \end{bmatrix} \quad \text{and} \quad N_2 \Delta= \begin{bmatrix} \sqrt{-\lambda_{n_1+1}} & 0 & \cdots & 0 \\ 0 & \sqrt{-\lambda_{n_1+2}} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sqrt{-\lambda_{n_1+n_2}} \end{bmatrix}, \]

then
\[ \Lambda_1 = N'_1 I_1 N_1 \quad \text{and} \quad \Lambda_2 = -N'_2 I_2 N_2 \]

where \( I_1 \) and \( I_2 \) are the \( n_1 \times n_1 \) and \( n_2 \times n_2 \) identity matrices respectively. Thus
\[ M'AM = N' \begin{bmatrix} I_1 & 0 & 0 \\ 0 & -I_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} N \]

where \( N \) is the \( n \times n \) nonsingular matrix
\[ N \Delta= \begin{bmatrix} N_1 & 0 & 0 \\ 0 & N_2 & 0 \\ 0 & 0 & I_3 \end{bmatrix} \]

and \( I_3 \) is the \((n - n_1 - n_2) \times (n - n_1 - n_2)\) identity matrix.

Hence if we now define
\[ T \Delta= MN^{-1} \]

then we can write
\[ T'AT = A_C \] (9.27)

where
\[ A_C \Delta= \begin{bmatrix} I_1 & 0 & 0 \\ 0 & -I_2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \] (9.28)

From this it clearly follows that if \( z \) is defined as in (9.25), then (9.26) will hold with \( c_1 = c_2 = \cdots = c_{n_1} = 1, \)
\( c_{n_1+1} = c_{n_1+2} = \cdots c_{n_1+n_2} = -1, \) and \( c_{n_1+n_2+1} = c_{n_1+n_2+2} = \cdots c_n = 0. \)

Let us note that the rank of \( A \), written \( r \), must equal the rank of \( T'AT \) because \( T \) is nonsingular. But the rank of \( T'AT \) must equal \( n_1 + n_2 \) because of (9.27) and the structure of \( A_C \) in (9.28). In other words
\[ r = n_1 + n_2 \]

Let us define the signature of \( A \), written \( s \), to be the difference between the number positive eigenvalues of \( A \) and the number of negative eigenvalues of \( A \). Thus
\[ s \Delta= n_1 - n_2 \]

It can be shown that \( s \), like \( r \), is invariant under congruent transformations. What’s more, the structure of \( A_C \) in (9.28) is uniquely determined by the congruence invariants \( r \) and \( s \) because
\[ n_1 = \frac{r + s}{2} \quad \text{and} \quad n_2 = \frac{r - s}{2} \]

\( A_C \) is accordingly called the congruence canonical form of \( A \).
A real \{symmetric\} quadratic form \(x'Ax\) is said to be \textit{positive semidefinite} if
\[
x'Ax \geq 0, \quad \forall x
\] (9.29)
If, in addition, the only vector for which \(x'Ax = 0\) is \(x = 0\), i.e.,
\[
x'Ax = 0 \implies x = 0,
\] (9.30)
then the quadratic form is \textit{positive definite}. \(A\) is a \textit{positive definite matrix} \{resp. \textit{positive semidefinite matrix}\} if \(x'Ax\) is a positive definite \{resp. positive semidefinite\} quadratic form. \(A\) is a \textit{negative definite matrix} \{resp. \textit{negative semidefinite matrix}\} if \(-A\) is a positive definite \{resp. positive semidefinite\} matrix. Symmetric matrices which are neither positive semidefinite nor negative semidefinite are called \textit{indefinite matrices}. We sometimes write \(A > 0\) or \(A \geq 0\) to indicate that \(A\) is positive definite or positive semidefinite respectively.

We claim that if \(A\) is positive definite, then so is any matrix \(B\) which is congruent to \(A\). For if \(A\) and \(B\) are any such matrices, then (9.29) and (9.30) must hold and there must be a nonsingular matrix \(T\) such that
\[
B = T'AT.
\]
Now if \(B\) were not positive definite, there would have to be a nonzero vector \(y\) such that
\[
y'By \leq 0
\]
But if such a vector were to exist, then with \(x \overset{\Delta}{=} Ty\) one would have
\[
x'Ax \leq 0 \quad \text{and} \quad x \neq 0
\]
which contradicts (9.29) and (9.30). Hence the claim is true.

Similar reasoning can be used to prove that all matrices congruent to a positive semidefinite matrix are positive semidefinite. In other words, positive definiteness and positive semidefiniteness are properties of symmetric matrices which are invariant under congruence transformations.

\subsection*{9.6.1 Conditions for Positive Definiteness}

Let \(A\) be a positive definite matrix and let \(T\) be an orthogonal matrix which diagonalizes \(T^{-1}AT\). Then
\[
T'AT = \begin{bmatrix}
\lambda_1 & 0 & \cdots & 0 \\
0 & \lambda_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \lambda_n
\end{bmatrix}
\]
where \(\lambda_1, \lambda_2, \ldots, \lambda_n\) are the eigenvalues of \(A\). We claim that \(T'AT\) \{and consequently \(A\)\} will be positive definite if and only if all of the \(\lambda_i\) are positive numbers. For if \(\lambda_k \leq 0\), then with the kth unit vector \(e_k\) we would have \(e_k'T'ATe_k = \lambda_k \leq 0\). Since \(e_k \neq 0\), this would mean that \(T'AT\) could not be positive definite. This proves that if \(A\) is positive definite, then its eigenvalues must be positive numbers. The reverse implication can be established in a similar manner. Thus we can conclude that a \textit{symmetric matrix is positive definite if and only if all of its eigenvalues are positive numbers} \footnote{What must be true of the signature and rank of a symmetric matrix in order for it to be positive definite?}. It can be shown in much the same manner that a \textit{symmetric matrix is positive semidefinite if and only if all of its eigenvalues are nonnegative}. 
Were one to rely on the preceding to test for positive definiteness of $A$, one would have to compute the eigenvalues of $A$. There is another way to test for positive definiteness which avoids this. We caution the reader to remember that the test we are about to describe applies only to symmetric matrices.

Suppose

$$A = \begin{bmatrix}
a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\
a_{12} & a_{22} & a_{23} & \cdots & a_{2n} \\
a_{13} & a_{23} & a_{33} & \cdots & a_{3n} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{1n} & \cdots & \cdots & \cdots & a_{nn}
\end{bmatrix}_{n \times n}$$

For $i \in \{1, 2, \ldots, n\}$, define the $i$th principal subarray of $A$ to be

$$D_i \triangleq \begin{bmatrix}
a_{11} & a_{12} & a_{13} & \cdots & a_{1i} \\
a_{12} & a_{22} & a_{23} & \cdots & a_{2i} \\
a_{13} & a_{23} & a_{33} & \cdots & a_{3i} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_{1i} & \cdots & \cdots & \cdots & a_{ii}
\end{bmatrix}_{i \times i}$$

The numbers $\det D_1, \det D_2, \ldots, \det D_n$ are called the $n$ principal leading minors of $A$. A proof of the following theorem can be found in most good texts on matrix algebra.

**Theorem 18** A symmetric matrix $A_{n \times n}$ is positive definite {resp. positive semidefinite} if and only if its $n$ leading principal minors are positive {resp. nonnegative}.

**Example 61** The matrix

$$\begin{bmatrix}1 & 2 \\ 2 & 1\end{bmatrix}$$

has leading minors $1$ and $-3$ and hence is neither positive definite nor positive semidefinite. Note that the matrix has a negative eigenvalue, namely $-1$.

**Example 62** The matrix

$$A = \begin{bmatrix}1 & 2 & 3 \\ 2 & 4 & 6 \\ 3 & 6 & 9\end{bmatrix}$$

has leading minors

$$\det 1 = 1, \quad \det \begin{bmatrix}1 & 2 \\ 2 & 4\end{bmatrix} = 0, \quad \det A = 0$$

and therefore is positive semidefinite but not positive definite.

### 9.6.2 Matrix Square Roots

Let $A$ be a positive semidefinite matrix. A real {square} matrix $B$ is called a square root of $A$ if

$$B^2 = A$$

Such a matrix $B$ is often denoted by $\sqrt{A}$.
Out aim is to show that each positive semidefinite matrix $A$ has a positive semidefinite square root $\sqrt{A}$. To construct $\sqrt{A}$ we first make use of the fact that there is an orthogonal matrix $T$ which diagonalizes $T^{-1}AT$. Thus

$$A = T\Lambda T'$$

where $\Lambda$ is a diagonal matrix whose diagonal elements, $\lambda_1, \lambda_2, \ldots, \lambda_n$ are $A$'s eigenvalues. Since $A$ is positive semidefinite, all of the $\lambda_i$ are nonnegative. Let $\sqrt{\Lambda}$ be that diagonal matrix whose diagonal elements are $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}$. Clearly $\sqrt{\Lambda}\sqrt{\Lambda} = \Lambda$. Hence if we now define

$$\sqrt{A} \triangleq T\sqrt{\Lambda}T'$$

then

$$(\sqrt{A})^2 = T\sqrt{\Lambda}T'T\sqrt{\Lambda}T' = T\sqrt{\Lambda}T^{-1}T\sqrt{\Lambda}T' = T\sqrt{\Lambda}\sqrt{\Lambda}T' = T\Lambda T' = A$$

which is as required.

**QUESTION:** If $M$ is any real, rectangular, $n \times m$ matrix, then why is $M'M$?

1. symmetric?
2. positive semidefinite?
3. positive definite if and only if rank $M = m$?

### 9.7 Simultaneous Diagonalization

Let $A$ and $B$ be real, $n \times n$ symmetric matrices with $A$ positive definite. In the sequel we will explain how to construct a nonsingular (but not necessarily orthogonal) matrix $T$ such that

$$T'AT = I$$  \hspace{1cm} (9.31)

$$T'BT = \Lambda$$  \hspace{1cm} (9.32)

where $I$ is the $n \times n$ identity matrix and $\Lambda$ is a diagonal matrix. *Caution:* The diagonal elements of $\Lambda$ will not necessarily be the eigenvalues of $B$!

To begin, let us note that $A$’s eigenvalues must all be positive because $A$ is positive definite. This means that $A$’s signature must be $n$ and thus $A$’s congruence canonical form must be the $n \times n$ identity matrix.

Let $P$ be any nonsingular matrix, constructed as in §9.5.4, which takes $A$ into its congruence canonical form; i.e.,

$$P'AP = I_{n \times n}$$  \hspace{1cm} (9.33)

Since $P'BP$ is symmetric, it is possible to construct an orthogonal matrix $Q$ which diagonalizes $Q^{-1}(P'BP)Q$. Set $\Lambda \triangleq Q^{-1}(P'BP)Q$. Since $Q$ is orthogonal.

$$Q'P'BPQ = \Lambda$$

Moreover because of (9.33),

$$Q'PAPQ = Q'Q = I$$

Thus if we define

$$T \triangleq PQ$$

then (9.31) and (9.32) both hold.
9.7.1 Constrained Optimization

Now suppose that $A$ and $B$ are $n \times n$ matrices and that $A$ is positive definite. In the sequel we shall consider the problem of finding a vector $x$, within the set of vectors which satisfy

$$x'Ax = 1,$$  \hspace{1cm} (9.34)

which maximizes the value of the quadratic form

$$q(x) \overset{\Delta}{=} x'Bx$$  \hspace{1cm} (9.35)

Towards this end, let us first note that if $T$ is any $n \times n$ nonsingular matrix, and $y$ is the vector of variables defined by

$$y \overset{\Delta}{=} T^{-1}x,$$  \hspace{1cm} (9.36)

then maximization of (9.35) with respect to $x$ satisfying constraint (9.34), is equivalent to maximization of

$$\bar{q}(y) \overset{\Delta}{=} y'T'BTy$$  \hspace{1cm} (9.37)

with respect to $y$ satisfying the constraint

$$y'T'ATy = 1$$  \hspace{1cm} (9.38)

In other words, if $x_0$ is a vector satisfying constraint (9.34), which at the same time maximizes the quadratic form in (9.35), then $y_0 \overset{\Delta}{=} T^{-1}x_0$ maximizes the quadratic form in (9.37) subject to (9.38) and conversely. Moreover, for any such $x_0$ and $y_0$, $q(x_0) = \bar{q}(y_0)$. We leave it to the reader to verify that these statements are correct.

The preceding suggests that we might significantly simplify the maximization problem by choosing $T$ so that

$$T'AT = I$$ \hspace{1cm} (9.39)

$$T'BT = \Lambda$$ \hspace{1cm} (9.40)

where $\Lambda$ is a diagonal matrix with entries $\lambda_1, \lambda_2, \ldots, \lambda_n$. For with $T$ so chosen, (9.37) and (9.38) become

$$\bar{q}(y) = \sum_{i=1}^{n} \lambda_i y_i^2$$ \hspace{1cm} (9.41)

and

$$\sum_{i=1}^{n} y_i^2 = 1$$ \hspace{1cm} (9.42)

respectively, where $y_i$ is the $i$th component of $y$. Suppose $\lambda_k$ is the largest number in the set $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. It is then reasonably clear that the $k$th unit vector $e_k$ is a value of $y$ which maximizes (9.41) subject to (9.42). Thus with $x_0 \overset{\Delta}{=} Te_k$,

$$x_0'Bx_0 = \max_{x'Ax = 1} x'Bx = \lambda_k$$

9.7.2 A Special Case

Now suppose that in the above problem, $A$ is the $n \times n$ identity matrix. Then $T$ must be an orthogonal matrix because of (9.39). Therefore the $\lambda_i$ must be the eigenvalues of $B$ because of (9.40). It follows that in this case, the maximum value of $x'Bx$ subject to $x'x = 1$ is the largest eigenvalue $\lambda_{\max}$ of $B$. Moreover if $z$ is a normalized eigenvector corresponding to $\lambda_{\max}$, then $z'z = 1$ and $z'Bz = \lambda_{\max} z'z = \lambda_{\max}$. Therefore $z$ is a value of $x$ which solves the maximization problem.
Theorem 19 Let $B$ be an $n \times n$ symmetric matrix and let $\lambda_{\text{max}}$ denote the largest eigenvalue of $B$. Then $\lambda_{\text{max}}$ is the maximum value of the quadratic form $q(x) \triangleq x'Bx$ over the set of all vectors $x$ satisfying $x'x = 1$. Moreover, $q(x)$ attains this maximum value at $x = x_0$, where $x_0$ is any normalized eigenvector of $B$ corresponding to the eigenvalue $\lambda_{\text{max}}$.

Using similar reasoning as above, one can prove just as easily the following theorem.

Theorem 20 Let $B$ be an $n \times n$ symmetric matrix and let $\lambda_{\text{min}}$ denote the smallest eigenvalue of $B$. Then $\lambda_{\text{min}}$ is the minimum value of the quadratic form $q(x) \triangleq x'Bx$ over the set of all vectors $x$ satisfying $x'x = 1$. Moreover, $q(x)$ attains this minimum value at $x = x_0$, where $x_0$ is any normalized eigenvector of $B$ corresponding to the eigenvalue $\lambda_{\text{min}}$.

The following are direct consequences of the preceding theorems.

$$\max_x \frac{x'Bx}{x'x} = \lambda_{\text{max}}$$

$$\min_x \frac{x'Bx}{x'x} = \lambda_{\text{min}}$$

These expressions imply that

$$\lambda_{\text{min}}||x||^2 \leq x'Bx \leq \lambda_{\text{max}}||x||^2, \quad \forall x \in \mathbb{R}^n$$

These inequalities are especially useful in the important special case when $B$ is positive definite and consequently $\lambda_{\text{max}}$ and $\lambda_{\text{min}}$ are positive numbers.

Example 63 Suppose $B \triangleq \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$ in which case $\det(sI - B) = (s - 1)(s - 3)$ so $B$ has eigenvalues of 1 and 3. Therefore

$$\max_x \frac{x'Bx}{x'x} = 3 \quad \text{and} \quad \min_x \frac{x'Bx}{x'x} = 1$$

Since $z \triangleq \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ and $w \triangleq \begin{bmatrix} -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$ are normalized eigenvectors of $B$ for the eigenvalues 3 and 1 respectively,

$$z'Bz = 3, \quad ||z|| = 1, \quad w'Bw = 1, \quad \text{and} \quad ||w|| = 1.$$ 

Note that for any fixed positive number $c$, the set of $x$ satisfying

$$x'Bx = c$$

is an ellipse in $\mathbb{R}^2$. The set of $x$ satisfying

$$||x|| = 1$$
is of course the unit circle in $\mathbb{R}^2$. With 
\[ x \triangleq \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \]
these equations can be written as
\[ 2x_1^2 + 2x_1x_2 + 2x_2^2 = c \]
and
\[ x_1^2 + x_2^2 = 1 \]
respectively. In the plane, the situation is as follows.
Part II

Linear Systems
Chapter 10

Introductory Concepts

The term “linear system” arises in many fields and can mean anything from a linear algebraic equation to a linear map between suitably defined Hilbert spaces. The aim of this chapter is to define what is meant by a linear dynamical system.

10.1 Linear Systems

We shall be primarily concerned with physical systems which can be represented by linear, finite dimensional dynamical models. By a real $n$-dimensional continuous-time linear dynamical system with input $u$, state $x$ and output $y$ is meant a system of equations of the form

\[
\dot{x} = A(t)x + B(t)u
\]
\[
y = C(t)x + D(t)u
\]

where $t$ takes values in the set of non-negative reals, $u, x,$ and $y$ are vector-valued functions of $t$, and overdot $\cdot$ means differentiation with respect to time; i.e.,

\[
\dot{x}(t) \triangleq \frac{d}{dt}x(t)
\]

Similarly, by a real $n$-dimensional discrete-time linear dynamical system with input $u$, state $x$ and output $y$ is meant a system of equations of the form

\[
\overset{\circ}{x} = A(t)x + B(t)u
\]
\[
y = C(t)x + D(t)u
\]

where $t$ takes values in the set of natural numbers $\mathbb{N} \triangleq \{0, 1, 2, 3, \cdots\}$, $u, x,$ and $y$ are vector-valued functions of $t$, and overdiamond $\circ$ means time advance; i.e.,

\[
\overset{\circ}{x}(t) \triangleq x(t + 1)
\]

In either the continuous or discrete time case, $u, x,$ and $y$ take values in $\mathbb{R}^m, \mathbb{R}^n$ and $\mathbb{R}^p$ respectively and $A_{n \times n}(t), B_{n \times m}(t), C_{p \times n}(t),$ and $D_{p \times m}(t)$ are real matrix-valued functions. In the event that $A(t), B(t), C(t),$ and $D(t)$ are constant matrices, not depending on $t$, the corresponding system is said to be time-invariant or stationary. Otherwise the system is time-varying.
In the sequel we will often use the coefficient matrix quadruple \( \{A(t), B(t), C(t), D(t)\} \) to denote the continuous-time system (10.1), (10.2) or the discrete-time system (10.3), (10.4), depending on context. The coefficient matrix triple \( \{A(t), B(t), C(t)\} \) denotes either system (10.1), (10.2) or (10.3), (10.4), again depending on context, for the case when \( D(t) = 0 \).

10.2 Continuous-Time Linear Systems

When dealing a continuous-time linear system \( \{A(t), B(t), C(t), D(t)\} \) we shall invariably assume that its input \( u \) and coefficient matrices \( A(t), B(t), C(t), D(t) \) are at least piecewise continuous functions of \( t \). Thus the variation of constants formula applies and the solution to the state equation (10.1) can be written at

\[
x(t) = \Phi(t,t_0)x_0 + \int_{t_0}^t \Phi(t,\tau)B(\tau)u(\tau)d\tau, \quad t \geq t_0 \geq 0
\]

where \( \Phi(t, \tau) \) is the state transition matrix of \( A(t) \) and \( x_0 \) is the state of (10.1) at \( t = t_0 \). Using this and the readout equation (10.2) we thus obtain the input-output formula

\[
y(t) = C(t)\Phi(t,t_0)x_0 + \int_{t_0}^t C(t)\Phi(t,\tau)B(\tau)u(\tau)d\tau + D(t)u(t), \quad t \geq t_0 \geq 0
\]

which completely characterizes the input-output behavior of the system. Let us note that if \( x_0 = 0 \), the preceding defines a bona fide linear function

\[
u \mapsto \int_{t_0}^t C(t)\Phi(t,\tau)B(\tau)u(\tau)d\tau + D(t)u(t)
\]

from the function space of piecewise continuous inputs \( [t_0, \infty) \to \mathbb{R}^m \) to function space of piecewise continuous outputs \( [t_0, \infty) \to \mathbb{R}^p \).

10.2.1 Time-Invariant, Continuous-Time, Linear Systems

In the time-invariant case, coefficient matrices \( A, B, C, \) and \( D \) are constant. Thus in this case, the state transition matrix of \( A \) is of the for

\[
\Phi(t, \tau) = e^{(t-\tau)A}
\]

where \( e^{tA} \) is the matrix exponential

\[
e^{tA} \Delta \sum_{i=0}^{\infty} \frac{1}{i!}(tA)^i
\]

Hence in this case, (10.5) and (10.6) become

\[
x(t) = e^{(t-t_0)A}x_0 + \int_{t_0}^t e^{(t-\tau)A}Bu(\tau)d\tau, \quad t \geq t_0 \geq 0
\]

and

\[
y(t) = Ce^{(t-t_0)A}x_0 + \int_{t_0}^t Ce^{(t-\tau)A}Bu(\tau)d\tau + Du(t), \quad t \geq t_0 \geq 0
\]

respectively.
For time invariant systems, one is often interested in the situation when \( t_0 = 0 \) and \( x_0 = 0 \). Under these conditions, (10.9) simplifies to the convolution integral
\[
y(t) = \int_0^t H(t - \tau)u(\tau)d\tau, \quad t \geq 0
\]
(10.10)
where
\[
H(t) \triangleq Ce^{tA}B + \delta(t)D
\]
and \( \delta \) is the unit impulse at \( t = 0 \). Note that in the special case when \( m = 1 \) and \( u \) is a unit impulse applied at time \( t = 0 \), \( \{A, B, C, D\} \)'s output response is \( H(t) \). It is thus natural to call \( H(t) \) the impulse response matrix of \( \{A, B, C, D\} \).

For time-invariant linear systems there is much to be gained by working in the frequency domain with Laplace Transforms. Let us write
\[
\mathcal{L}\{\dot{x} \mid x(0) = x_0\} = X(s) \text{ and } \mathcal{L}\{y \mid u(0) = u_0\} = Y(s)
\]
Hence by solving for \( X(t) \) we get
\[
X(s) = (sI - A)^{-1}x_0 + (sI - A)^{-1}BU(s)
\]
Substituting \( X(s) \) into the expression for \( Y(s) \) we thus obtain
\[
Y(s) = C(sI - A)^{-1}x_0 + T(s)U(s)
\]
where
\[
T(s) \triangleq C(sI - A)^{-1}B + D
\]
(10.12)
We call \( T(s) \) the transfer matrix of \( \Sigma \) or of \( \{A, B, C, D\} \). We see that \( T(s) \) characterizes completely the relationship between \( U(s) \) and \( Y(s) \) for the case when \( x_0 = 0 \). It is perhaps not surprising then that \( T(s) \) is the Laplace transform of the impulse response matrix \( H(t) \).

The formula in (10.12) clearly shows that the transfer functions in \( T(s) \) are rational functions of \( s \). It is not difficult to see that they are proper rational functions\(^1\). Note that \( T(s) \) will be strictly proper just in case just in case \( D = 0 \).

\( T(s) \) and \( H(t) \) are Laplace transform pairs and should be regarded as two mathematically equivalent input-output descriptions of \( \Sigma \). There is a third input-output description which we will find useful later. What we are referring to is the sequence of matrices
\[
M_0 \triangleq D, \quad M_i \triangleq \frac{d}{dt} \left( H(t) - D\delta(t) \right)_{\mid t=0}, \quad i \geq 1
\]
which is called \( \Sigma \)'s \{ or \( H(s) \)'s or \( T(s) \)'s\} Markov sequence. It is easy to verify that the \( M_i \) relate to the coefficient matrix of \( \Sigma \) by the formulas
\[
M_0 = D, \quad M_i = CA^{i-1}B, \quad i \geq 1
\]
The \( M_i \) are also the coefficient matrices of a formal power series expansion of \( T(s) \); i.e.,
\[
T(s) \approx M_0 + M_1 s^{-1} + M_2 s^{-2} + \cdots
\]
It is possible to prove that two linear systems have the same transfer matrix if and only if they have the same Markov sequence.

\(^1\)A rational function in one variable is proper if the degree of its numerator is no greater than the degree of its denominator; a proper rational function is strictly proper if the degree of its numerator is strictly less than the degree of its denominator. A matrix of rational functions is proper if each of its entries is proper and strictly proper if each of its entries is strictly proper.
10.3 Discrete-Time Linear Systems

The state equation (10.3) of a discrete-time linear system also admits a solution which is similar in form to (10.5). For the discrete case, the equation is of the form

$$x(t) = \Phi(t,t_0)x(t_0) + \sum_{\tau=t_0}^{t-1} \Phi(t-1,\tau)B(\tau)u(\tau)$$  \hspace{1cm} (10.13)

where $\Phi : \mathbb{N} \times \mathbb{N} \rightarrow \mathbb{R}^{n \times n}$ is the discrete-time state transition matrix

$$\Phi(t,\tau) = \begin{cases} A(t)A(t-1)\cdots A(\tau+1) & \text{if } t > \tau \\ I & \text{if } t = \tau \end{cases}$$  \hspace{1cm} (10.14)

Hence from (10.4)

$$y(t) = C(t)\Phi(t,t_0)x(t_0) + \sum_{\tau=t_0}^{t-1} C(t)\Phi(t-1,\tau)B(\tau)u(\tau) + D(t)u(t)$$  \hspace{1cm} (10.15)

10.3.1 Time-Invariant Discrete Time Systems

In the time invariant case, $A(t)$ is constant so from (10.14) we get

$$\Phi(t,\tau) = A(t-\tau), \quad t \geq \tau$$  \hspace{1cm} (10.16)

Hence in the time-invariant case (10.13) and (10.15) simplify to

$$x(t) = A(t-t_0)x(t_0) + \sum_{\tau=t_0}^{t-1} A(t-\tau-1)Bu(\tau)$$  \hspace{1cm} (10.17)

and

$$y(t) = CA(t-t_0)x(t_0) + \sum_{\tau=t_0}^{t-1} CA(t-\tau-1)Bu(\tau) + Du(t)$$  \hspace{1cm} (10.18)

respectively. Just as in the continuous case, for time invariant discrete time systems, one is often interested in the situation when $t_0 = 0$ and $x_0 = 0$. Under these conditions, (10.18) simplifies to the convolution sum

$$y(t) = \sum_{\tau=0}^{t} H(t-\tau)u(\tau), \quad t \geq 0$$  \hspace{1cm} (10.19)

where

$$H(t) \triangleq \begin{cases} D & \text{if } t = 0 \\ CA(t-1)B & \text{if } t > 0 \end{cases}$$  \hspace{1cm} (10.20)

Note that in the special case when $m = 1$ and $u$ is a unit pulse applied at time $t = 0$, $\{A, B, C, D\}$’s output response is $H(t)$. It is thus natural to call $H(t)$ the pulse response matrix of $\{A, B, C, D\}$.

For time-invariant discrete-time linear systems there is much to be gained by working in the frequency domain with z-transforms. Let us write $U(z)$, $X(z)$, and $Y(z)$ for the z-transforms of $u$, $x$, and $y$ respectively. Recall that the z-transform is a linear map and that the z-transform of $\hat{x}$ is $zX(z) - x(0)$. From this and the fact that $x = Ax + Bu$ and $y = Cx + Du$ it follows that

$$zX(z) - x(0) = AX(z) + BU(z) \quad \text{and} \quad Y(z) = CX(z) + DU(z)$$
Hence by solving for $X(z)$ we get
\[ X(z) = (zI - A)^{-1}x_0 + (zI - A)^{-1}BU(z) \]
Eliminating $X(z)$ from the expression for $Y(z)$ we thus obtain
\[ Y(z) = C(zI - A)^{-1}x_0 + T(z)U(z) \]
where
\[ T(z) \overset{\Delta}{=} C(zI - A)^{-1}B + D \] (10.21)
We call $T(z)$ the \textit{discrete-time transfer matrix} of $\Sigma$ or of $\{A,B,C,D\}$. We see that $T(z)$ characterizes completely the relationship between $U(z)$ and $Y(z)$ for the case when $x_0 = 0$. It is perhaps not surprising then that $T(z)$ is the $z$-transform of the pulse response matrix $H(t)$. Note that the formula for the transfer matrix of a discrete-time linear system has exactly algebraic form as the transfer matrix for a continuous-time linear system.

$T(z)$ and $H(t)$ are $z$-transform pairs and should be regarded as two mathematically equivalent input-output descriptions of $\Sigma$. There is a third input-output description which we will find useful later. What we are referring to is the sequence of matrices
\[ M_i \overset{\Delta}{=} H(i), \quad i \geq 0 \]
which is called $\Sigma$’s \{ or $H(t)$’s or $T(s)$’s\} \textit{discrete-time Markov sequence}. It is easy to verify that the $M_i$ relate to the coefficient matrix of $\Sigma$ by the formulas
\[ M_0 = D, \quad M_i = CA^{i-1}B, \quad i \geq 1 \]
The $M_i$ are also the coefficient matrices of a formal power series expansion of $T(z)$; i.e.,
\[ T(z) \approx M_0 + M_1z^{-1} + M_2z^{-2} + \cdots \]
It is possible to prove that two discrete-time linear systems have the same transfer matrix if and only if they have the same Markov sequence.

\subsection{Sampled Data Systems}

Perhaps the most important example of a discrete time system is that which results when the output of a time-invariant, continuous time system
\[ \dot{x} = Ax + Bu \quad y = Cx + Du \]
is sampled. Let $T$ be a fixed positive number called a \textit{sampling time} and let $U_T$ be the subspace of all piecewise constant input signals of the form
\[ u(t) = u_i, \quad iT \leq t \leq (i+1)T, \quad i = 0, 1, 2, \ldots \]
where $u_i \in \mathbb{R}^m$. If $u$ is any such input and if $x(t)$ is the resulting state-response, then for each $i \geq 0$,
\[ x((i+1)T) = e^{AT}x(iT) + \int_{iT}^{(i+1)T} e^{A((i+1)T-\tau)}Bu_i d\tau \]
From this it follows that if we define $s = \tau - iT$, $x_i \overset{\Delta}{=} x(iT)$ and $y_i$ as the limit of $y(t)$ at $t$ approaches $iT$ from above, then for $i \geq 0$
\[ x_{i+1} = (e^{AT})x_i + \left( \int_0^T e^{A(T-s)}Bds \right)u_i \]
\[ y_i = Cx_i + Du_i \]
which is a time-invariant, discrete time, linear system.
10.4 The Concept of a Realization

The continuous and discrete state systems discussed thus far should be regarded as candidate systems for modeling the relationships between physically significant variables, namely inputs and outputs.

\[ u \xrightarrow{\Sigma} y \]

Figure 10.1: Linear System \( \Sigma \)

As we have seen, each such system uniquely determines a linear mapping of inputs to outputs. For example, in the case of a continuous-time linear system \( \{A(t), B(t), C(t), D(t)\} \) initialized at the zero state at \( t = t_0 \), one has the relation

\[ y(t) = \int_{t_0}^{t} W(t, \tau)u(\tau)d\tau + D(t)u(t) \]

where

\[ W(t, \tau) \triangleq C(t)\Phi(t, \tau)B(\tau) \]

\( \Phi(t, \tau) \) being the state transition matrix of \( A(t) \). We call \( W(t, \tau) \) the weighting pattern of \( \{A(t), B(t), C(t)\} \).

Similarly, in the case of a discrete-time linear system \( \{A(t), B(t), C(t), D(t)\} \) initialized at the zero state at \( t = t_0 \), one has the relation

\[ y(t) = \sum_{\tau=t_0}^{t-1} W(t, \tau)u(\tau) + D(t)u(t) \]

where

\[ W(t, \tau) \triangleq C(t)\Phi(t-1, \tau)B(\tau) \]

\( \Phi(t, \tau) \) being the discrete-time state transition matrix of \( A(t) \). In this case, \( W(t, \tau) \) the discrete-time weighting pattern of \( \{A(t), B(t), C(t)\} \).

Consider the continuous time case. Suppose that we are given an arbitrary of \( p \times n \) matrix \( Q(t, \tau) \) which depend at least piecewise continuously on its arguments. By a realization of \( Q(t, \tau) \) is meant any continuous-time linear system \( \{A(t), B(t), C(t)\} \) whose weighting pattern is equal to \( Q(t, \tau) \).

Exactly the same ideas apply in the discrete-time case: Suppose that we are given an arbitrary of \( p \times n \) matrix \( Q : \mathbb{N} \times \mathbb{N} \to \mathbb{R}^{p \times m} \) By a discrete-time realization of \( Q(t, \tau) \) is meant any discrete-time linear system \( \{A(t), B(t), C(t)\} \) whose weighting pattern is equal to \( Q(t, \tau) \).

Consideration of either the continuous-time or discrete-time case leads naturally to a number of questions:

1. **Existence**: Under what conditions does \( Q(t, \tau) \) admit a finite-dimensional realization?

2. **Uniqueness**: To what extent are realizations of \( Q(t, \tau) \) unique?

3. **Time-Invariance**: What must be true of \( Q(t, \tau) \) in order for it to admit a time-invariant linear realization?

4. **Minimal-Realization**: What must be true of a realization of \( Q(t, \tau) \) in order for its \{state-space\} dimension to be minimal over the set of all realizations of \( Q(t, \tau) \)?

Linear realization theory is concerned with development of answers to all of these questions. Answers for the continuous-time case and for the discrete-time case are of course going to be different. Issues
in continuous-time case are typically somewhat more challenging technically than the discrete-time case, because for continuous-time systems one has to deal with differential equations. On the other hand, the continuous-time theory is somewhat more concise than the discrete-time theory; this partially counterbalances the relative technical simplicity the discrete-time case has over the continuous-time case. In any event, it almost always turns out that once one understands an issue in continuous-time, extension to the discrete-time case is a simple matter. For this reason, in the sequel we will usually limit developments primarily to the continuous-time case. The reader is encouraged to work out the discrete-time analog of each continuous-time result which follows.

Transfer Matrix Realizations

In the time-invariant case, the weighting pattern of a linear system \( \{ A, B, C \} \) is of the form

\[
W(t, \tau) = Ce^{(t-\tau)A}B
\]

Note that

\[
W(t, \tau) = H(t - \tau)
\]

where \( H(t) \) is the impulse response of \( \{ A, B, C \} \). Thus realizing a given \( W \) amounts to finding matrices \( A, B, C \) for which

\[
Ce^{tA}B = H(t)
\]

Prompted by this, we also use the word realization to describe any constant linear system \( \Sigma \triangleq \{ A, B, C, D \} \) for which either

\[
Ce^{tA}B + D\delta(t) = H(t)
\]

or

\[
C(sI - A)^{-1}B + D = T(s)
\]

where \( H(t) \) and \( T(s) \) are given matrices and \( \delta(t) \) is the unit impulse at \( t = 0 \). In the former case we say that \( \Sigma \) is an impulse matrix realization, whereas in the latter \( \Sigma \) is a transfer matrix realization. Of course, any transfer matrix realization is also an impulse response realization and vice versa.

Recall that the formula for the transfer matrix of a discrete-time system is identical in form to that for a continuous-time system. Because of this, the transfer matrix realization questions for discrete and continuous time systems are identical. We will address these questions later in the notes.

10.4.1 Existence of Continuous-Time Linear Realizations

The following theorem settles the main existence question for continuous-time linear systems.

**Theorem 21** A \( p \times m \) piecewise-continuous matrix \( Q(t, \tau) \) is the weighting pattern of a linear system, if and only if there exists a finite positive integer \( n \) and piecewise-continuous matrices \( M(t) \) and \( N(t) \) such that

\[
Q(t, \tau) = M(t)N(\tau)
\]

Moreover, if such \( M_{p \times n}(t) \) and \( N_{n \times m}(t) \) exist, \( Q(t, \tau) \) is realized by the system \( \{ A(t), B(t), C(t) \} \), where \( A(t) \triangleq 0_{n \times n}, B(t) \triangleq N(t) \) and \( C(t) \triangleq M(t) \).

**Proof:** The state-transition matrix of \( A(t) \triangleq 0 \) is \( \Phi(t, \tau) = I, \ \forall t, \tau \geq 0 \). Hence the weighting pattern of \( \{ A(t), B(t), C(t) \} \) is

\[
W(t, \tau) = M(t)I_{n \times n}N(\tau) = Q(t, \tau).
\]
Although Theorem 21 answers the existence question, it leaves many closely related issues unresolved. For example, the realization used in the proof of the theorem will typically be a time-varying system, even in the case when \( Q(t, \tau) \) admits a time-invariant realization. Moreover, the dimension of the proof’s realization may well be much larger than that of some other, more carefully constructed realization of \( Q(t, \tau) \). To deal with these issues, we will need to delve into the underlying properties of continuous-time linear systems. Among these is are the properties of controllability and observability, which after stability, are probably the most important concept in linear system theory. Controllability is discussed in the next chapter.
Chapter 11

Controllability

The aim of this chapter is to define and characterize the concept of a controllable continuous-time linear system. We begin with the closely related idea of reachability.

11.1 Reachable States

Let \( z(t, \tau, x; v) \) denote the state of \( \Sigma \) at time \( t \) assuming \( \Sigma \) was in the state \( x \) at time \( \tau \) and \( u(t) = v(t) \) for \( t \in [\tau, t] \). Thus by the variation of constants formula,

\[
z(t, \tau, x; v) = \Phi(t, \tau)x + \int_{\tau}^{t} \Phi(t, s)B(s)v(s)ds
\]

(11.1)

where \( \Phi(t, \tau) \) is the state transition matrix of \( A(t) \).

Let \( t_0 \) and \( T \) be fixed times with \( T > t_0 \). Let \( x_0 \in \mathbb{R}^n \) be given. Let us agree to say that a state \( x \in \mathbb{R}^n \) of \( \Sigma \) is reachable on \( [t_0, T] \) from \( x_0 \) if there exists an input \( v \) defined on \( [t_0, T] \) such that

\[
x = z(T, t_0, x_0; v)
\]

(11.2)

Thus \( x \) is reachable from \( x_0 \) on \( [t_0, T] \) just in case there is an input \( v \) which transfers \( \Sigma \) from state \( x_0 \) at time \( t_0 \) to state \( x \) at time \( T \).

Let \( \mathcal{R}[t_0, T] \) denote the set of all states of \( \Sigma \) which are reachable from the zero state on \( [t_0, T] \). Note that \( \mathcal{R}[t_0, T] \) is nonempty because the zero state of \( \Sigma \) is reachable from the zero state on \( [t_0, T] \) with the input \( v(t) = 0, \ t \in [t_0, T] \). Now suppose that \( x_1 \) and \( x_2 \) are any two states in \( \mathcal{R}[t_0, T] \). This means that there must be inputs \( v_1 \) and \( v_2 \) such that

\[
x_i = z(T, t_0, 0; v_i), \ i \in \{1, 2\}
\]

Then by linearity

\[
r_1x_1 + r_2x_2 = r_1z(T, t_0, 0; v_1) + r_2z(T, t_0, 0; v_2) = z(T, t_0, 0; r_1v_1 + r_2v_2), \ \forall r_1, r_2 \in \mathbb{R}
\]

Thus for any numbers \( r_1 \) and \( r_2 \), the linear combination \( r_1x_1 + r_2x_2 \) is reachable from the zero state on \( [t_0, T] \). Evidently \( \mathcal{R}[t_0, T] \) is not only nonempty, but also closed under vector addition and scalar multiplication. This proves that \( \mathcal{R}[t_0, T] \) is a linear subspace of \( \mathbb{R}^n \). We call \( \mathcal{R}[t_0, T] \) the reachable space of \( \Sigma \) {or of the pair \( (A(t), B(t)) \)} on \( [t_0, T] \). It is possible to characterize \( \mathcal{R}[t_0, T] \) as follows.
Theorem 22

\[ \mathcal{R}[t_0, T] = \text{image } G_C(t_0, T) \]  
(11.3)

where

\[ G_C(t_0, T) = \int_{t_0}^{T} \Phi(T, \tau)B(\tau)B'(\tau)\Phi'(T, \tau)d\tau \]  
(11.4)

The matrix \( G_C(t_0, T) \) defined by (11.4) is called the controllability Gramian of \( (A(t), B(t)) \) or of \( \Sigma \) on \([t_0, T]\). Note that \( G_C(t_0, T) \) is a symmetric, positive semi-definite matrix; in addition if \( G_C(t_0, T) \) is positive definite, then so is \( G_C(t_0, T) \) for any \( T \geq 0 \). From this and Theorem 22 it follows that if every state of \( \Sigma \) is reachable from the zero state on \([t_0, T]\), then every state of \( \Sigma \) is also reachable from the zero state on \([t_0, T] \) provided \( T \geq T \).

Proof of Theorem 22: First suppose that \( x \in \mathcal{R}[t_0, T] \). Then \( x \) is reachable from the zero state on \([t_0, T]\) so there must be an input \( v \) such that

\[ x = \int_{t_0}^{T} \Phi(T, \tau)B(\tau)v(\tau)d\tau \]  
(11.5)

Write \( G_C \) for \( G_C(t_0, T) \). In general, \( \text{(image } G_C)^\perp = \text{kernel } G_C' \), but since \( G_C \) is symmetric, this identity becomes \( \text{(image } G_C)^\perp = \text{kernel } G_C \). Hence \( \mathbb{R}^n \) admits an orthogonal decomposition of the form

\[ \mathbb{R}^n = \text{image } G_C \oplus \text{kernel } G_C \]

Therefore there must be vectors \( g \in \text{image } G_C \) and \( \bar{g} \in \text{kernel } G_C \) such that

\[ x = g + \bar{g} \]  
(11.6)

Since \( \bar{g} \) and \( g \) are orthogonal

\[ \bar{g}'x = ||\bar{g}||^2 \]

where, for all \( y \in \mathbb{R}^n \), \( ||y|| \) denotes the Euclidean norm \( ||y|| = \sqrt{y'y} \). From this and (11.5) it follows that

\[ ||\bar{g}||^2 = \int_{t_0}^{T} \bar{g}'\Phi(T, \tau)B(\tau)v(\tau)d\tau \]  
(11.7)

Using the definition of \( G_C \) in (11.4),

\[ \bar{g}'G_C\bar{g} = \int_{t_0}^{T} \bar{g}'\Phi(T, \tau)B(\tau)B'(\tau)\Phi'(T, \tau)\bar{g}d\tau = \int_{t_0}^{T} ||B'(\tau)\Phi'(T, \tau)\bar{g}||^2d\tau \]  
(11.8)

But

\[ \bar{g}'G_C\bar{g} = 0 \]

because \( \bar{g} \) is in the kernel of \( G_C \). From this and (11.8) it follows that

\[ \int_{t_0}^{T} ||B'(\tau)\Phi'(T, \tau)\bar{g}||^2d\tau = 0 \]

This implies that

\[ B'(\tau)\Phi'(T, \tau)\bar{g} = 0 \quad \forall \tau \in [t_0, T] \]

because \( B'(\tau)\Phi'(T, \tau)\bar{g} \) is continuous on \([t_0, T] \). This shows that the integrand in (11.7) vanishes identically on \([t_0, T] \) so \( \bar{g} = 0 \). In view of (11.6), it must be true that \( x = g \) and therefore that \( x \in \text{image } G_C \). Since \( x \) was chosen arbitrarily, we have proved that

\[ \mathcal{R}[t_0, T] \subset \text{image } G_C(t_0, T) \]  
(11.9)
To prove the reverse inclusion, now fix \( x \in \text{image } G_C \). Then there must be a vector \( q \in \mathbb{R}^n \) such that

\[
x = G_C q
\]

Consider the input

\[
v(t) = B'(t)\Phi(T, t)q, \quad t \in [t_0, T]
\]

With \( v \) so defined

\[
z(T, t_0, 0; v) = \int_{t_0}^{T} \Phi(T, \tau)B(\tau)v(\tau)d\tau = \int_{t_0}^{T} \Phi(T, \tau)B'(\tau)\Phi'(T, \tau)q d\tau = G_C(t_0, T)q = x
\]

Thus \( x \) is reachable from the zero state on \( [t_0, T] \). Since \( x \) is arbitrary, this must be true for every \( x \) in the image of \( G_C \). Therefore

\[
\text{image } G_C(t_0, T) \subset \mathcal{R}[t_0, T]
\]

From this and (11.9), it now follows that (11.3) is true. \( \blacksquare \)

### 11.2 Controllability

A linear system \( \Sigma \triangleq \{ A(t), B(t), C(t), D(t) \} \) \{or matrix pair \( (A(t), B(t)) \) \} is said to be controllable on \([t_0, T]\) if for each pair of states \( x_1, x_2 \in \mathbb{R}^n \) there is a control input \( v \) which transfers \( \Sigma \) from state \( x_1 \) at \( t = t_0 \) to state \( x_2 \) at time \( t = t_2 \). In other words, \( \Sigma \) is controllable on \([t_0, T]\) just in case for each pair \( x_1, x_2 \in \mathbb{R}^n \), there is a \( v \) such that

\[
x_2 = z(T, t_0, x_1, v)
\]

where, as before,

\[
z(t, \tau, x; v) \triangleq \Phi(t, \tau)x + \int_{\tau}^{T} \Phi(t, s)B(s)v(s)ds
\]

Thus a controllable linear system on a given time interval, is a linear system within which it is possible to transfer from any state to any other state on the given time interval.

Let us note that if \( \Sigma \) is controllable on \([t_0, T]\), then every possible state of \( \Sigma \) must be reachable from the zero state on \([t_0, T]\). In other words, if \( \Sigma \) is controllable on \([t_0, T]\), then it must be true that \( \mathcal{R}[t_0, T] = \mathbb{R}^n \). The converse is also true:

**Theorem 23** A linear system \( \Sigma \) with coefficient matrix quadruple \( \{ A(t), B(t), C(t), D(t) \} \) is controllable on \([t_0, T]\) if and only if

\[
\mathcal{R}[t_0, T] = \mathbb{R}^n
\]

In view of Theorem 22, we see that controllability of \( \Sigma \) on \([t_0, T]\) is equivalent to image \( G_C(t_0, T) = \mathbb{R}^n \). Since the latter, in turn, is equivalent to \( \det G_C(t_0, T) \neq 0 \), we are led to the following alternative characterization of controllability.

**Corollary 6** A linear system with coefficient matrix quadruple \( \{ A(t), B(t), C(t), D(t) \} \) is controllable on \([t_0, T]\) if and only if its controllability Gramian \( G_C(t_0, T) \) is nonsingular.

**Proof of Theorem 23**: As we’ve already explained, just before the statement of Theorem 23, controllability of \( \Sigma \) on \([t_0, T]\) implies that \( \mathcal{R}[t_0, T] = \mathbb{R}^n \). To prove the converse, suppose that (11.12) holds, and that \( x_1 \)
Hence (11.10) holds which proves that $v$ has the desired property. 

11.2.1 Controllability Reduction

Let $\Sigma \triangleq \{A(t), B(t), C(t)\}$ be a given $n$-dimensional continuous time linear system and let

$$W(t, \tau) = C(t)\Phi(t, \tau)B(\tau)$$

be its weighting pattern where $\Phi(t, \tau)$ is the state transition matrix of $A(t)$. Let $[t_0, T]$ be any interval of positive length on which $\Sigma$ is defined. The aim of this section is to prove that every weighting pattern has a controllable realization. We will do this constructively by explaining how to use $\Sigma$ to compute a linear system $\bar{\Sigma} \triangleq \{\bar{A}(t), \bar{B}(t), \bar{C}(t)\}$ which is controllable on $[t_0, T]$ and which has the same weighting pattern as $\Sigma$. Since $\bar{\Sigma}$’s dimension will turn out to be less than $\Sigma$’s, if $\Sigma$ is not controllable, this construction is referred to as controllability reduction. The fact that $W(t, \tau)$ can always be realized by a controllable system is one of the cornerstones of the theory upon which the existence of minimal dimensional realizations of $W(t, \tau)$ is based.

To begin, let us suppose right away that $\Sigma$ is not controllable. For if $\Sigma$ were controllable, we could simply define $\bar{\Sigma}$ to be $\Sigma$ and we would have at once a controllable realization of $W(t, \tau)$. To proceed, let $G_C$ denote $\Sigma$’s controllability Gramian on $[t_0, T]$. In other words

$$G_C \triangleq \int_{t_0}^{T} \Phi(T, \tau)B(\tau)B'(\tau)\Phi'(T, \tau) d\tau$$

As noted earlier, the symmetry of $G_C$ implies that $\mathbb{R}^n$ admits an orthogonal decomposition of the form

$$\mathbb{R}^n = \text{image } G_C \oplus \text{kernel } G_C$$

Let $P$ denote the orthogonal projection matrix on image $G_C$. Hence $PG_C = G_C$. Because $G_C$ is positive-semidefinite it is possible\(^1\) to construct a full rank matrix $R_{n \times \bar{n}}$, with $\bar{n} \triangleq \text{rank } G_C$, such that

$$G_C = RR'$$

\(^1\)This can easily be deduce by first noting that $G_C$ is congruent to a block diagonal matrix of the form $\begin{bmatrix} I_{\bar{n} \times \bar{n}} & 0 \\ 0 & 0 \end{bmatrix}$. In other words, for some nonsingular matrix $Q$, $G_C = Q \begin{bmatrix} I_{\bar{n} \times \bar{n}} & 0 \\ 0 & 0 \end{bmatrix} Q'$. Defining $R$ to be the submatrix consisting of the first $\bar{n}$ columns of $Q$ produces the desired factorization.
Our construction relies on the fact that \( P \) can be written as
\[
P = R(R')^{-1}R'
\]

{The reader should verify that this basic algebraic property is so.} Now consider the new system \( \bar{\Sigma} \) defined for \( t \in [t_0, T] \) by
\[
\bar{A}(t) = 0 \quad \bar{B}(t) = S\Phi(T, t)B(t) \quad \bar{C}(t) = C(t)\Phi(t, T)R
\]
where
\[
S \overset{\Delta}{=} (R'R)^{-1}R'
\]
It will now be shown that \( \bar{\Sigma} \) is controllable and that its weighting pattern is \( W(t, \tau) \). We establish controllability first. Toward this end, note that \( \bar{A} \)'s state transition matrix is \( \bar{\Phi}(t, \tau) = I_{n \times n} \). Using this fact we can write \( \bar{\Sigma} \)'s controllability Gramian
\[
\bar{G}_C = \int_{t_0}^{T} \bar{\Phi}(T, t)\bar{B}(t)\bar{B}'(t)\bar{\Phi}'(T, t)dt
\]
as
\[
\bar{G}_C = \int_{t_0}^{T} \bar{B}(t)\bar{B}'(t)dt
\]
Hence, using the definitions of \( \bar{B}(t) \) and \( \bar{C}(t) \) we can write
\[
\bar{G}_C = \int_{t_0}^{T} S\Phi(T, t)B(t)B'(t)\Phi(T, t)'S'dt
\]
\[
= S \left( \int_{t_0}^{T} \Phi(T, t)B(t)B'(t)\Phi(T, t)'dt \right) S'
\]
\[
= SG_CS'
\]
\[
= (R'R)^{-1}R'R'R(R'R)^{-1}
\]
\[
= I_{n \times n}
\]
This prove that \( \bar{G}_C(t_0, T) \) is nonsingular. Therefore by Corollary 6, \( \bar{\Sigma} \) is controllable on \([t_0, T]\) as claimed.

To prove that \( \bar{\Sigma} \)'s weighting pattern is the same as \( \Sigma \)'s, we first need to show that
\[
P\Phi(T, t)B(t) = \Phi(T, t)B(t), \quad t \in [t_0, T]
\]
(11.13)
Now this will be true provided
\[
\int_{t_0}^{T} \|(B'(t)\Phi(T, t)'P' - B'(t)\Phi'(T, t))z\|^2dt = 0, \quad \forall z \in \mathbb{R}^n
\]
(11.14)
Fixing \( z \) arbitrarily and expanding the integral on the left above we get
\[
\int_{t_0}^{T} \|(B'(t)\Phi(T, t)'P' - B'(t)\Phi'(T, t))z\|^2dt = \int_{t_0}^{T} z'P\Phi(T, t)B(t)B'(t)\Phi'(T, t)P'zdt
\]
\[
- \int_{t_0}^{T} z'\Phi(T, t)B(t)B'(t)\Phi'(T, t)P'zdt
\]
\[
- \int_{t_0}^{T} z'P\Phi(T, t)B(t)B'(t)\Phi'(T, t)zdt
\]
\[
+ \int_{t_0}^{T} z'P\Phi(T, t)B(t)B'(t)\Phi'(T, t)zdt
\]
\[
= z'PG_Cz - z'G_CP'z - z'PG_Cz + z'G_Cz
\]
\[
= z'G_Cz - z'G_Cz + z'G_Cz
\]
\[
= 0
\]
Hence (11.14) is true. Consequently (11.13) is true as well.

It remains to be shown that \( \bar{\Sigma} \)'s weighting pattern is the same as \( \Sigma \)'s. This is now easily accomplished by noting that \( \bar{\Sigma} \)'s weighting pattern \( \bar{W} \) satisfies
\[
\bar{W}(t, \tau) = \bar{C}(t) \bar{\Phi}(t, \tau) \bar{B}(\tau)
\]
\[
= C(t) \Phi(t, T) R S \Phi(T, \tau) B(\tau)
\]
\[
= C(t) \Phi(t, T) \Phi(T, \tau) B(\tau)
\]
\[
= C(t) \Phi(t, \tau) B(\tau)
\]
\[
= W(t, \tau)
\]

Therefore \( \bar{\Sigma} \) and \( \Sigma \) have the same weighting pattern as claimed. ■

We summarize.

**Theorem 24** Every continuous-time weighting pattern admits a controllable realization.

### 11.3 Time-Invariant Continuous-Time Linear Systems

While the constructions and results regarding controllability discussed so far, apply to both time-varying and time-invariant linear systems, the consequential simplifications of the time-invariant case are significant enough to justify a separate treatment.

Let \( \Sigma = \{A, B, C, D\} \) be a time-invariant system. Then \( \Sigma \)'s controllability Gramian on \([t_0, T]\) is
\[
G_C(t_0, T) = \int_{t_0}^{T} e^{(T-\tau)A} BB' e^{(T-\tau)A'} d\tau
\]

By introducing the change of variables \( s = \tau - t_0 \) we see that
\[
G_C(t_0, T) = \int_{0}^{T-t_0} e^{(T-t_0-s)A} BB' e^{(T-t_0-s)A'} ds
\]

Since the right hand side of the preceding is also the controllability Gramian of \( \Sigma \) on \([0, T-t_0]\), we can conclude that in the time-invariant case the controllability Gramian of \( \Sigma \) on \([t_0, T]\) depends only on the difference \( T-t_0 \), and not independently on \( T \) and \( t_0 \). For this reason, in the time-invariant case \( t_0 \) is usually taken to be zero. We will adopt this practice in the sequel.

As we’ve already seen, the image of \( \Sigma \)'s controllability Gramian on a given interval \([t_0, T]\) of positive length is the subspace \( R[t_0, T] \) of states of \( \Sigma \) reachable from the zero state on \([t_0, T] \). Evidently this subspace too, only depends on the time difference \( T-t_0 \), and not separately on \( t_0 \) and \( T \). In fact, \( R[t_0, T] \) doesn’t even depend on \( T-t_0 \), provided this difference is a positive number. Our aim is to explain why this is so and at the same time to provide an explicit algebraic characterization of \( R[t_0, T] \) expressed directly in terms of \( A \) and \( B \). Towards this end, let us define the subspace
\[
< A|B > = B + AB + \cdots A^{(n-1)} B
\]

where \( n \) is the dimension of \( \Sigma \) and for \( i \geq 0 \) \( A^i B \) is the image or column span of \( A^i B \). We call \( < A|B > \) the controllable space of \( \Sigma \) or of the pair \((A, B)\). Let us also agree to call the \( n \times nm \) matrix
\[
[ B \ AB \ \cdots \ A^{(n-1)} B ]
\]
the controllability matrix of $\Sigma$ or of the pair $(A, B)$. We see at once that the controllable space of $(A, B)$ is the image or column span of the controllability matrix of $(A, B)$. The following theorem establishes the significance these definitions.

**Theorem 25** For all $T > t_0 \geq 0$

$$\mathcal{R}[t_0, T] = < A|B >$$

(11.15)

The theorem implies that the set of states of a time-invariant which are reachable from the zero state on any interval of positive length, does not depend on either the length of the interval or on when the interval begins. The proof of Theorem 25 depends on the following lemma.

**Lemma 2** For each fixed time $t_1 > 0$

$$< A|B > = \text{image} \int_0^{t_1} e^{At}BB'e^{A't} dt$$

(11.16)

**Proof of Theorem 25:** By definition,

$$G_C(t_0, T) = \int_{t_0}^T e^{(T-\tau)}BB'e^{A'(T-\tau)}d\tau$$

By making the change of variable $t = T - \tau$ we see that

$$G_C(t_0, T) = \int_0^{(T-t_0)} e^{At}BB'e^{A't}dt$$

From this and Theorem 22 we see that

$$\mathcal{R}[t_0, T] = \text{image} \int_0^{(T-t_0)} e^{At}BB'e^{A't}dt$$

Application of Lemma 2 with $t_1 \triangleq T - t_0$ then yields (11.15) which is the desired result.

**Proof of Lemma 2:** Fix $t_1 > 0$ and define

$$M = \int_0^{t_1} e^{At}BB'e^{A't}dt$$

(11.17)

As pointed out earlier in the math notes, $e^{tA}$ can be expressed as

$$e^{tA} = \sum_{i=1}^n \gamma_i(t)A^{(i-1)}$$

for suitable functions $\gamma_i(t)$. It is therefore possible to write $M$ as

$$M = \int_0^{t_1} \sum_{i=1}^n \gamma_i(t)A^{(i-1)}BB'e^{A't}dt$$

Therefore

$$M = \sum_{i=1}^n A^{(i-1)}BN_i$$
where
\[ N_i = \int_0^{t_1} \gamma_i(t) B' e^{A't} dt, \quad i \in \{1, 2, \ldots, n\} \]

It follows that
\[ \text{image} \ M = \text{image} \sum_{i=1}^{n} A^{(i-1)} B N_i \subset \sum_{i=1}^{n} \text{image} \ A^{(i-1)} B N_i \subset \sum_{i=1}^{n} \text{image} \ A^{(i-1)} B = \langle A|B \rangle \]

For the reverse inclusion, let \( x \) be any vector in \( \text{kernel} \ M \). Then \( Mx = 0, \ x'Mx = 0, \)
\[ x' \left( \int_0^{t_1} e^{A't} BB' e^{A't} dt \right) x = 0 \]
and so
\[ \int_0^{t_1} ||B' e^{A't} x||^2 dt = 0 \]
It follows that
\[ B' e^{A't} x = 0, \ \forall t \in [0, t_1] \]
Repeated differentiation yields
\[ B'(A')^{(i-1)} e^{A't} x = 0, \ \forall t \in [0, t_1], \ i \in \{1, 2, \ldots, n\} \]
Evaluation at \( t = 0 \) thus provides
\[ B'(A')^{(i-1)} x = 0, \ i \in \{1, 2, \ldots, n\} \]
Clearly
\[ R' x = 0 \]
where \( R \) is the controllability matrix
\[ R = [B \ AB \ \cdots \ A^{(n-1)}] \]
Therefore \( x \in \text{kernel} \ R' \) and since \( x \) was chosen arbitrarily in \( \text{kernel} \ M \), it must be true that \( \text{kernel} \ M \subset \text{kernel} \ R' \). Therefore, taking orthogonal complements we have
\[ \text{image} \ R \subset \text{image} \ M' \]
But \( M \) is symmetric and \( \text{image} \ R = \langle A|B \rangle \) so
\[ \langle A|B \rangle \subset \text{image} \ M \]
which is the required reverse containment.

In the light of Theorems 23 and 25 we can now state

**Theorem 26** An \( n \)-dimensional time-invariant linear system \( \{A, B, C, D\} \) is controllable if and only if
\[ \langle A|B \rangle = \mathbb{R}^n \]

From this and the fact that the controllable space of \( (A, B) \) is the image of the controllability matrix of \( (A, B) \) we can also state the following.
We’d like to illustrate next that controllability is a “generic” or typical property of a time-invariant linear system. To make this a little more precise, suppose we write \( M \) for the linear vector space of all real matrix pairs \( (A_{n \times n}, B_{n \times m}) \), with addition and scalar multiplication of such pairs defined in the most natural way. We’d like to show that “almost every” pair in \( M \) is controllable. To do this we first provide a characterization of those pairs which are not controllable. To do this, let us first note that each \( n \)th order minor of the controllability matrix of \( (A, B) \) is a polynomial function of the \( n^2 + nm \) entries of \( A \) and \( B \). Let \( \mu \) denote the sum of the squares of all such minors. Then \( \mu \) is also a polynomial function of the \( n^2 + nm \) entries of \( A \) and \( B \). Moreover \( \mu \) will vanish for some pair \( (A, B) \), just in case all of the \( n \)th order minors of \( (A, B) \)’s controllability matrix vanish. Now for \( (A, B) \) not to be controllable, all \( n \)th order minors must vanish. This follows from Corollary 7 and from the definition of rank. Thus the set of pairs \( (A, B) \) which are not controllable is the same as the set of \( (A, B) \) pairs for which \( \mu \) vanishes. This set is thus a strictly proper algebraic subset of \( M \). Its complement – let’s call it \( S \) – is the set of all controllable pairs in \( M \). Because \( \mu \) is a continuous function, \( S \) must be an open subset in \( M \). Moreover, from the fact that \( \mu \) is a polynomial it can also be shown that \( S \) is dense\(^2\) in \( M \). In this sense controllable pairs in \( M \) are “almost all” pairs in \( M \).

11.3.1 Properties of the Controllable Space of \( (A, B) \)

The controllable space of \( (A, B) \) turns out to have several important properties which we will make use of in the next section. The first, and perhaps most important of these is that \(< A\mid B >\) is an \( A \)-invariant subspace.

Lemma 3 For each matrix pair \((A_{n \times n}, B_{n \times m})\),
\[
A < A\mid B > \subset < A\mid B >
\] (11.18)

Proof: Fix \( x \in < A\mid B > \). Thus there must be vectors \( b_i \in B, i \in \{1, 2, \ldots, n\} \) such that
\[
x = b_1 + Ab_2 + \cdots + A^{(n-2)}b_{n-1} + A^{(n-1)}b_n
\]
Then
\[
Ax = Ab_1 + A^2b_2 + \cdots + A^{(n-1)}b_{n-1} + A^n b_n
\]
Now by the Cayley-Hamilton theorem
\[
A^n = -a_1 I - a_2 A - \cdots - a_n A^{(n-1)}
\]
where \( s^n + a_n s^{(n-1)} + \cdots + a_2 s + a_1 \) is the characteristic polynomial of \( A \). Therefore
\[
Ax = -a_1 b_n + A(b_1 - a_2 b_n) + A^2 (b_2 - a_3 b_n) + \cdots + A^{(n-1)} (b_{n-1} - a_n b_n)
\]
Thus \( Ax \in < A\mid B > \). Since this is true for all \( x \in < A\mid B > \), (11.18) must be true. \( \blacksquare \)

By the controllability index of \((A_{n \times n}, B_{n \times m})\) is meant the smallest positive integer \( n_c \) for which
\[
< A\mid B > = B + AB + \cdots + A^{(n_c-1)} B
\]
Clearly \( n_c \leq n \), with the inequality typically being strict if \( m > 1 \). The following is a consequence of the preceding lemma and the Cayley-Hamilton theorem.

\(^2\)This means that each matrix pair in \( M \) can be approximated with arbitrarily small error by a controllable pair in \( M \).
Lemma 4 For each matrix pair $(A_{n \times n}, B_{n \times m})$,

$$< A|B > = B + AB + \ldots A^{(k-1)}B, \quad \forall k \geq n_c$$  \hspace{1cm} (11.19)

where $n_c$ is the controllability index of $(A, B)$.

Proof: In view of the definition of $n_c$ it is enough to prove that

$$< A|B > = B + AB + \ldots A^{(k-1)}B,$$  \hspace{1cm} (11.20)

holds for all $k > n$. To prove that this is so we will use induction. We know, first of all that (11.20) is true if $k = n$. Suppose therefore that $i \geq n$ is a fixed integer such that (11.20) holds for all $n \leq k \leq i$. Our goal is use this to prove that that (11.20) also holds for $k = i + 1$. For this, we first note that

$$B + AB + \ldots A^i B = B + A(B + AB + \ldots A^{(i-1)}B)$$

But by the inductive hypothesis,

$$B + AB + \ldots A^{(i-1)}B = < A|B >$$

so

$$B + AB + \ldots A^i B = B + A < A|B >$$

From this and the $A$-invariance of $< A|B >$, there follows

$$B + AB + \ldots A^i B \subset B + < A|B >$$

But since $B \subset < A|B >$, it must be true that $B + < A|B > = < A|B >$. Therefore

$$B + AB + \ldots A^i B \subset < A|B >$$

Since the reverse containment is clearly true, it follows that (11.20) must hold for $k = i + 1$. Hence, by induction (11.19) is true. \hfill \Box

There is an interesting way to characterize a controllable space, which differs from what we’ve discussed so far. Let $(A, B)$ be given and fixed. Suppose we define $\mathcal{C}$ to be the class of all subspaces of $\mathbb{R}^n$ which are $A$-invariant and which contain $B$. That is

$$\mathcal{C} \triangleq \{ S : AS \subset S, \quad B \subset S \}$$

Note that $\mathbb{R}^n \in \mathcal{C}$ so $\mathcal{C}$ is nonempty. Moreover it is easy to show that if $S \in \mathcal{C}$ and $T \in \mathcal{C}$, then $S \cap T \in \mathcal{C}$. In other words, $\mathcal{C}$ is closed under subspace intersection. Because of this and the finite dimensionality of $\mathbb{R}^n$, it is possible to prove that $\mathcal{C}$ contains a unique smallest element which is a subspace of every other subspace in $\mathcal{C}$. In fact, $< A|B >$ turns out to be this unique smallest subspace. In other words, $< A|B >$ contains $B$, is $A$-invariant, and is contained in every other subspace which contains $B$ and is $A$-invariant. The reader may wish to prove that this is so.

11.3.2 Control Reduction for Time-Invariant Systems

Although the control reduction technique described earlier also works in the time-invariant case, it is possible to accomplish the same thing in a more algebraic manner without having to use the controllability gramian. What we shall do is to explain how to construct from a given time-invariant linear system, a new time-invariant linear system which is controllable and which has the same transfer matrix as the system we started with. To carry out this program we shall make use of the following fact, which is of interest in its own right.
Lemma 5 Let \( \{A_{n \times n}, B_{n \times m}, C_{p \times n}\} \) and \( \{\bar{A}_{\bar{n} \times \bar{n}}, \bar{B}_{\bar{n} \times m}, \bar{C}_{p \times \bar{n}}\} \) be two constant matrix triples. If there exists a matrix \( T_{n \times \bar{n}} \) such that
\[
AT = T\bar{A} \quad B = T\bar{B} \quad CT = \bar{C}
\]
then
\[
C(sI - A)^{-1}B = \bar{C}(sI - \bar{A})^{-1}\bar{B}
\]

Proof: Since \( AT = T\bar{A} \), it must be true that \( (sI - A)T = T(sI - \bar{A}) \) and thus that
\[
T(sI - \bar{A})^{-1} = (sI - A)^{-1}T
\]
Therefore
\[
C(sI - A)^{-1}B = C(sI - A)^{-1}TB = CT(sI - \bar{A})^{-1}\bar{B} = \bar{C}(sI - \bar{A})^{-1}\bar{B}
\]
which completes the proof. \( \blacksquare \)

The lemma implies that any two time-invariant linear systems whose coefficient matrices are related as in (11.21), have the same transfer matrix. It is important to note that such systems need not have the same dimension, and that \( T \) need not be nonsingular or even square. Suppose that that (11.21) holds and that \( \Sigma \)'s is a system of the form
\[
\dot{x} = \bar{A}x + \bar{B}u \quad y = \bar{C}x
\]
If we then define
\[
x \triangleq T\bar{x}
\]
then it is possible to write
\[
\dot{x} = Ax + Bu \quad y = Cx
\]
which is a new linear system \( \Sigma \) modeling the same input-output relationship between \( u \) and \( y \). In the event that \( \Sigma \) and \( \bar{\Sigma} \) do have the same dimension \( \{\text{i.e., } n = \bar{n}\} \) and \( T \) is nonsingular, (11.21) yields
\[
T^{-1}AT = \bar{A} \quad TB = \bar{B} \quad CT^{-1} = \bar{C}
\]
Under these conditions it is natrual to say that \( \Sigma \) and \( \bar{\Sigma} \) are similar linear systems.

We now turn to the problem of constructing from a given time-invariant linear system \( \Sigma \triangleq \{A, B, C\} \), a new time-invariant linear system \( \bar{\Sigma} \) which is controllable and which has the same transfer matrix as \( \Sigma \). As a first step, define \( \bar{\Sigma} \)'s dimension \( \bar{n} \triangleq \dim < A|B > \). Next pick a basis matrix \( R_{n \times \bar{n}} \) for \( < A|B > \). Because \( < A|B > \) is \( A \)-invariant, the equation
\[
AR = R\bar{A}
\]
has a unique solution \( \bar{A}_{\bar{n} \times \bar{n}} \). Moreover, because \( B \subset < A|B > \), the equation
\[
B = R\bar{B}
\]
also has a unique solution \( \bar{B} \). Define
\[
\bar{C} \triangleq CR
\]
It follows from (11.22)-(11.24) and Lemma 5 that \( \bar{\Sigma} \triangleq \{\bar{A}, \bar{B}, \bar{C}\} \) has the same transfer matrix as \( \Sigma \). Now from (11.22), \( A^iR = R\bar{A}^i \), \( i \geq 1 \). This and (11.23) imply that
\[
A^iB = R\bar{A}^i\bar{B}, \quad i \geq 0
\]
Therefore
\[
[B \ AB \ldots \ A^{(n-1)}]_{n \times m} = R_{n \times \bar{n}} \begin{bmatrix} \bar{B} & \bar{A}\bar{B} & \ldots & \bar{A}^{(n-1)}\bar{B} \end{bmatrix}_{\bar{n} \times m}
\]
This implies that
\[
\text{rank } [B \ AB \ \cdots \ A^{(n-1)}]_{n \times m} \leq \text{rank } [\bar{B} \ \bar{A} \bar{B} \ \cdots \ \bar{A}^{(n-1)}\bar{B}]_{\bar{n} \times m}
\]
But
\[
\text{rank } [B \ AB \ \cdots \ A^{(n-1)}]_{n \times m} = \bar{n}
\]
and
\[
\text{rank } [\bar{B} \ \bar{A} \bar{B} \ \cdots \ \bar{A}^{(n-1)}\bar{B}]_{\bar{n} \times m} = \bar{n}
\]
This is equivalent to
\[
\bar{B} + \bar{A} \bar{B} + \cdots + \bar{A}^{(n-1)} \bar{B} = \bar{A} \bar{B}
\]
Therefore
\[
< \bar{A} | \bar{B} > = \mathbb{R}^{\bar{n}}
\]
so \((\bar{A}, \bar{B})\) is a controllable pair. Hence \(\bar{\Sigma}\) is a controllable system with the same transfer matrix as \(\Sigma\). □

### 11.3.3 Controllable Decomposition

The preceding discussion was concerned solely with the problem of constructing from a given linear system, a new controllable linear system with the same transfer matrix as the original. Our aim now is to represent the original system
\[
\dot{x} = Ax + Bu \quad y = Cx
\]
in a new coordinate system within which \(\{\bar{A}, \bar{B}, \bar{C}\}\) appears naturally as a subsystem. Toward this end, let \(S\) be any \(n \times (n - \bar{n})\) matrix for which \(T^{-1}AS\) is nonsingular. Define
\[
\begin{bmatrix}
\bar{A} \n \times (n - \bar{n}) \\
\bar{A}(n - \bar{n}) \times (n - \bar{n})
\end{bmatrix} = T^{-1}AS \quad \hat{C} = CS
\]
Recall that
\[
AR = R\bar{A}, \quad B = R\bar{B} \quad CR = \bar{C}
\]
From these expressions it follows that
\[
T^{-1}AT = \begin{bmatrix}
\bar{A} & \hat{A} \\
0 & \hat{A}
\end{bmatrix} \quad T^{-1}B = \begin{bmatrix}
\bar{B} \\
0
\end{bmatrix} \quad CT = \begin{bmatrix}
\bar{C} & \hat{C}
\end{bmatrix}
\]
Hence if we define \(z = T^{-1}x\), and partition \(z\) as
\[
z \triangleq \begin{bmatrix}
\bar{z} \\
\hat{z}
\end{bmatrix}
\]
where \(\bar{z}\) is a \(\bar{n}\)-vector, then we can write
\[
\dot{\bar{z}} = \bar{A}\bar{z} + \hat{A}\bar{z} + \bar{B}u \quad (11.26)
\]
\[
\dot{\hat{z}} = \hat{A}\hat{z} \quad (11.27)
\]
\[
y = \bar{C}\bar{z} + \hat{C}\hat{z} \quad (11.28)
\]
Note that \(u\) has no influence on \(\hat{z}\) and moreover that if \(\hat{z}(0) = 0\), then the input-output relationship is determined solely by the controllable subsystem
\[
\dot{\bar{z}} = \bar{A}\bar{z} + \bar{B}u \quad y = \bar{C}\bar{z}
\]
The spectrum of \(\bar{A}\) is is uniquely determined by \((A, B)\) and is sometimes called the *uncontrollable spectrum* of \((A, B)\).
Single-Input Systems

Let \((A_n, b_n)\) be a controllable matrix pair. This means that the controllability matrix \(Q \triangleq [b \ A \ b \ldots \ A^{(n-1)}b]\) has rank \(n\) and thus that \(\{b, Ab, \ldots, A^{(n-1)}b\}\) is a basis for \(\mathbb{R}^n\). In other words,

**Lemma 6** An \(n\)-dimensional, single input pair \((A, b)\) is controllable if and only if \(A\) is cyclic and \(b\) is a generator for \(\mathbb{R}^n\).

By means of similarity, we can transform such a pair \((A, b)\) into an especially useful form. Towards this end define the row vector

\[
h \triangleq [0 \ 0 \ldots \ 0 \ 1]_{1 \times n} Q^{-1}
\]

Then

\[
hQ = [0 \ 0 \ldots \ 0 \ 1]_{1 \times n}
\]

so

\[
hb = 0, \ hAb = 0, \ldots, hA^{n-2}b = 0 \quad (11.29)
\]

and

\[
hA^{n-1}b = 1 \quad (11.30)
\]

Define the matrix

\[
T \triangleq \begin{bmatrix}
  h \\
  hA \\
  \vdots \\
  hA^{n-1}
\end{bmatrix}_{n \times n} \quad (11.31)
\]

and note that

\[
Tb = \begin{bmatrix}
  0 \\
  0 \\
  \vdots \\
  1
\end{bmatrix}_{n \times 1} \quad (11.32)
\]

We claim that \(T\) is nonsingular. Suppose it were not; then it would be possible to find numbers \(g_i\) such that

\[
hA^{n-1} = \sum_{i=1}^{n-1} g_i hA^{i-1}
\]

This would imply that

\[
hA^{n-1}b = \sum_{i=1}^{n-1} g_i hA^{i-1}b = 0
\]

because of (11.29). But this contradicts (11.30), so \(T\) cannot be singular. □

In view of the definition of \(T\) in (11.31)

\[
TA = \begin{bmatrix}
  hA \\
  hA^2 \\
  \vdots \\
  hA^n
\end{bmatrix} \quad (11.33)
\]

and

\[
hA^i = e_{i+1}'T, \quad i \in \{1, 2, \ldots, n-1\} \quad (11.34)
\]
where \( e_{i+1} \) is the \( i \)th unit vector in \( \mathbb{R}^n \). Meanwhile, by the Cayley-Hamilton Theorem

\[
hA^n = -a_1 h - a_2 hA - \cdots - a_n hA^{n-1}
\]

where

\[
\alpha(s) = s^n + a_n s^{n-1} + \cdots + a_2 s + a_1
\]
is the characteristic polynoma of \( A \). Hence

\[
hA^n = \begin{bmatrix} -a_1 & -a_2 & \cdots & -a_n \end{bmatrix} T
\]  

(11.35)
because of the definition of \( T \) in (11.31). From (11.33) - (11.35) we thus obtain the expression

\[
TA = \begin{bmatrix} e'_1T \\ e'_2T \\ \vdots \\ e'_nT \\ [-a_1 & -a_2 & \cdots & -a_n]T \end{bmatrix}
\]

Factoring out \( T \) on the right we get

\[
TA = A_C T
\]  

(11.36)

where

\[
A_C = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -a_1 & -a_2 & -a_3 & \cdots & -a_n \end{bmatrix}_{n \times n}
\]

From this and (11.32) we conclude that

\[
A_C \triangleq TAT^{-1} \quad b_C = Tb
\]  

(11.37)

where

\[
b_C \triangleq \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{n \times 1}
\]

\((A_C, b_C)\) is uniquely determined by \((A, b)\) and is called the control canonical form of the single-input controllable pair \((A, b)\). Note that \((A_C, b_C)\) is also uniquely determined by \( A \)'s characteristic polynomial \( \alpha(s) \). Control canonical forms find applications in several different places. One is in realizing transfer functions Another is in defining a state-feedback to assign to a “closed-loop” linear system a prescribed spectrum. We will discuss theses applications in later chapters.
Chapter 12

Observability

In this chapter we explore the extent to which the state of a continuous time linear system can be recovered from available measured signals, namely $u$ and $y$. In doing the we will come up with a concept, namely observability, which plays a role dual to controllability in the realization and representation of a linear system. We begin with a discussion of unobservable states.

12.1 Unobservable States

Let $\theta(t, t_0, x_0; v)$ denote the value of the output of the $n$-dimensional linear system

$$\Sigma: \quad y = C(t)x + D(t)u \quad \dot{x} = A(t)x + B(t)u$$

(12.1)

at time $t$, assuming the system was in state $x_0$ at time $t_0$ and that $v(t)$ was the input applied on the time interval $[t_0, t]$. Then, by the variation of constants formula,

$$\theta(t, t_0, x_0; v) = C(t)\Phi(t, t_0) x_0 + \int_{t_0}^{t} W(t, \tau) v(\tau) d\tau + D(t)v(t)$$

(12.2)

where $\Phi(t, \tau)$ is the state transition matrix of $A(t)$ and $W(t, \tau)$ is the weighting pattern

$$W(t, \tau) = C(t)\Phi(t, \tau)B(\tau)$$

Fix $T > t_0 \geq 0$. Two states $x_1$ and $x_2$ of $\Sigma$ are said to be indistinguishable on $[t_0, T]$, if for all inputs $v(t)$ which are defined and piecewise-continuous in $[t_0, T]$,

$$\theta(t, t_0, x_1; v) = \theta(t, t_0, x_2; v), \quad \forall t \in [t_0, T]$$

(12.3)

Thus two states are indistinguishable just in case both determine the same mapping of $\Sigma$’s inputs into its outputs on $[t_0, T]$. This means that if $x_1$ were $x_2$ are indistinguishable, it would impossible to decide from input-output measurements on $[t_0, T]$ whether $\Sigma$ was in state $x_1$ or $x_2$ at time $t_0$. It is easy to verify that indistinguishability is an equivalence relation on $\Sigma$’s state space $\mathbb{R}^n$. The most one could expect to learn from input-output measurements on $[t_0, T]$ about the state $x_0$ of $\Sigma$ at time $t_0$, is thus the indistinguishability equivalence class within which $x_0$ resides.

In the light of (12.2), it is clear that $x_1$ and $x_2$ are indistinguishable if and only if

$$C(t)\Phi(t, t_0)(x_1 - x_2) = 0 \quad \forall t \in [t_0, T]$$

(12.4)
This suggests that the set of states indistinguishable from the zero state ought to be easy to characterize. Accordingly, write
\[ \mathcal{N}[t_0, T] \triangleq \{ x : C(t)\Phi(t, t_0)x = 0, \ \forall t \in [t_0, T] \} \]
and note that \( \mathcal{N}[t_0, T] \) is a subspace of \( \mathbb{R}^n \). \( \mathcal{N}[t_0, T] \) is thus the set of states of \( \Sigma \) which produce the zero output on \( [t_0, T] \) when \( \Sigma \) is initialized at any such state at time \( t_0 \) and which the zero input is applied on \( [t_0, T] \). We call \( \mathcal{N}[t_0, T] \) the unobservable space of \( \Sigma \) (or of \((C(t), A(t))\)) on \([t_0, T]\). Note that two states \( x_1 \) and \( x_2 \) of \( \Sigma \) are indistinguishable on \([t_0, T]\) just in case their difference \( x_1 - x_2 \) is an unobservable state in \( \mathcal{N}[t_0, T] \).

Mathematically \( \mathcal{N}[t_0, T] \) possesses a set of properties “dual” to those of the reachable space of \( \Sigma \) discussed earlier.

**Theorem 27**

\[ \mathcal{N}[t_0, T] = \text{kernel } G_O(t_0, T) \]  
(12.5)

where
\[ G_O(t_0, T) \triangleq \int_{t_0}^{T} \Phi'(\tau, t_0)C'(\tau)C(\tau)\Phi((\tau, t_0))d\tau \]  
(12.6)

The matrix \( G_O(t_0, T) \) defined by (12.6) is called the observability Gramian of \((C(t), A(t))\) or of \( \Sigma \) on \([t_0, T]\). Note that \( G_O(t_0, T) \) is a symmetric, positive semi-definite matrix; in addition it is and a non-decreasing function of \( T \) is the sense that
\[ G_O(t_0, T_2) \geq G_O(t_0, T_1) \text{ if } T_2 \geq T_1 \]

From this and Theorem 27 it follows that \( \Sigma \)'s unobservable space is a non-increasing function of \( T \) in the sense that
\[ \mathcal{N}[t_0, T_2] \subset \mathcal{N}[t_0, T_1] \text{ if } T_2 \geq T_1 \]

This means that any state unobservable on \([t_0, T_2] \) is also unobservable on \([t_0, T_1] \) provided \( T_2 \geq T_1 \).

**Proof of Theorem 27:** First suppose that \( x \in \mathcal{N}[t_0, T] \). Then
\[ C(t)\Phi(t, t_0)x = 0, \ \forall t \in [t_0, T] \]

Thus
\[ \Phi'(t, t_0)C'(t)C(t)\Phi(t, t_0)x = 0, \ \forall t \in [t_0, T] \]

Therefore
\[ \int_{t_0}^{T} \Phi'(\tau, t_0)C'(\tau)C(\tau)\Phi(\tau, t_0)x d\tau = 0 \]

From this and the definition of \( G_O(t_0, T) \) in (12.6) it follows that \( G_O(t_0, T)x = 0 \). Therefore \( x \in \text{kernel } G_O(t_0, T) \).

Since \( x \) is chosen arbitrarily, we have proved that
\[ \mathcal{N}[t_0, T] \subset \text{kernel } G_O(t_0, T) \]  
(12.7)

To prove the reverse inclusion, now fix \( x \in \text{kernel } G_O(t_0, T) \). Then
\[ \left( \int_{t_0}^{T} \Phi'(\tau, t_0)C'(\tau)C(\tau)\Phi(\tau, t_0)d\tau \right)x = 0 \]

Therefore
\[ \int_{t_0}^{T} |C(\tau)\Phi(\tau, t_0)x|^2 d\tau = 0 \]

so
\[ C(t)\Phi(t, t_0)x = 0, \ \forall t \in [t_0, T] \]

It follows that \( x \in \mathcal{N}[t_0, T] \). Thus kernel \( G_O(t_0, T) \subset \mathcal{N}[t_0, T] \). Therefore (12.5) is true. ■
12.2 Observability

A linear system $\Sigma \overset{Δ}{=} \{A(t), B(t), C(t), D(t)\}$ (or matrix pair $(C(t), A(t))$) is said to be observable on $[t_0, T]$ if for each initial state $x(t_0) = x_0$ and each input $v$, defined on $[t_0, T]$, it is possible to uniquely determine $x_0$ from from knowledge of $v$ on $[t_0, T]$ and $\Sigma$’s response $y$ on $[t_0, T]$. In other words, $\Sigma$ is observable on $[t_0, T]$ just in case for each pair $x_1, x_2 \in \mathbb{R}^n$ and each input $v$,

$$x_1 = x_2 \text{ whenever } \theta(t, t_0, x_1; v) = \theta(t, t_0, x_2; v), \quad \forall t \in [t_0, T]$$

where, as before,

$$\theta(t, t_0, x; v) = C(t)\Phi(t, t_0)x_0 + \int_{t_0}^{t} W(t, \tau)v(\tau)d\tau + D(t)v(t)$$

(12.8)

Thus an observable linear system on a given time interval, is a linear system for which it is possible to uniquely determine the system’s state at the beginning of the interval from measurements of the system’s input and output on the interval.

Let us note that if $\Sigma$ is observable on $[t_0, T]$, then $\mathcal{N}[t_0, T]$ must be the zero subspace – for if $\mathcal{N}[t_0, T]$ contained two distinct vectors, and if $\Sigma$ were initialized at one of them at $t = t_0$, then it would be impossible to uniquely recover this initial state from knowledge of $u$ and $y$, because distinct states in $\mathcal{N}[t_0, T]$ are indistinguishable on $[t_0, T]$. In other words, if $\Sigma$ is observable on $[t_0, T]$, then it must be true that $\mathcal{N}[t_0, T] = 0$. The converse is also true:

**Theorem 28** A linear system $\Sigma$ with coefficient matrix quadruple $\{A(t), B(t), C(t), D(t)\}$ is observable on $[t_0, T]$ if and only if

$$\mathcal{N}[t_0, T] = 0$$

(12.9)

In view of Theorem 27, we see that observability of $\Sigma$ on $[t_0, T]$ is equivalent to kernel $G_O(t_0, T) = 0$. Since the latter, in turn, is equivalent to $\det G_O(t_0, T) \neq 0$, we are led to the following alternative characterization of observability.

**Corollary 8** A linear system with coefficient matrix quadruple $\{A(t), B(t), C(t), D(t)\}$ is observable on $[t_0, T]$ if and only if its observability Gramian $G_O(t_0, T)$ is nonsingular.

**Proof of Theorem 28:** As we’ve already explained, just before the statement of Theorem 28, observability of $\Sigma$ on $[t_0, T]$ implies that $\mathcal{N}[t_0, T] = 0$. To prove the converse, suppose that $\mathcal{N}[t_0, T] = 0$, that $x_0$ is any given initial state of $\Sigma$ at $t = t_0$, and that $v$ is any input which is defined and piece-wise continuous on $[t_0, T]$. Then by the variation of constants formula

$$y(t) = C(t)\Phi(t, t_0)x_0 + \int_{t_0}^{t} W(t, \tau)v(\tau)d\tau + D(t)v(t), \quad t \in [t_0, T]$$

Therefore

$$z(t) = C(t)\Phi(t, t_0)x_0, \quad t \in [t_0, T]$$

(12.10)

where

$$z(t) \overset{Δ}{=} y(t) − \int_{t_0}^{t} W(t, \tau)v(\tau)d\tau + D(t)v(t), \quad t \in [t_0, T]$$

Since $y$ and $v$ are known signals, so is $z$. To complete the proof, it is enough to show that we can solve (12.10) uniquely for $x_0$. Towards this end, we multiply both sides of (12.10) by $\Phi'(t, t_0)C'(t)$ and then integrate. What results is

$$\int_{t_0}^{T} \Phi'(t, t_0)C'(t)z(t)d\tau = \int_{t_0}^{T} \Phi'(t, t_0)C'(t)C(t)\Phi(t, t_0)x_0d\tau$$
In view of the definition of the observability Gramian in (12.6), this equation becomes

$$\int_{t_0}^{T} \Phi'(\tau, t_0) C'(\tau) z(\tau) d\tau = G_O(t_0, T)x_0$$  \hspace{1cm} (12.11)$$

But \(N[t_0, T] = 0\) by assumption and kernel \(G_O[t_0, T] = N[t_0, T]\) because of Theorem 27. Hence \(G_O[t_0, T]\) is nonsingular and thus invertible. It follows from (12.11) that \(x_0\) is uniquely recovered by the formula

$$x_0 = G_O^{-1}(t_0, T) \int_{t_0}^{T} \Phi'(\tau, t_0) C'(\tau) z(\tau) d\tau$$

Thus \(\Sigma\) is observable.

\[\textbf{12.2.1 Observability Reduction}\]

Let \(\Sigma \triangleq \{A(t), B(t), C(t)\}\) be a given \(n\)-dimensional continuous time linear system and let

$$W(t, \tau) = C(t)\Phi(t, \tau)B(\tau)$$

be its weighting pattern where \(\Phi(t, \tau)\) is the state transition matrix of \(A(t)\). Let \([t_0, T]\) be any interval of positive length on which \(\Sigma\) is defined. The aim of this section is to prove that every weighting pattern has an observable realization. We will do this constructively by explaining how to use \(\Sigma\) to compute a linear system \(\bar{\Sigma} \triangleq \{\bar{A}(t), \bar{B}(t), \bar{C}(t)\}\) which is observable on \([t_0, T]\) and which has the same weighting pattern as \(\Sigma\). Since \(\bar{\Sigma}\)'s dimension will turn out to be less than \(\Sigma\)'s, if \(\Sigma\) is not observable, this construction is referred to as an \textit{observability reduction}.

To begin, let us suppose right away that \(\Sigma\) is not observable. For if \(\Sigma\) were observable, we could simply define \(\bar{\Sigma}\) to be \(\Sigma\) and we would have at once an observable realization of \(W(t, \tau)\). To proceed, let \(G_O\) denote \(\Sigma\)'s observability Gramian on \([t_0, T]\). In other words

$$G_O \triangleq \int_{t_0}^{T} \Phi'(\tau, t_0) C'(\tau) C(\tau) \Phi(\tau, t_0) d\tau$$

As noted earlier when we discussed controllability, the symmetry of \(G_O\) implies that \(\mathbb{R}^n\) admits an orthogonal decomposition of the form

$$\mathbb{R}^n = \text{image } G_O \oplus \text{kernel } G_O$$

Let \(P\) denote the orthogonal projection matrix on image \(G_O\). Hence \(PG_O = G_O\) and therefore \(G_O P' = G_O\). Because \(G_O\) is positive-semidefinite it is possible to construct a full rank matrix \(R_{n \times \bar{n}}\), with \(\bar{n} \triangleq \text{rank } G_O\), such that

$$G_O = RR'$$

Our construction relies on the fact that \(P\) can be written as

$$P = R(R'R)^{-1}R'$$

Now consider the new system \(\bar{\Sigma}\) defined for \(t \in [t_0, T]\) by

$$\bar{A}(t) = 0 \quad \bar{B}(t) = R'\Phi(t_0, t)B(t) \quad \bar{C}(t) = C(t)\Phi(t, t_0)S$$

where

$$S \triangleq R(R'R)^{-1}$$
It will now be shown that $\bar{\Sigma}$ is observable and that its weighting pattern is $W(t, \tau)$. We establish observability first. Toward this end, note that $\bar{A}$'s state transition matrix is $\bar{\Phi}(t, \tau) = I_{\bar{n} \times \bar{n}}$. Using this fact we can write $\bar{\Sigma}$'s observability Gramian

\[ \bar{G}_O = \int_{t_0}^{T} \Phi'(t, t_0) \bar{C}'(t) \bar{C}(t) dt \]

as

\[ \bar{G}_O = \int_{t_0}^{T} \bar{C}'(t) \bar{C}(t) dt \]

Hence, using the definition of $\bar{C}(t)$

\[ \bar{G}_O = \int_{t_0}^{T} S' \Phi'(t, t_0) C'(t) C(t) \Phi(t, t_0) S dt \]

\[ = S' \left( \int_{t_0}^{T} \Phi'(t, t_0) C'(t) C(t) \Phi(t, t_0) dt \right) S \]

\[ = S' \bar{G}_O S \]

\[ = (R'R)^{-1} R' RR (R'R)^{-1} \]

\[ = I_{\bar{n} \times \bar{n}} \]

This prove that $\bar{G}_O(t_0, T)$ is nonsingular. Therefore by Corollary 8, $\bar{\Sigma}$ is observable on $[t_0, T]$ as claimed.

To prove that $\bar{\Sigma}$'s weighting pattern is the same as $\Sigma$'s, we first need to show that

\[ C(t) \Phi(t, t_0) P = C(t) \Phi(t, t_0), \quad t \in [t_0, T] \] (12.12)

Now this will be true provided

\[ \int_{t_0}^{T} ||(C(t) \Phi(t, t_0)' P - C(t) \Phi(t, t_0)) z||^2 dt = 0, \quad \forall z \in \mathbb{R}^n \] (12.13)

Fixing $z$ arbitrarily and expanding the integral on the left above we get

\[ \int_{t_0}^{T} ||(C(t) \Phi(t, t_0) P - C(t) \Phi(t, t_0)) z||^2 dt = \int_{t_0}^{T} z' P' \Phi'(t, t_0) C'(t) C(t) \Phi(t, t_0) P z dt \]

\[ - \int_{t_0}^{T} z' \Phi'(t, t_0) C'(t) C(t) \Phi'(t, t_0) P z dt \]

\[ - \int_{t_0}^{T} z' P' \Phi'(t, t_0) C'(t) C(t) \Phi'(t, t_0) z dt \]

\[ + \int_{t_0}^{T} z' \Phi'(t, t_0) C'(t) C(t) \Phi'(t, t_0) z dt \]

\[ = z' P' \bar{G}_O P z - z' \bar{G}_O P z - z' P' \bar{G}_O z + z' \bar{G}_O z \]

\[ = z' \bar{G}_O z - z' \bar{G}_O z + z' \bar{G}_O z \]

\[ = 0 \]

Hence (12.13) is true. Consequently (12.12) is true as well.

It remains to be shown that $\bar{\Sigma}$'s weighting pattern is the same as $\Sigma$'s. This is now easily accomplished by noting that $\bar{\Sigma}$'s weighting pattern $\bar{W}$ satisfies

\[ \bar{W}(t, \tau) = \bar{C}(t) \bar{\Phi}(t, \tau) \bar{B}(\tau) \]

\[ = \bar{C}(t) \bar{B}(\tau) \]
Therefore \( \bar{\Sigma} \) and \( \Sigma \) have the same weighting pattern as claimed.

We summarize.

**Theorem 29** Every continuous-time weighting pattern admits an observable realization.

Suppose now that \( \Sigma \) is controllable. Our aim is to show that controllability is not lost under an observability reduction. In other words, we want to verify that \( \bar{\Sigma} \) is controllable. To accomplish this we look at \( \bar{\Sigma} \)'s controllability Gramian

\[
\bar{G}_C = \int_{t_0}^{T} \bar{B}(t)\bar{B}'(t)dt
\]

Since \( \bar{\Phi}(t, \tau) = I \),

\[
\bar{G}_C = \int_{t_0}^{T} \bar{B}(t)\bar{B}'(t)dt
\]

Hence, using the definition of \( \bar{B}(t) \)

\[
\bar{G}_C = \int_{t_0}^{T} R\Phi(t_0, t)B(t)B'(t)\Phi'(t_0, t)Rdt
\]

\[
= \int_{t_0}^{T} R\Phi(t_0, T)\Phi(T, t)B(t)B'(t)\Phi'(t_0, T)Rdt
\]

\[
= R\Phi(t_0, T)\left( \int_{t_0}^{T} \Phi(t_0, t)B(t)B'(t)\Phi'(t_0, t)dt \right) \phi'(t_0, T)R
\]

\[
= R\Phi(t_0, T)G_C(t_0, T)\Phi'(t_0, T)R
\]

But \( R_{n \times n} \) has full rank and \( \Phi(t_0, T) \) is nonsingular so \( \Phi'(t_0, T)R \) has full rank as well. Moreover, controllability of \( \Sigma \) implies that \( G_C(t_0, T) \) is positive definite. Hence \( R\Phi(t_0, T)G_C(t_0, T)\Phi'(t_0, T)R \) is also positive definite and thus nonsingular. This proves that \( G_C(t_0, T) \) is nonsingular and thus that \( \bar{\Sigma} \) is controllable. Thus we see that reducing a controllable system to an observable system results in a system which is controllable as well.

It is now clear that we have a procedure for realizing any weighting pattern as a controllable and observable system. First realize the weighting pattern. Second reduce the realization to one which is controllable. Third reduce the controllable realization to one which is observable. It can be shown that the process also works if steps two and three are interchanged. We summarize.

**Theorem 30** Each weighting pattern admits a controllable and observable realization.

In the next section we prove that this procedure also provides us with a “minimal realization” of any weighting pattern.

### 12.3 Minimal Systems

Let us agree to call a system \( \Sigma \overset{\Delta}{=} \{A(t), B(t), C(t)\} \) minimal if the dimension of \( \Sigma \)'s state space is at least as small as the dimension of the state space of any linear system with the same weighting pattern as \( \Sigma \). Thus
\( \Sigma \) is minimal if its dimension is the least among the dimensions of all linear systems with the same weighting pattern. The following theorem is the centerpiece of linear realization theory.

**Theorem 31** A linear system is minimal if and only if it is controllable and observable.

**Proof:** If \( \Sigma \) is minimal then it must be both controllable and observable. For it if were not controllable \{observable\}, a lower dimensional system with the same weighting pattern could be constructed via a controllability \{observability\} reduction.

To prove the converse, suppose that \( \Sigma \) is controllable and observable and that \( \bar{\Sigma} \triangleq \{ \bar{A}(t), \bar{B}(t), \bar{C}(t) \} \) any linear system with the same weighting pattern as \( \Sigma \). Let \( n \) and \( \bar{n} \) denote the state space dimensions of \( \Sigma \) and \( \bar{\Sigma} \) respectively. It is enough to prove that \( n \leq \bar{n} \).

Since \( \Sigma \) and \( \bar{\Sigma} \) have the same weighting pattern,

\[
C(t)\Phi(t, \tau)B(\tau) = \bar{C}(t)\bar{\Phi}(t, \tau)\bar{B}(\tau)
\]

Using the composition rule for state transition matrices we can write

\[
C(t)\Phi(t, t_0)\Phi(t_0, T)\Phi(T, \tau)B(\tau) = \bar{C}(t)\bar{\Phi}(t, t_0)\bar{\Phi}(t_0, T)\bar{\Phi}(T, \tau)\bar{B}(\tau)
\]

Multiplying both sides of this equation on the left by \( \Phi'(t, t_0)C'(t) \) and on the right by \( B'(\tau)\Phi'(T, \tau) \) one gets

\[
\Phi'(t, t_0)C'(t)C(t)\Phi(t, t_0)\Phi(t_0, T)\Phi(T, \tau)B(\tau)B'(\tau)\Phi'(T, \tau) = \Phi'(t, t_0)C'(t)\bar{C}(t)\bar{\Phi}(t, t_0)\bar{\Phi}(t_0, T)\bar{\Phi}(T, \tau)\bar{B}(\tau)B'(\tau)\Phi'(T, \tau)
\]

Integrating both sides with respect to both \( t \) and \( \tau \) there follows

\[
G_O(t_0, T)\Phi(t_0, T)G_C(t_0, T) = M\bar{\Phi}(t_0, T)N \tag{12.14}
\]

where

\[
G_O(t_0, T) \triangleq \int_{t_0}^{T} \Phi'(t, t_0)C'(t)C(t)\Phi(t, t_0)dt \quad G_C(t_0, T) \triangleq \int_{t_0}^{T} \Phi(t, \tau)B(\tau)B'(\tau)\Phi'(T, \tau)d\tau
\]

\[
M_{n \times \bar{n}} \triangleq \int_{t_0}^{T} \Phi'(t, t_0)C'(t)\bar{C}(t)\bar{\Phi}(t, t_0)dt \quad N_{\bar{n} \times n} \triangleq \int_{t_0}^{T} \bar{\Phi}(T, \tau)\bar{B}(\tau)B'(\tau)\Phi'(T, \tau)d\tau
\]

Since \( \Sigma \) is controllable and observable, both \( G_C(t_0, T) \) and \( G_O(t_0, T) \) are nonsingular \( n \times n \) matrices. Since \( \Phi(t_0, T) \) also is nonsingular, the matrix on the left in (12.14) is nonsingular. It follows that \( M\bar{\Phi}(t_0, T)N \) is nonsingular or equivalently that

\[
\text{rank } M\bar{\Phi}(t_0, T)N = n
\]

But \( \bar{\Phi}(t_0, T) \) is an \( \bar{n} \times \bar{n} \) matrix so \( n \leq \bar{n} \).

### 12.4 Time-Invariant Continuous-Time Linear Systems

While the constructions regarding observability discussed so far, apply to both time-varying and time-invariant linear systems, the consequential simplications of the time-invariant case are significant enough to justify a separate treatment.

Let \( \Sigma \triangleq \{ A, B, C, D \} \) be a time-invariant system. Then \( \Sigma \)'s observability Gramian on \([t_0, T]\) is
\[ G_C(t_0, T) \overset{\Delta}{=} \int_{t_0}^{T} e^{(\tau-t_0)A'}C'Ce^{(\tau-t_0)A}d\tau \]

By introducing the change of variables \( s = \tau - t_0 \) we see that

\[ G_C(t_0, T) = \int_{0}^{T-t_0} e^{sA'C'Ce^{sA}}ds \]

Since the right hand side of the preceding is also the observability Gramian of \( \Sigma \) on \([0, T-t_0]\), we can conclude that in the time-invariant case the observability Gramian of \( \Sigma \) on \([t_0, T]\) depends only on the difference \( T-t_0 \), and not independently on \( T \) and \( t_0 \). For this reason, in the time-invariant case \( t_0 \) is usually taken to be zero. We will adopt this practice in the sequel.

As we’ve already seen, the kernel of \( \Sigma \)’s observability Gramian on a given interval \([t_0, T]\) of positive length is the subspace of unobservable states \( N[t_0, T] \) of \( \Sigma \) on \([t_0, T]\). Evidently this subspace too, only depends on the time difference \( T-t_0 \), and not separately on \( t_0 \) and \( T \). In fact, \( N[t_0, T] \) doesn’t even depend on \( T-t_0 \), provided this difference is a positive number. Our aim is to explain why this is so and at the same time to provide an explicit algebraic characterization of \( N[t_0, T] \) expressed directly in terms of \( C \) and \( A \). Towards this end, let us define the subspace

\[ [C|A] \overset{\Delta}{=} \bigcap_{i=1}^{n} \ker CA^{(i-1)} \]

where \( n \) is the dimension of \( \Sigma \).

Let us also agree to call the \( pn \times n \) matrix

\[
\begin{bmatrix}
C \\
CA \\
\vdots \\
CA^{(n-1)}
\end{bmatrix}
\]

the \textit{observability matrix} of \( \Sigma \) or of the pair \((C, A)\). We see at once that \([C|A]\) is the kernel of the observability matrix of \((C, A)\). The following theorem establishes the significance these definitions.

\textbf{Theorem 32} For all \( T > t_0 \geq 0 \)

\[ N[t_0, T] = [C|A] \quad (12.15) \]

The theorem implies that the set of states of a time-invariant which are unobservable on any interval of positive length, does not depend on either the length of the interval or on when the interval begins. In view of Theorem 32 we will also call \([C|A]\) the \textit{unobservable space} of \( \Sigma \) or of the pair \((C, A)\).

The proof of Theorem 32 depends on the following lemma.

\textbf{Lemma 7} For each fixed time \( t_1 > 0 \)

\[ [C|A] = \ker \int_{0}^{t_1} e^{At'C'Ce^{At}}dt \quad (12.16) \]
Proof of Theorem 32: By definition,

\[ G_O(t_0, T) = \int_{t_0}^{T} e^{(\tau-t_0)A'}C'e^{(\tau-t_0)A}d\tau \]

By making the change of variable \( t = \tau - t_0 \) we see that

\[ G_O(t_0, T) = \int_{0}^{(T-t_0)} e^{A't}C'e^{A}dt \]

From this and Theorem 27 we see that

\[ N[t_0, T] = \text{kernel} \int_{0}^{(T-t_0)} e^{A't}C'e^{A}dt \]

Application of Lemma 7 with \( t_1 \triangleq T - t_0 \) then yields (12.15) which is the desired result. \( \blacksquare \)

Proof of Lemma 7: Fix \( t_1 > 0 \) and define

\[ M = \int_{0}^{t_1} e^{A't}C'C'e^{A}dt \]  (12.17)

As pointed out earlier in the math notes, \( e^{tA} \) can be expressed as

\[ e^{tA} = \sum_{i=1}^{n} \gamma_i(t)A^{(i-1)} \]

for suitable functions \( \gamma_i(t) \). It is therefore possible to write \( M \) as

\[ M = \int_{0}^{t_1} \sum_{i=1}^{n} \gamma_i(t)e^{A't}CA^{(i-1)}dt \]

Therefore

\[ M = \sum_{i=1}^{n} R_iCA^{(i-1)} \]

where

\[ R_i = \int_{0}^{t_1} \gamma_i(t)e^{A't}C'dt, \quad i \in \{1, 2, \ldots, n\} \]

It follows that

\[ \text{kernel} \ M = \text{kernel} \sum_{i=1}^{n} R_iCA^{(i-1)} \supset \bigcap_{i=1}^{n} \text{kernel} \ R_iCA^{(i-1)} \supset \bigcap_{i=1}^{n} \text{kernel} \ CA^{(i-1)} = [C|A] \]

Thus \( \text{kernel} \ M \supset [C|A] \).

For the reverse inclusion, let \( x \) be any vector in \( \text{kernel} \ M \). Then \( Mx = 0 \), \( x'Mx = 0 \),

\[ x'\left( \int_{0}^{t_1} e^{A't}C'e^{A}dt \right) x = 0 \]

and so

\[ \int_{0}^{t_1} ||Ce^{A}x||^2 dt = 0 \]

It follows that

\[ Ce^{A}x = 0, \quad \forall t \in [0, t_1] \]
Repeated differentiation yields

\[ C(A)^{(i-1)}e^{At}x = 0, \quad \forall t \in [0, t_1], \quad i \in \{1, 2, \ldots, n\} \]

Evaluation at \( t = 0 \) thus provides

\[ C(A)^{(i-1)}x = 0, \quad i \in \{1, 2, \ldots, n\} \]

Therefore \( x \in [C|A] \) and since \( x \) was chosen arbitrarily in kernel \( M \), it must be true that kernel \( M \subset [C|A] \). This is the required reverse containment.

In the light of Corollary 8 we can now state

**Theorem 33** An \( n \)-dimensional time invariant linear system \( \{A, B, C, D\} \) is observable if and only if

\[ [C|A] = 0 \]

From this and the fact that the unobservable space of \( (C, A) \) is the kernel of the observability matrix of \( (C, A) \) we can also state the following.

**Corollary 9** An \( n \)-dimensional time invariant linear system \( \{A, B, C, D\} \) is observable if and only if

\[
\begin{align*}
\text{rank} \begin{bmatrix}
C \\
CA \\
CA^2 \\
\vdots \\
CA^{(n-1)}
\end{bmatrix} &= n
\end{align*}
\]

It has by now no doubt occured to the reader that there are certain algebraic similarities between the ideas we’ve been discussing in this chapter and those discussed earlier in the chapter on controllability. For example, the transpose of the observability matrix of \( (C, A) \) is the controllability matrix of the transposed pair \( (A', C') \). The unobservable space of \( (C, A) \) is the orthogonal complement of the controllable space of \( (A', C') \). The controllability Gramian of \( (A, B) \) and the observability Gramian of \( (C, A) \) are similar in form. While these similarities are especially useful in gaining insights into proofs and constructions, they actually have no *system theoretic* significance. We sometimes say that controllability and observability are dual concepts. It is possible to be more precise about this by appealing to an idea from linear algebra called “duality,” but the abstraction one must envoke hardly seems worth the effort\(^1\).

### 12.4.1 Properties of the Unobservable Space of \( (C, A) \)

Like the controllable space of \( (A, B) \), the unobservable space of \( (C, A) \) has several important properties which we will make use of in the next section. The first, and perhaps most important of these is that \( [C|A] \) is an \( A \)-invariant subspace.

**Lemma 8** For each matrix pair \( (C_{p \times n}, A_{n \times n}) \),

\[ A[C|A] \subset [C|A] \quad (12.18) \]

\(^1\)The concept of duality being eluded to here, should not be confused with the very different and very useful concept of duality which arises in convex analysis and linear programming.
Proof: Fix $x \in [C|A]$. Thus
\[ CA^{(i-1)}x = 0, \quad i \in \{1, 2, \ldots, n\} \] (12.19)

Thus
\[ CA^{(i-1)}(Ax) = 0, \quad i \in \{1, 2, \ldots, n-1\} \] (12.20)

By the Cayley-Hamilton theorem $A^n = -a_n A^{(n-1)} - \cdots - a_2 A - a_1 I$, where $s^n + a_n s^{(n-1)} + \cdots a_2 s + a_1$ is the characteristic polynomial of $A$. Thus
\[ CA^{(n-1)}(Ax) = -a_n CA^{(n-1)}x - \cdots - a_2 CAx - a_1 Cx \]
so from (12.19), $CA^{(n-1)}(Ax) = 0$. From this and (12.20) it follows that
\[ CA^{(i-1)}(Ax) = 0, \quad i \in \{1, 2, \ldots, n\} \]
and thus that $Ax \in [C|A]$. Since this is true for all $x \in [C|A]$, (12.18) must be true. ■

By the observability index of $(C_{p \times n}, A_{n \times n})$ is meant the smallest positive integer $n_o$ for which
\[ [C|A] = \bigcap_{i=1}^{n_o} \text{kernel } CA^{(i-1)} \]

Clearly $n_o \leq n$, with the inequality typically being strict if $p > 1$. The following is a consequence of the preceding lemma and the Cayley-Hamilton theorem.

**Lemma 9** For each matrix pair $(C_{p \times n}, A_{n \times n})$,
\[ \text{rank } \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{(k-1)} \end{bmatrix} = \text{rank } \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{(n-1)} \end{bmatrix} \quad \forall k \geq n_o \] (12.21)

where $n_o$ is the observability index of $(C, A)$.

This lemma can be proved by exploiting duality and appealing to Lemma 4. The first step in this direction would be to replace (12.21) with the equivalent expression
\[ <A'|C'> = C' + A'C' + \cdots + (A')^{(k-1)}C' \]

where $C' \triangleq \text{image } C'$. The reader may wish to develop a proof along these lines.

There is an interesting way to characterize an unobservable space, which differs from what we’ve discussed so far. Let $(C, A)$ be given and fixed. Suppose we define $\mathcal{C}$ to be the class of all subspaces of $\mathbb{R}^n$ which are $A$-invariant and which are contained in $\text{kernel } C$. That is
\[ \mathcal{C} \triangleq \{ S : AS \subset S, \quad S \subset \text{kernel } C \} \]

Note that the zero subspace is in $\in \mathcal{C}$ so $\mathcal{C}$ is nonempty. Moreover it is easy to show that if $S \in \mathcal{C}$ and $T \in \mathcal{C}$, then $S + T \in \mathcal{C}$. In other words, $\mathcal{C}$ is closed under subspace addition. Because of this and the finite dimensionality of $\mathbb{R}^n$, it is possible to prove that $\mathcal{C}$ contains a unique largest element which contains every other subspace in $\mathcal{C}$. In fact, $[C|A]$ turns out to be this unique largest subspace. In other words, $[C|A]$ is contained in kernel $C$, is $A$-invariant, and contains in every other subspace which is contained in kernel $C$ and is $A$-invariant. The reader may wish to prove that this is so.
12.4.2 Observability Reduction for Time-Invariant Systems

Although the observability reduction technique described earlier, also works in the time-invariant case, it is possible to accomplish the same thing in a more algebraic manner without having to use the observability Gramian. What we shall do is to explain how to construct from a given time-invariant linear system, a new time-invariant linear system which is observable and which has the same transfer matrix as the system we started with. To carry out this program we shall make use of the following fundamental algebraic fact.

**Lemma 10** Let $M_{n \times m}$ and $N_{p \times m}$ be given real matrices. There exists a solution $X_{p \times n}$ to the linear algebraic equation

$$XM = N$$

if and only if

$$\ker M \subseteq \ker N$$

**Proof:** The solvability condition for the transposed equation $M'X' = N'$, is $\text{image } M' \supset \text{image } N'$. Taking orthogonal complements, this condition becomes

$$(\text{image } M')^\perp \subseteq (\text{image } N')^\perp$$

But $(\text{image } M')^\perp = \ker M$ and $(\text{image } N')^\perp = \ker N$. Substitution of these identities into (12.24) yields (12.23) which is the desired result. ■

We now turn to the problem of constructing from a given time-invariant linear system $\Sigma = \{A, B, C\}$, a new time-invariant linear system $\bar{\Sigma}$ which is observable and which has the same transfer matrix as $\Sigma$. As a first step, define $\bar{\Sigma}$’s dimension $\bar{n} \triangleq n - \dim [C|A]$. Next pick and full rank matrix $P_{\bar{n} \times n}$ such that $\ker P \subseteq \ker C$. An easy way to do this is to define $P'$ to be a basis matrix for $<A'|C'|$. We claim that $\ker P \subseteq \ker C \ker PA \subseteq \ker P$.

The first containment is a consequence of the fact that $[C|A] \subseteq \ker C$. The second containment is a consequence of the fact that $[C|A]$ is $A$-invariant. In any event, these two containments plus Lemma 10 imply that the equations

$$PA = \bar{A}P$$

$$C = \bar{C}P$$

have {unique} solutions $\bar{A}_{\bar{n} \times \bar{n}}$ and $\bar{C}_{p \times \bar{n}}$ respectively. Define $\bar{B} \triangleq PB$. In summary, we have defined a linear system $\bar{\Sigma} \triangleq \{\bar{A}, \bar{B}, \bar{C}\}$ whose coefficient matrices are related to those of $\Sigma = \{A, B, C\}$ by the equations

$$PA = \bar{A}P$$

$$\bar{B} = PB$$

$$C = \bar{C}P$$

In view of Lemma 5, we see at once that $\Sigma$ and $\bar{\Sigma}$ have the same transfer matrix.

We claim that $(\bar{C}, \bar{A})$ is an observable pair. To establish this, we first note from the fact $PA = \bar{A}P$, that $PA^i = \bar{A}^iP, \ i \geq 0$. Hence $CA^i = \bar{C}PA^i = \bar{C}\bar{A}^iP, \ i \geq 0$. Therefore

$$\begin{bmatrix}
C \\
CA \\
\vdots \\
CA^{(n-1)}
\end{bmatrix} = \begin{bmatrix}
\bar{C} \\
\bar{C}A \\
\vdots \\
\bar{C}A^{(n-1)}
\end{bmatrix} P$$

Thus

$$\text{rank} \begin{bmatrix}
C \\
\bar{C}A \\
\vdots \\
\bar{C}A^{(n-1)}
\end{bmatrix} \geq \text{rank} \begin{bmatrix}
C \\
CA \\
\vdots \\
CA^{(n-1)}
\end{bmatrix} = \bar{n}$$
But the matrix on the left has \( \bar{n} \) columns so its rank cannot be larger than \( \bar{n} \). Therefore

\[
\text{rank} \begin{bmatrix}
\bar{C} \\
\bar{C}A \\
\vdots \\
\bar{C}A^{(n-1)}
\end{bmatrix} = \bar{n}
\]

Hence by Lemma 9

\[
\text{rank} \begin{bmatrix}
\bar{C} \\
\bar{C}A \\
\vdots \\
\bar{C}A^{(\bar{n}-1)}
\end{bmatrix} = \bar{n}
\]

so \((\bar{C}, \bar{A})\) is an observable pair. Therefore \( \bar{\Sigma} \) is an observable system with the same transfer matrix as \( \Sigma \).

Suppose now that \( \Sigma \) is controllable. Our aim is to show that controllability is not lost under an observability reduction, just as before in the time-varying case. In other words, we want to verify that \( \bar{\Sigma} \) is controllable. To accomplish this we again use the fact that \( PA^i = \bar{A}^iP, \ i \geq 0 \) together with the previous definition of \( \bar{B} \), namely \( \bar{B} = PB \), to write

\[
\begin{bmatrix}
\bar{B} \\
\bar{A}\bar{B} \\
\vdots \\
\bar{A}^{(n-1)}\bar{B}
\end{bmatrix} = P \begin{bmatrix}
B \\
AB \\
\vdots \\
A^{(n-1)}B
\end{bmatrix}
\]

Since \( \Sigma \) is controllable, \( \begin{bmatrix}
B \\
AB \\
\vdots \\
A^{(n-1)}B
\end{bmatrix} \) has linearly independent rows. So does \( P \). Since the product of two matrices which each have linearly independent rows, is a matrix with linearly independent rows, \( \begin{bmatrix}
\bar{B} \\
\bar{A}\bar{B} \\
\vdots \\
\bar{A}^{(\bar{n}-1)}\bar{B}
\end{bmatrix} \) must have linearly independent rows. This means that

\[
\text{rank} \begin{bmatrix}
\bar{B} \\
\bar{A}\bar{B} \\
\vdots \\
\bar{A}^{(\bar{n}-1)}\bar{B}
\end{bmatrix} = \bar{n}
\]

Hence by Lemma 4

\[
\text{rank} \begin{bmatrix}
\bar{B} \\
\bar{A}\bar{B} \\
\vdots \\
\bar{A}^{(\bar{n}-1)}\bar{B}
\end{bmatrix} = \bar{n}
\]

which establishes the controllability of \( \bar{\Sigma} \). Thus we see that reducing a controllable system to an observable system in this manner results in a system which is controllable, just as in the time-varying case.

It is now clear that once we have we have a procedure for realizing a transfer matrix as a time-invariant linear system, we will also have a procedure for constructing a realization which is both controllable and observable. First realize the transfer matrix as a time-invariant system.. Second reduce the realization to one which is controllable. Third reduce the controllable realization to one which is observable. It can be shown that the process also works if steps two and three are interchanged. We will return to this important topic in the next chapter where we explain how to realize a transfer matrix.

### 12.4.3 Observable Decomposition

The preceding discussion was concerned solely with the problem of constructing from a given linear system, a new observable linear system with the same transfer matrix as the original. Our aim now is to represent the original system

\[
\dot{x} = Ax + Bu \\
y = Cx
\]

in a new coordinate system within which \( \{\bar{A}, \bar{B}, \bar{C}\} \) appears naturally as a subsystem. Toward this end, let \( S \) be any \((n-\bar{n}) \times n\) matrix for which \( T = \begin{bmatrix} P \\ S \end{bmatrix} \) is nonsingular. Define

\[
\begin{bmatrix}
\bar{A}_{(n-\bar{n}) \times \bar{n}} \\
\hat{A}_{(n-\bar{n}) \times (n-\bar{n})}
\end{bmatrix} \overset{\Delta}{=} SAT^{-1} \\
\hat{B} \overset{\Delta}{=} SB
\]
Recall that
\[ PA = \bar{A}P, \quad \bar{B} = PB \quad C = \bar{C}P \]

From these expressions it follows that
\[ TAT^{-1} = \begin{bmatrix} \bar{A} & 0 \\ \bar{A} & \bar{A} \end{bmatrix} \quad TB = \begin{bmatrix} \bar{B} \\ \bar{B} \end{bmatrix} \quad CT^{-1} = \begin{bmatrix} \bar{C} & 0 \end{bmatrix} \]

Hence if we define \( z = Tx \), and partition \( z \) as
\[ z \triangleq \begin{bmatrix} \bar{z} \\ \hat{z} \end{bmatrix} \]
where \( \bar{z} \) is a \( \bar{n} \)-vector, then we can write
\[
\dot{\bar{z}} = \bar{A}\bar{z} + \bar{B}u \quad (12.27) \\
\dot{\hat{z}} = \hat{A}\bar{z} + \hat{A}\hat{z} + \hat{B}u \quad (12.28) \\
y = \bar{C}\bar{z} \quad (12.29)
\]

Note that \( y \) does not depend on \( \hat{z} \) and moreover that if \( \hat{z}(0) = 0 \), then the input-output relationship is determined solely by the observable subsystem
\[ \dot{\bar{z}} = \bar{A}\bar{z} + \bar{B}u \quad y = \bar{C}\bar{z} \]

The spectrum of \( \hat{A} \) is is uniquely determined by \((C, A)\) and is sometimes called the unobservable spectrum of \((C, A)\).
Chapter 13

Transfer Matrix Realizations

The transfer matrix of a \( m \)-input, \( p \)-output, \( n \)-dimensional, time-invariant linear system \( \{A, B, C, D\} \) is a \( p \times m \) matrix of proper rational functions defined by the formula

\[
T(s) \triangleq C(sI - A)^{-1}B + D
\]  

Any \( m \)-input, \( p \)-output linear system \( \{A, B, C, D\} \) of any dimension for which (13.1) holds, is a realization of \( T(s) \). The aim of this chapter is to explain how to go about constructing realizations of any given matrix of proper rational functions. We begin with transfer functions.

13.1 Realizing a Transfer Function

Suppose we are given a proper transfer function

\[
t(s) = \frac{\beta(s)}{\alpha(s)}
\]

with \( \alpha(s) \) monic, which we wish to realize. The first step is to separate it into its strictly proper part and what’s left: Define

\[
d = \lim_{s \to \infty} \frac{\beta(s)}{\alpha(s)}
\]

Note that \( d \) must be finite because of properness. Moreover if we define

\[
\gamma(s) = \beta(s) - d
\]

then

\[
\frac{\beta(s)}{\alpha(s)} = \frac{\gamma(s)}{\alpha(s)} + d
\]

and

\[
\frac{\gamma(s)}{\alpha(s)}
\]

is strictly proper. There are many ways to realize strictly proper rational functions. We shall discuss two methods.
13.1.1 Realization Using the Control Canonical Form

Our aim is to realize the strictly proper transfer function

\[ \frac{\gamma(s)}{\alpha(s)} \]

For this, let \((A, b)\) be the control canonical form determined by

\[ \alpha(s) = s^n + a_n s^{n-1} + \cdots + a_2 s + a_1 \]

That is

\[
A = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
-a_1 & -a_2 & -a_3 & \cdots & -a_n
\end{bmatrix}_{n \times n}, \quad b = \begin{bmatrix} 0 \\
0 \\
\vdots \\
0 \\
1 \end{bmatrix}_{n \times 1}
\]

Note that because of the form of this pair,

\[
(sI - A)^{-1}b = \frac{1}{\alpha(s)} \begin{bmatrix} 1 \\
s \\
\vdots \\
s^{n-1} \end{bmatrix}
\]

which is a formula well worth remembering.

Suppose

\[ \gamma(s) = g_n s^{(n-1)} + \cdots + g_2 s + g_1 \]

Define

\[ c = \begin{bmatrix} g_1 & g_2 & \cdots & g_n \end{bmatrix} \]

Then because of (13.2),

\[ c(sI - A)^{-1}b = \frac{\gamma(s)}{\alpha(s)} \]

which means that \(\{A, b, c, d\}\) is a realization of the original transfer function

\[ \frac{\beta(s)}{\alpha(s)} \]

13.1.2 Realization Using Partial Fractions

Another way to realize a transfer function, which is fundamentally different that the way just described, is to expand the transfer function in partial fractions, then realize each term in the expansion and then put them together in the appropriate way. The disadvantage of this method compared with what was just done
13.1. REALIZING A TRANSFER FUNCTION

is clear: the latter requires one to find all the roots of \( \alpha(s) \), whereas the former does not. Despite this, the partial fraction approach provides several insights into the realization process and for this reason will be sketched in the sequel.

Suppose for simplicity, that \( \alpha(s) \) has distinct \( n \) roots \( \lambda_1, \ldots, \lambda_n \). That is,

\[
\alpha(s) = (s - \lambda_1)(s - \lambda_2) \cdots (s - \lambda_n)
\]

where \( \lambda_i \neq \lambda_j \) if \( i \neq j \). Then the partial fraction expansion of the above transfer function would be

\[
\frac{\beta(s)}{\alpha(s)} = d + \frac{r_1}{(s - \lambda_1)} + \cdots + \frac{r_n}{(s - \lambda_n)}
\]

where \( d \) is as determined before and \( r_i \) is the residue

\[
r_i = \frac{\beta(s)}{\alpha(s)} \bigg|_{s=\lambda_i}
\]

The next step is to realize each of the transfer functions in the sum. A simple one-dimensional realization of the \( i \)th such transfer function is

\[
A_i \triangleq \lambda_i \quad b_i \triangleq 1 \quad c_i \triangleq r_i
\]

(13.3)

Of course

\[
A_i \triangleq \lambda_i \quad b_i \triangleq r_i \quad c_i \triangleq 1
\]

will work just as well. In fact there are infinitely many ways to construct such a one-dimensional realization. We leave it to the reader to verify that no matter which of these realizations we use for each \( i \), if we define

\[
A \triangleq \text{diagonal } \{A_1, A_2, \ldots, A_n\}_{n \times n} \quad b \triangleq \text{column } \{b_1, b_2, \ldots, b_n\} \quad c \triangleq [c_1 \ c_2 \ \cdots \ c_n]
\]

then \( \{A, b, c, d\} \) realizes the given transfer function. Note that for any such realization

\[
A = \begin{bmatrix}
\lambda_1 & 0 & 0 & 0 \\
0 & \lambda_2 & 0 & 0 \\
\vdots & 0 & \ddots & \vdots \\
0 & \cdots & 0 & \lambda_n
\end{bmatrix}
\]

Note also that some of the \( \lambda_i \) and \( r_i \) may be complex valued, in which case the preceding will be a complex realization. If one starts with a real transfer function, complex realizations can always be avoided by first taking care of complex terms in the original expansion. Suppose for example that \( r_j \) and \( \lambda_j \) are non-real numbers. Then because the transfer function under consideration has real coefficients, there must also be an integer \( k \) such that \( \lambda_k = \lambda_j^* \) and \( r_k = r_j^* \) where \( * \) denotes complex conjugate. If this is so then it is always possible to write

\[
\frac{r_j}{(s - \lambda_j)} + \frac{r_k}{(s - \lambda_k)} = \frac{(f_1 s + f_2)}{(s^2 - 2\sigma s + \sigma^2 + \omega^2)}
\]

(13.4)

where the \( f_1, \sigma \) and \( \omega \) are real numbers and \( \lambda_j = \sigma + \omega \sqrt{-1} \). It can easily be seen that the real transfer function in (13.4) is realized \{for example\} by the linear system

\[
A = \begin{bmatrix}
\sigma & -\omega \\
\omega & \sigma
\end{bmatrix} \quad b = \begin{bmatrix} 0 \\
1
\end{bmatrix} \quad c = [f_1 \ \sigma f_2 + f_3]
\]

We leave it to the reader to work out how to realize a sum consisting of real transfer function of this type and real transfer function whose denominators are of degree one. The reader may also wish to consider the general case when \( \alpha(s) \) has repeated roots.
13.2 Realizing a $p \times 1$ Transfer Matrix Using the Control Canonical Form

Both of the preceding methods easily generalize to the realization of a column vector of transfer functions. We shall only discuss the extension of the first method.

Suppose we are given a $p \times 1$ column vector $T(s)$ of proper transfer functions. Without loss of generality we can assume that all of the transfer functions in $T(s)$ have the same monic common denominator $\alpha(s)$:

$$T(s) = \begin{bmatrix} \frac{\beta_1(s)}{\alpha(s)} \\ \frac{\beta_2(s)}{\alpha(s)} \\ \vdots \\ \frac{\beta_p(s)}{\alpha(s)} \end{bmatrix}$$

Proceding along the same lines as before in the case of a single transfer function, we can write

$$T(s) = H(s) + d$$

where $d$ is the constant $p$-vector

$$d \triangleq \lim_{s \to \infty} T(s)$$

and $H(s)$ is the strictly proper transfer matrix defined by

$$H(s) \triangleq T(s) - d$$

Suppose $H(s)$ of the form

$$H(s) = \begin{bmatrix} \frac{\gamma_1(s)}{\alpha(s)} \\ \frac{\gamma_2(s)}{\alpha(s)} \\ \vdots \\ \frac{\gamma_p(s)}{\alpha(s)} \end{bmatrix}$$

Let $A$ and $b$ be the control canonical form determined by $\alpha(s)$. For $i \in \{1, 2, \ldots, p\}$, define $c_i$ so that

$$c_i(sI - A)^{-1}b = \frac{\gamma_i(s)}{\alpha(s)}$$

much as was done before for a scalar transfer function. It follows that if we define

$$C \triangleq \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_p \end{bmatrix}$$

then $\{A, b, C\}$ will realize $H(s)$ and consequently $\{A, b, C, d\}$ will realize $T(s)$. Note that the realization is controllable.
13.3 Realizing a $p \times m$ Transfer Matrix

Now suppose that $T(s)$ is a $p \times m$ matrix. In other words

$$T(s) = \begin{bmatrix} T_1(s) & T_2(s) & \cdots & T_m(s) \end{bmatrix}$$

where each $T_i(s)$ is a $p \times 1$ matrix of proper transfer functions. Using (for example) the method of the preceding section, we can find, for each $i \in \{1, 2, \ldots, m\}$, a $n_i$-dimensional realization $\{A_i, b_i, C_i, d_i\}$ of $T_i(s)$. The reader should carefully verify that

$$A \triangleq \text{block diagonal } \{A_1, A_2, \ldots, A_m\}$$

$$B \triangleq \text{block diagonal } \{b_1, b_2, \ldots, b_m\}$$

$$C \triangleq \begin{bmatrix} C_1 & C_2 & \cdots & C_m \end{bmatrix}$$

$$D \triangleq \text{block diagonal } \{d_1, d_2, \ldots, d_m\}$$

is then a realization of $T(s)$. In fact this realization is even controllable. We have proved the following.

**Theorem 34** A matrix of rational functions is realizable by a finite-dimensional, time-invariant, linear dynamical system if and only if each of its elements is a proper rational function.

We now know how

1. to realize any transfer matrix.
2. to reduce any time-invariant linear system to a controllable time-invariant linear system with the same transfer matrix.
3. to reduce any time-invariant linear system to an observable time-invariant linear system with the same transfer matrix.

Moreover we know that controllability is not lost if we reduce a time-invariant controllable linear system to an observable time-invariant linear system with the same transfer matrix. Thus the preceding three steps, performed in sequence, will provide a controllable and observable time-invariant realization of any matrix of proper rational functions. It can be shown that the modified procedure obtained by interchanging steps 2 and 3, also produces a controllable and observable time-invariant realization of any matrix of proper rational functions. We summarize.

**Theorem 35** Each matrix of proper rational functions admits a time-invariant controllable and observable realization.

13.4 Minimal Realizations

Let us agree to say that $\Sigma \triangleq \{A, B, C, D\}$ is a minimal realization of a given transfer matrix $T(s)$ if $\Sigma$ realizes $T(s)$ and if the dimension of $\Sigma$’s state space is as least as small as the dimension of the state space of any other time-invariant linear system which realizes $T(s)$. Note that $\Sigma$ is minimal in the sense defined earlier in Chapter 3 because two time-invariant linear systems have the same transfer matrix just in case they have the same weighting pattern. In view of Theorem 31, we see that a minimal realization of $T(s)$ is the same as a controllable and observable realization of $\Sigma$. Thus the three steps enumerated at the end of the last section provide a procedure for constructing a minimal realization of any given transfer matrix.
13.5 Isomorphic Systems

Let us agree to say that two $m$-input, $n$-dimensional, $p$-output, time-invariant systems $\Sigma \triangleq \{A, B, C, D\}$ and $\bar{\Sigma} \triangleq \{\bar{A}, \bar{B}, \bar{C}, D\}$ are similar systems or isomorphic systems if there exists a nonsingular matrix $Q$ such that

$$QAQ^{-1} = \bar{A} \quad QB = \bar{B} \quad CQ^{-1} = \bar{C}$$

System similarity is an equivalence relation on the class of all $m$-input, $n$-dimensional, $p$-output, time-invariant systems. The transfer matrix of a linear system is invariant under system similarity – but it is not necessarily a complete invariant. In other words, while two similar linear systems necessarily have the same transfer matrix, it is not true in general that two systems with the same transfer matrix are similar, even if both have the same dimension. On the other hand, two controllable and observable linear systems with the same transfer matrix are in fact similar. This is a statement of the so-called “Isomorphism Theorem.”

**Theorem 36 (Isomorphism)** Two controllable and observable linear systems with the same transfer matrix are isomorphic.

The theorem implies that two controllable and observable linear systems with the same transfer matrix are “essentially the same” in that they differ from each other only by a change of state variables. The proof of Theorem 36 depends on the following lemma which is useful in its own right.

**Lemma 11** Let $\{\bar{A}, \bar{B}, \bar{C}\}$ and $\{A, B, C\}$ be the coefficient matrix triples of two linear systems with the same transfer matrix. Suppose that the former system is $\bar{n}$-dimensional and controllable and that the latter is $n$-dimensional and observable. Then

$$\bar{C} = CT \quad T\bar{A} = AT \quad B = T\bar{B}$$  \hspace{1cm} (13.5)

where $T$ the $n \times \bar{n}$ matrix

$$T \triangleq RR^{-1},$$

$R^{-1}$ being a right inverse of the the controllability matrix of $(\bar{A}, \bar{B})$, and $R \triangleq [B \ AB \cdots A^{\bar{n}-1}B]$.

**Proof of Theorem 36:** Let $\Sigma \triangleq \{\bar{A}, \bar{B}, \bar{C}\}$ and $\Sigma \triangleq \{A, B, C\}$ be two controllable and observable systems with the same transfer matrix. Then both systems are minimal with the same state space dimension $\bar{n} = n$. Let $T$ be the $n \times n$ matrix $T \triangleq RR^{-1}$, where $R^{-1}$ is a right inverse of the the controllability matrix of $(\bar{A}, \bar{B})$, and $R$ is the controllability matrix of $(A, B)$. By Lemma 11,

$$\bar{C} = CT \quad T\bar{A} = AT \quad B = T\bar{B}$$  \hspace{1cm} (13.6)

It is therefore enough to prove that $T$ is nonsingular.

From the latter two equations in (13.6) there follows $A^iB = T\bar{A}^i\bar{B}$, $i \geq 0$. Therefore $R = TR$. But rank $R = n$, so rank $T \geq n$. Since $T$ is $n \times n$, rank $T = n$ so $T$ is nonsingular. \hspace{1cm} $\blacksquare$

**Proof of Lemma 11:** Equality of the transfer matrices of $\{\bar{A}, \bar{B}, \bar{C}\}$ and $\{A, B, C\}$ implies that the transfer matrices’ Markov matrices are equal. Thus

$$\bar{C}\bar{A}^i\bar{B} = CA^iB, \quad i \geq 0$$

From the relations there follow

$$\bar{N}\bar{R} = NR \quad \bar{C}\bar{R} = CR \quad \bar{N}\bar{A}\bar{R} = NAR \quad \bar{N}\bar{B} = NB$$
where $N$ is the observability matrix of $(C, A)$ and

$$
\tilde{N} \triangleq \begin{bmatrix}
\bar{C} \\
\bar{C}A \\
\vdots \\
\bar{C}A^{n-1}
\end{bmatrix}
$$

These equations respectively imply that

$$
N^{-1}\tilde{N} = RR^{-1} \quad \bar{C} = CRR^{-1} \quad N^{-1}\tilde{N}\bar{A} = ARR^{-1} \quad B = N^{-1}\tilde{N}B
$$

where $N^{-1}$ is a left inverse of $N$. Since $T \triangleq RR^{-1}$, it follows from the preceding that the lemma is true. \square
Chapter 14

Stability

Intuitively speaking, a system is stable if nothing “blows up.” The aim of stability theory is to define stability in precise mathematical terms to capture this idea. This turns out to be a bit of a challenge as there are many different ways to define stability which are all useful. For example, one could say that a linear system is stable if each bounded input produces a bounded output response. Or one could say a linear system is stable if the state of the system remains bounded, no matter what the initial state, assuming the input is zero. The former is an example of an “external” or input-output view of stability while the latter is an “internal” view of stability. Under certain conditions the two views are equivalent. The aim of this chapter is to define various notions of stability. We begin with the idea of internal stability.

14.1 Uniform Stability

Let \( \Sigma \) denote a linear system with state equation

\[
\dot{x} = A(t)x + B(t)u
\]

where \( A(t) \) and \( B(t) \) are bounded and piecewise continuous matrices defined on the time interval \([0, \infty)\). We say that \( \Sigma \) or \( A(t) \) is uniformly stable if there is a positive constant \( \beta \) such that

\[
||\Phi(t, \tau)|| \leq \beta, \quad \forall t \geq \tau, \quad \forall \tau \geq 0 \tag{14.1}
\]

Here \( \Phi(t, \tau) \) is the state transition matrix of \( A(t) \) and \( || \cdot || \) is any norm or induced norm on \( \mathbb{R}^{n \times n} \) for which the submultiplicative property holds\(^1\). It can be shown quite easily that uniform stability means that any solution \( x(t) \) to \( \dot{x} = A(t)x \) is bounded in the sense that \( ||x(t)|| \leq \beta ||x(t_0)|| \) for all \( t \geq t_0 \) and all \( t_0 \geq 0 \). The modifier “uniform” is used to emphasize the requirement that for uniform stability \( \beta \) must be independent of \( t_0 \).

Uniform stability turns out to be less than one might hope for in at least three ways. First, even though uniform stability ensures boundedness of all solutions to \( \dot{x} = A(t)x \), it does not ensure that all such solutions will tend to zero as \( t \to \infty \). Second, there are uniformly stable systems whose state responses to bounded inputs grow without bound. Third, there are uniformly stable \( A(t) \)’s which, after arbitrarily small perturbations of their elements, are no longer uniformly stable. Each of these situations are exemplified by the scalar state equation \( \dot{x} = u \). This system is uniformly stable because its state-transition matrix, namely \( \Phi(t, \tau) = 1 \) satisfies (14.1) with \( \beta \overset{\Delta}{=} 1 \). Note first that the only solution to \( \dot{x} = 0 \) which tends to zero

\[^1\text{For concreteness, one can think in terms of the infinity norm } ||M_{p \times q}|| \overset{\Delta}{=} \max_{i,j} |m_{ij}|.\]
as \( t \to \infty \), is the zero solution itself! Note second that the bounded input \( u(t) = 1, \ t \geq 0 \) produces the unbounded state response \( x(t) = x(0) + t \). Third note that if \( A \) is perturbed to \( A = \epsilon \) where \( \epsilon \) is any positive number, then \( \Phi(t, \tau) = e^{\epsilon(t-\tau)} \); thus no matter how small \( \epsilon \) is, \( A \) will no longer be uniformly stable because \( ||\Phi(t, \tau)|| \) will grow without bound as \( t \to \infty \).

The preceding suggest that what’s needed is a stronger notion of stability which precludes all of these undesirable properties. We introduce such a notion in the next section.

### 14.2 Exponential Stability

Again let \( \Sigma \) denote a linear system with state equation

\[
\dot{x} = A(t)x + B(t)u
\]

where \( A(t) \) and \( B(t) \) are bounded and piecewise continuous matrices defined on the time interval \([0, \infty)\). Let us agree to call \( \Sigma \) or \( A(t) \) exponential stable if there is are positive constants \( \beta \) and \( \lambda \) such that

\[
||\Phi(t, \tau)|| \leq \beta e^{-\lambda(t-\tau)}, \ \forall t \geq \tau, \ \forall \tau \geq 0
\]  \quad (14.2)

It can be shown quite easily that exponential stability means that any solution \( x(t) \) to \( \dot{x} = A(t)x \) tends to zero exponentially fast in the sense that \( ||x(t)|| \leq \beta||x(t_0)||e^{-\lambda(t-t_0)} \) for all \( t \geq t_0 \) and all \( t_0 \geq 0 \). The modifier “uniform” is sometimes used before the word exponential to emphasize the requirement that \( \beta \) and \( \lambda \) must be independent of \( t_0 \). It should be clear that exponential stability implies uniform stability, but not the converse.

At we have said, an exponentially stable \( A(t) \) is a matrix for which all solutions to \( \dot{x} = A(t)x \) tend to zero exponentially fast. But what if we drop the requirement that the convergence be exponential? More specifically, suppose we say that \( \Sigma \) or \( A(t) \) is uniformly asymptotically stable if for any positive number \( \epsilon \) there exists a time \( T_\epsilon \) such that

\[
||\Phi(t, t_0)|| \leq \epsilon \ \forall t \geq T_\epsilon + t_0, \ \forall t_0 \geq 0
\]

In other words, uniform asymptotically stability means that for any tolerance \( \epsilon \) we might pick and no matter what the value of \( t_0 \), we can be sure that \( ||x(t)|| \) will be at least as small as \( \epsilon ||x(t_0)|| \) provided \( t \geq t_0 + T_\epsilon \). This of course implies that \( x(t) \) is converging to zero at \( t \to \infty \). The modifier uniform in this case means that \( T_\epsilon \) must be independent of \( t_0 \).

Let us note right away that exponential stability of \( A(t) \) implies uniform asymptotic stability of \( A(t) \). What’s perhaps a little surprising is that the converse is also true.

**Theorem 37** A bounded, piecewise-continuous matrix \( A(t) \) defined on \([0, \infty)\) is uniformly asymptotically stable if and only if it is exponentially stable.

The theorem implies that if all solutions to \( \dot{x} = A(t)x \) converge to zero, uniformly in \( t_0 \), then the convergence must be exponential.

**Proof of Theorem 37:** It is enough to prove that uniform asymptotic stability implies exponential stability. For this let \( \epsilon \) be any positive number less than 1, and let \( T \) be any positive time such that

\[
||\Phi(t, \tau)|| \leq \epsilon \ \forall t \geq T + \tau, \ \forall \tau \geq 0
\]

Let

\[
\lambda \triangleq -\frac{1}{T} \log \epsilon
\]
Since the norm we are using is submultiplicative, we have
\[ ||\Phi(t + \tau, \tau)|| \leq e^{-\lambda T} \quad \forall \tau \geq 0 \quad (14.3) \]

Next let
\[ a \triangleq \sup_{t \in [0, \infty)} ||A(t)|| \]
Since \( A(t) \) is bounded, \( a \leq \infty \). By definition,
\[ \dot{\Phi}(t, \tau) = A(t)\Phi(t, \tau), \quad \Phi(\tau, \tau) = I, \quad \forall t \geq \tau, \quad \forall \tau \geq 0 \]
Thus
\[ \Phi(t, \tau) = I + \int_{\tau}^{t} A(s)\Phi(s, \tau)ds, \quad \forall t \geq \tau, \quad \forall \tau \geq 0 \]
Therefore for all \( t \geq \tau \) and \( \tau \geq 0 \)
\[
||\Phi(t, \tau)|| = ||I + \int_{\tau}^{t} A(s)\Phi(s, \tau)ds|| \\
\leq ||I|| + ||\int_{\tau}^{t} A(s)\Phi(s, \tau)ds|| \\
\leq 1 + \int_{\tau}^{t} ||A(s)\Phi(s, \tau)||ds
\]
Hence
\[ ||\Phi(t, \tau)|| \leq 1 + \int_{\tau}^{t} a||\Phi(s, \tau)||ds, \quad \forall t \geq \tau, \quad \forall \tau \geq 0 \]
Thus by the Bellman-Gronwall Lemma,
\[ ||\Phi(t, \tau)|| \leq e^{a(t-\tau)}, \quad \forall t \geq \tau, \quad \forall \tau \geq 0 \]
Clearly
\[ ||\Phi(t, \tau)|| \leq e^{(a+\lambda)T-\lambda(t-\tau)}, \quad \forall T \geq t-\tau, \quad \forall t-\tau \geq 0, \quad \forall \tau \geq 0 \quad (14.4) \]
Now fix \( t_0 > 0 \) and \( t \geq t_0 \). Let \( n \) be a nonnegative integer and \( t_1 \) a number in \([0, T]\) such that \( t - t_0 = nT + t_1 \). Then by the composition rule for state-transition matrices
\[ \Phi(t, t_0) = \Phi(t_1 + nT + t_0, nT + t_0)\Phi(nT + t_0, (n-1)T + t_0) \cdots \Phi(T + t_0, t_0) \]
Since the norm we are using is submultiplicative,
\[ ||\Phi(t, t_0)|| \leq ||\Phi(t_1 + nT + t_0, nT + t_0)|| ||\Phi(nT + t_0, (n-1)T + t_0)|| \cdots ||\Phi(T + t_0, t_0)|| \]
From this, (14.3) and (14.4) there follows
\[ ||\Phi(t, t_0)|| \leq e^{(a+\lambda)T-\lambda(t_1 + nT + t_0 - nT - t_0)}e^{-\lambda(nT + t_0 - (n-1)T - t_0)} \cdots e^{-\lambda(T + t_0 - t_0)} \]
Hence
\[ ||\Phi(t, t_0)|| \leq e^{(a+\lambda)T - \lambda(t_1 + nT - t_0)} = e^{(a+\lambda)(T - \lambda(t - t_0))} \]
Setting \( \beta \triangleq e^{(a+\lambda)T} \) we thus arrive at the inequality
\[ ||\Phi(t, t_0)|| \leq \beta e^{-\lambda(t-t_0)} \]
Since this holds for all \( t \geq t_0 \) and all \( t_0 \geq 0 \), \( A(t) \) is exponentially stable. \( \blacksquare \)
Example 64 The general solution to the differential equation
\[ \dot{x} = -\frac{1}{1+t} x \]
is
\[ x(t) = \frac{1 + t_0}{1 + t} x(t_0) \]
Clearly any such solution tends to zero as \( t \to \infty \) but not exponentially fast. Note that given any positive number \( \epsilon < 1 \), it is not possible to find another number \( T_\epsilon \), not depending on \( t_0 \), for which
\[ ||x(t)|| \leq \epsilon ||x(t_0)||, \quad \forall t \geq t_0 + T_\epsilon, \quad \forall t_0 > 0 \]
Said differently, the system is not uniformly asymptotically stable even though \( x \) tends to zero.

14.3 Time-Invariant Systems

For time-invariant systems, one drops the modifier uniform and refers to uniformly stable systems simply as stable systems. Time-invariant linear systems which are not stable are said to be unstable linear systems. For time invariant case it is possible to completely characterize both stability and exponential stability in terms of the eigenvalues of \( A \) and there multiplicities. Let us begin with a brief review.

14.3.1 The Structure of the Matrix Exponential

Assume that \( A \) is a given constant \( n \times n \) matrix, that \( J \) is its Jordan Normal Form and that \( T \) is a nonsingular matrix such that
\[ A = TJT^{-1} \]
Then as we’ve already explained in Chapter 8 of the math notes,
\[ e^{tA} = Te^{tJ}T^{-1} \]
Now since \( J \) is a block diagonal matrix of the form
\[ J = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & J_m \end{bmatrix} \]
we can write
\[ e^{tJ} = \begin{bmatrix} e^{tJ_1} & 0 & \cdots & 0 \\ 0 & e^{tJ_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{tJ_m} \end{bmatrix} \]
Thus to characterize the kinds of time functions which can appear in \( e^{tA} \), what we really need to do it to characterize the kinds of time functions which appear in matrix exponentials of the form \( e^{tJ_i} \) where \( J_i \) is a typical Jordan Block. Whatever these time functions are, they are all going to appear in \( e^{tA} \) in linear combination with similar time functions from \( J \)'s other Jordan Blocks. This is because for any matrix \( B \), each entry of the matrix \( TB^{-1} \) is a linear combination of the entries of the matrix \( B \).

\(^2\text{Named in honor of this course’s distinguished spokesperson.}\)
Now it is quite clear that if
\[ J_i = [\lambda_i]_{1 \times 1} \]
then \( e^{tJ_i} \) is simply the scalar exponential
\[ e^{tJ_i} = e^{t\lambda_i} \]
Consider next the case when
\[ J_i = \begin{bmatrix} \lambda_i & 1 \\ 0 & \lambda_i \end{bmatrix}_{2 \times 2} \]
Then
\[ e^{tJ_i} = \begin{bmatrix} e^{t\lambda_i} & te^{t\lambda_i} \\ 0 & e^{t\lambda_i} \end{bmatrix} \]
This can easily be verified by checking to see that with \( e^{tJ_i} \) so defined,
\[ \frac{d}{dt} e^{tJ_i} = J_i e^{tJ_i}, \quad \text{and} \quad e^{tJ_i \mid_{t=0}} = I \]
Similarly, for
\[ J_i = \begin{bmatrix} \lambda_i & 1 & 0 \\ 0 & \lambda_i & 1 \\ 0 & 0 & \lambda_i \end{bmatrix} \]
one finds that
\[ e^{tJ_i} = \begin{bmatrix} e^{t\lambda_i} & te^{t\lambda_i} & \frac{t^2}{2} e^{t\lambda_i} \\ 0 & e^{t\lambda_i} & te^{t\lambda_i} \\ 0 & 0 & e^{t\lambda_i} \end{bmatrix} \]
More generally, if
\[ J_i = \begin{bmatrix} \lambda_i & 1 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_i & 1 & \cdots & 0 & 0 \\ 0 & 0 & \lambda_i & \ddots & \ddots & \ddots \\ \vdots & \vdots & \ddots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \cdots & \lambda_i & 1 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_i \end{bmatrix}_{m_i \times m_i} \]
then
\[ e^{tJ_i} = \begin{bmatrix} e^{t\lambda_i} & te^{t\lambda_i} & \frac{t^2}{2} e^{t\lambda_i} & \cdots & \frac{(m_i-2)!}{(m_i-2)!} e^{t\lambda_i} & \frac{(m_i-1)!}{(m_i-1)!} e^{t\lambda_i} \\ 0 & e^{t\lambda_i} & te^{t\lambda_i} & \cdots & \frac{(m_i-3)!}{(m_i-3)!} e^{t\lambda_i} & \frac{(m_i-2)!}{(m_i-2)!} e^{t\lambda_i} \\ 0 & 0 & e^{t\lambda_i} & \cdots & \frac{(m_i-4)!}{(m_i-4)!} e^{t\lambda_i} & \frac{(m_i-3)!}{(m_i-3)!} e^{t\lambda_i} \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & \cdots & e^{t\lambda_i} & te^{t\lambda_i} \\ 0 & 0 & 0 & \cdots & 0 & e^{t\lambda_i} \end{bmatrix}_{m_i \times m_i} \]
Thus we see that the Jordan Block of the elementary divisor \((s-\lambda_i)^{m_i}\) contains within it, linear combinations \(\{\text{actually just scalar multiples}\}\) of the time functions
\[ e^{t\lambda_i}, te^{t\lambda_i}, \ldots, t^{(m_i-1)} e^{t\lambda_i} \]
We’ve arrived at the following characterization of the state transition matrix of \(A\).
**Theorem 38** Let $A$ be an $n \times n$ matrix with elementary divisors

$$(s - \lambda_1)^{m_1}, (s - \lambda_2)^{m_2}, \ldots, (s - \lambda_q)^{m_q}$$

Then each of the entries of $e^{tA}$ is a linear combination of the time functions

$$e^{t\lambda_i}, te^{t\lambda_i}, \ldots, t^{(m_i-1)}e^{t\lambda_i}, \ i \in \{1, 2, \ldots, q\}$$

Thus the qualitative behavior of $e^{tA}$ and consequently all solutions to $\dot{x} = Ax$, is completely characterized by $A$’s eigenvalues and the sizes of their Jordan blocks.

**Real Matrices**

Suppose $A$ is real, that $\lambda$ is one of its eigenvalues and that $m$ is the size of its Jordan block. Then as we’ve just noted, linear combinations of the time functions

$$e^{\lambda t}, te^{\lambda t}, \ldots, t^{m-1}e^{\lambda t}$$

must appear in $e^{tA}$. Suppose $\lambda$ is not real; i.e.,

$$\lambda = a + j\omega$$

where $a$ and $\omega$ are real numbers and $\omega \neq 0$. Then

$$\lambda^* = a - j\omega$$

must also be an eigenvalue of $A$ with the same size Jordan block at $\lambda$. Because $A$ is real, it can easily be shown that the linear combinations of the time functions

$$e^{\lambda t}, te^{\lambda t}, \ldots, t^{m-1}e^{\lambda t}, e^{\lambda^* t}, te^{\lambda^* t}, \ldots, t^{m-1}e^{\lambda^* t}$$

which appear in $e^{tA}$ can also be written as linear combinations of time functions of the forms

$$e^{at}\cos \omega t, e^{at}\sin \omega t, te^{at}\cos \omega t, te^{at}\sin \omega t, \ldots, t^{m-1}e^{at}\cos \omega t, t^{m-1}e^{at}\sin \omega t$$

**14.3.2 Asymptotic Behavior**

Let $A$ be either real or complex valued. On the basis of Theorem 38 we can draw the following conclusions.

1. If all of $A$’s eigenvalues have negative real parts, then $\|e^{tA}\| \to 0$ exponentially fast, as $t \to \infty$. Such $A$ are said to be asymptotic stability matrices.

2. If at least one eigenvalue of $A$

   (a) has a positive real part $a$, then

   $$\|e^{tA}\|$$

   must grow without bound, as fast as $e^{at}$, as $t \to \infty$.

---

$^3$Note that for any finite integer $i \geq 0$, and any positive number $\mu$, $t^i e^{-\mu t} \to 0$ as $t \to \infty$. 
(b) has a zero real part and a Jordan block of size \( m > 1 \), then
\[
||e^{tA}||
\]
must grow without bound as fast \( t^{m-1} \), as \( t \to \infty \).

In either case, such \( A \) are said to be \emph{unstable matrices}.

3. If all of \( A \)'s eigenvalues have nonpositive real parts, and those with zero real parts all have Jordan blocks of size 1, then
\[
||e^{tA}||
\]
will remain finite as \( t \to \infty \) but will not approach 0. Such \( A \) are said to be \emph{stable matrices}.

The key observation here is summarized as follows:

**Theorem 39** A \( n \times n \) matrix \( A \) is asymptotically stable; i.e.,
\[
\lim_{t \to \infty} e^{tA} = 0
\]
if and only if all of \( A \)'s eigenvalues have negative real parts.

### 14.3.3 Discrete Systems

Just about everything we've discussed in the preceding section for the linear differential equation
\[
\dot{x} = Ax
\]
extends an a natural way to a linear recursion equation of the form
\[
x(t+1) = Ax(t), \quad t = 0, 1, 2, \ldots
\]
For example, for the recursion equation we can write
\[
x(t) = A^{(t-\tau)}x(\tau), \quad \forall t, \tau \geq 0
\]
For this reason \( A^{(t-\tau)} \) is called the \{discrete-time\} \emph{state transition matrix} of \( A \). Note that if \( J \) is \( A \)'s Jordan normal form and \( A = TJT^{-1} \), then
\[
A^t = TJ^tT^{-1},
\]
so the asymptotic behavior of \( A^t \) is characterized by \( A \)'s eigenvalues, just as in the continuous time case discussed above. We leave it to the reader to verify the following:

1. If all of \( A \)'s eigenvalues have magnitude less than 1, then \( ||A^t|| \to 0 \) exponentially fast, as \( t \to \infty \).
   Such \( A \) are said to be \emph{discrete-time asymptotic stability matrices}.

2. If at least one eigenvalue of \( A \)
   (a) has magnitude \( a \) greater than 1, then
   \[
   ||A^t||
   \]
   must grow without bound, as fast as \( |a|^t \), as \( t \to \infty \).

\footnote{Note that for any finite integer \( i \geq 0 \), and any real number \( \mu \) with magnitude less than 1, \( t^i \mu^t \to 0 \) as \( t \to \infty \).}
(b) has magnitude 1 and multiplicity \( m > 1 \), then

\[
||A^t||
\]

must grow without bound as fast \( t^{m-1} \), as \( t \to \infty \).

In either case, such \( A \) are said to be \textit{discrete-time unstable matrices}.

3. If all of \( A \)'s eigenvalues have magnitudes no greater than 1, and those with magnitude 1 are all have Jordan blocks of size 1, then

\[
||A^t||
\]

will remain finite as \( t \to \infty \) but will not approach 0. Such \( A \) are said to be \textit{discrete-time stable matrices}.

The discrete-time version of Theorem 39 is as follows:

\begin{theorem}
An \( n \times n \) matrix \( A \) is discrete-time asymptotically stable; i.e.,

\[
\lim_{t \to \infty} A^t = 0
\]

if and only if all of \( A \)'s eigenvalues have magnitudes less than 1.
\end{theorem}

### 14.4 Lyapunov Stability

The preceding discussion of stability and asymptotic stability for time-varying systems leaves unanswered the question of how one might go about determining whether or not a given \( A(t) \) is uniformly stable or asymptotically stable. In fact, for general time-varying linear systems, there is no constructive means for doing this. Nonetheless, there are various ways to characterize uniform and asymptotic stability which prove useful in certain applications. One of the most useful ideas developed so far is do to the late 19th century engineer A. M. Lyapunov. Lyapunov’s idea applies not only to linear differential equations, but also to more general nonlinear differential equations of the form

\[
\dot{x} = f(x, t)
\]

(14.5)

where \( f \) satisfies \( f(0, t) = 0 \) and is at least locally Lipschitz in \( x \) and piecewise-continuous and bounded \( t \) for every fixed \( x \). Note that for the \( A(t) \)'s we’ve been considering, \( f(x, t) \equiv A(t)x \) satisfies all of these requirements.

Lyapunov’s idea is roughly as follows. As a first step one tries to find a continuously differentiable\(^5\), non-negative, scalar-valued, function \( V(x, t) \) which is bounded in \( t \) on \([0, \infty)\) for each fixed \( x \) and which is radially unbounded\(^6\) in \( x \) for each fixed \( t \). One further requires \( V \) to have the property that that along each solution \( \phi \) to (14.5), the total derivative the time function \( V_t \equiv V(\phi(t), t) \) is non-positive. For if this is so, then \( V_t \) is monotone non-increasing, and therefore never larger than its initial value \( V(x_0, t_0) \). Because of this, and \( V \)'s radial unboundedness, it would then follow that \( \phi(t) \) is bounded where ever it exists and consequently that \( \phi(t) \) exists and is bounded on \([t_0, \infty)\).

To recap, if one can find a continuously differentiable, non-negative, scalar valued, function \( V(x, t) \), which is radially unbounded in \( x \) for each fixed \( t \) and bounded in \( t \) for each fixed \( x \), and whose time derivative

\(^5\)A function \( U(x, t) \) is \textit{continuously differentiable} if its first derivatives with respect to \( x \) and \( t \) are all continuous in \( x \) and \( t \).

\(^6\)A nonnegative, scalar-valued function \( U(x) \) is \textit{radially unbounded} if for each positive number \( r \) there is a positive number \( \hat{r} \) such that \( U(x) > r \) whenever \( ||x|| > \hat{r} \). The definition implies that \( ||x|| \) is finite if \( U(x) \) is.
along any solution to (14.5) is non-positive, then each maximal solution \( \phi \) to (14.5) must be bounded, and therefore defined and bounded on \([t_0, \infty)\). V’s with these properties are typically called Lyapunov functions\(^7\).

What makes Lyapunov’s idea so appealing, is that it is often possible to show that \( \dot{V} \leq 0 \), without having to explicitly compute a solution \( \phi \) to (14.5). In particular, suppose \( V \) satisfies the condition

\[
\frac{\partial V(x,t)}{\partial t} + \frac{\partial V(x,t)}{\partial x} f(x,t) \leq 0, \quad \forall x \in \mathbb{R}^n, \forall t \geq 0
\]

Then along any solution \( \phi \)

\[
\dot{V} = \left( \frac{\partial V(x,t)}{\partial t} + \frac{\partial V(x,t)}{\partial x} \phi \right) \bigg|_{x=\phi(t)} = \left( \frac{\partial V(x,t)}{\partial t} + \frac{\partial V(x,t)}{\partial x} f(x,t) \right) \bigg|_{x=\phi(t)} \leq 0
\]

In other words, to show that \( \dot{V} \leq 0 \), its enough to check that (14.6) is true - and this can be done without computing \( \phi \).

To exploit Lyapunov’s idea, it is thus necessary to construct a “candidate Lyapunov function” \( V \) in such a way that that (14.6) holds. Unfortunately there is in general no systematic method for doing this. Indeed, except for special cases such as \( f(x,t) = Ax \) with \( A \) constant, the construction of Lyapunov functions is very much of an art.

### 14.4.1 Linear Time-Varying Differential Equations

For linear time-varying differential equations of the form

\[
\dot{x} = A(t)x
\]

with \( A(t) \) piecewise-continuous and bounded on \([0, \infty)\), one typically considers candidate Lyapunov functions of the form

\[
V(x,t) = x'P(t)x
\]

where \( P(t) \) is a symmetric, positive definite, continuously differentiable, bounded matrix\(^8\) on \([0, \infty)\). To avoid difficulties in the limit at \( t \to \infty \) one typically demands that \( P(t) \) remain “positive definite in the limit.” This is accomplished by stipulating that there should be a positive number \( \mu \) such that

\[
P(t) \geq \mu I, \quad \forall t \geq 0
\]

What this requirement means is that \( P(t) \)’s smallest eigenvalue, \( \lambda_{\text{min}}(t) \), can never get be smaller than \( \mu \) for any value of \( t \).

Suppose that such a \( P(t) \) can be found and that \( P(t) \) miraculously satisfies a matrix differential equation of the form

\[
\dot{P}(t) + A'(t)P + P(t)A(t) + Q(t) = 0, \quad t \in [0, \infty)
\]

where \( Q(t) \) is a symmetric, bounded, piecewise-continuous, positive semidefinite matrix on \([0, \infty)\). We claim that under these conditions, \( A(t) \) is uniformly stable. To understand why this is so, let us note that along a solution \( x(t) \) to (14.7), \( V(x,t) \) must satisfy

\[
\dot{V}(t) \triangleq \frac{d}{dt} V(x(t), t) = x'(t)\dot{P}(t)x(t) + x'(t)A'(t)P(t)x(t) + x'(t)P(t)A(t)x(t) = x'(t)(\dot{P}(t) + A'(t)P(t) + P(t)A(t))x(t)
\]

\(^7\)The standard definition of a Lyapunov function also requires \( V \) to satisfy \( V(0,t) = 0 \), \( \forall t \geq 0 \).

\(^8\)This would be a good time to review the sections in the Linear Algebra Notes which deal with norms, simultaneous diagonalization, and quadratic forms.
Hence from (14.10)\[ \dot{V} = -x'(t)Q(t)x(t) \] (14.11)

But $Q(t)$ is positive semidefinite which means that $\dot{V} \leq 0$, so $V(x(t), t)$ is a nonincreasing function of $t$. Thus $V$ is bounded above by its initial value $V(x(t_0), t_0)$. That is

$$V(x(t), t) \leq V(x(t_0), t_0), \ \forall t \geq t_0$$

In view of (14.9),

$$||x(t)||^2 \leq \frac{1}{\mu} V(x(t_0), t_0), \ \forall t \geq t_0$$

where $|| \cdot ||$ is the Euclidean norm $||z|| = \sqrt{z^T z}$. To establish uniform stability we need finally to use the assumption that $P(t)$ is bounded on $[0, \infty)$ Because $P$'s eigenvalues depend continuously on its entries and because $P$'s entries are all bounded on $[0, \infty)$, we can conclude that $P$'s largest eigenvalue $\lambda_{\text{max}}(t)$ is also bounded on $[0, \infty)$. Let $\nu$ denote this bound. In other words $\lambda_{\text{max}}(t) \leq \nu$, $\forall t \in [0, \infty)$. This implies that

$$P(t) \leq \nu I, \ \forall t \in [0, \infty)$$

(14.12)

and thus that $V(x(t_0), t_0) \leq \nu ||x(t_0)||^2$, $\forall t \geq t_0$. Therefore $||x(t)||^2 \leq \frac{\nu}{\mu} ||x(t_0)||^2$ or

$$||x(t)|| \leq \sqrt{\frac{\nu}{\mu}} ||x(t_0)||$$

Since this holds for all $t_0$ and neither $\mu$ nor $\nu$ depend on $t_0$, $A(t)$ must be uniformly stable as claimed.

**Theorem 41 (Uniform Stability)** Let $A(t)$ be bounded and piecewise continuous on $[0, \infty)$. Suppose there is a positive number $\mu$ and bounded, symmetric matrices $P(t)$ and $Q(t)$ defined on $[0, \infty)$ with $P(t)$ continuously differentiable and positive definite and $Q(t)$ piecewise-continuous and positive semidefinite, such that $P(t) \geq \mu I$, $\forall t \geq 0$ and

$$\dot{P}(t) + A'(t)P + P(t)A(t) + Q(t) = 0, \ t \in [0, \infty)$$

(14.13)

Then $A(t)$ is uniformly stable.

By making various stronger assumptions about $Q(t)$, it is possible to say even more. We discuss two different assumptions of this type.

$Q(t)$ is Positive Definite on $[0, \infty)$

Suppose we assume in addition to what’s required in Theorem 41, that there is a positive number $\rho$ such that

$$Q(t) \geq \rho I, \ \forall t \geq 0$$

(14.14)

Then using this followed by (14.9) we can write

$$x'(t)Q(t)x(t) \geq \rho ||x(t)||^2 \geq \frac{\rho}{\nu} V(x(t), t), \ \forall t \geq 0$$

From this and (14.11) there follows

$$\frac{d}{dt} V(x(t), t) \leq -\frac{\rho}{\nu} V(x(t), t), \ \forall t \geq 0$$

Thus

$$\frac{d}{dt} \left\{ e^{\frac{\rho}{\nu} t} V(x(t), t) \right\} = e^{\frac{\rho}{\nu} t} \left\{ \frac{d}{dt} V(x(t), t) + \frac{\rho}{\nu} V(x(t), t) \right\} \leq 0$$
Clearly
\[ e^{\frac{\mu}{\nu}t}V(x(t), t)) \leq e^{\frac{\mu}{\nu}t_0}V(x(t_0), t_0), \quad \forall t \geq t_0 \]
so
\[ V(x(t), t)) \leq e^{-\frac{\mu}{\nu}(t-t_0)}V(x(t_0), t_0), \quad \forall t \geq t_0 \]
Using (14.9) and (14.12) we thus arrive at the inequality
\[ ||x(t)||^2 \leq \frac{\mu}{\nu}e^{-\frac{\mu}{\nu}(t-t_0)}||x(t_0)||^2, \quad \forall t \geq t_0 \]
or
\[ ||x(t)|| \leq \sqrt{\frac{\mu}{\nu}e^{-\frac{\mu}{\nu}(t-t_0)}}||x(t_0)||, \quad \forall t \geq t_0 \]
Since \( \mu, \nu \) and \( \rho \) don’t depend on \( t_0 \), we can conclude that \( A(t) \) is exponentially stable. We summarize.

**Corollary 10 (Exponential Stability)** Let the hypotheses of Theorem 41 hold and suppose that there is a positive number \( \rho \) such that \( Q(t) \geq \rho I, \quad \forall t \geq 0 \). Then \( A(t) \) is exponentially stable.

\[ \dot{Q}(t) \text{ Exists and is Bounded on } [0, \infty) \]

Without assuming anything more about \( Q(t) \) than we did in the statement of Theorem 41 we can deduce more by exploiting the fact from calculus that every bounded monotone nonincreasing function converges in the limit. We have such a function, namely \( V(x(t), t) \). We’ve already established that \( V \) is monotone nonincreasing and bounded above. Moreover, we know that \( V \) is bounded below by zero because \( P(t) \) is a positive definite matrix. Thus we can conclude that the limit
\[ V_\infty \triangleq \lim_{t \to \infty} V(x(t), t) \]
exists and is finite. This means that
\[ \lim_{t \to \infty} \int_{t_0}^t \dot{V}(x(s), s)ds = V_\infty - V(x(t_0), t_0) \]
Hence from (14.11) we see that
\[ \lim_{t \to \infty} \int_{t_0}^t x'(s)Q(s)x(s)ds = V(x(t_0), t_0) - V_\infty < \infty \quad (14.15) \]
We claim that if \( \dot{Q}(t) \) exists and is bounded on \([t_0, \infty)\) then (14.15) implies that
\[ \lim_{t \to \infty} x'(t)Q(t)x(t) = 0 \quad (14.16) \]
This equation can be used to deduce, to some extent, what happens to \( x(t) \) in the limit as \( t \to \infty \). For example, if \( Q(t) = Q \) is a constant matrix, then the preceding implies that \( x \to \) kernel \( Q \). On the other hand if \( x'(t)x \) is the norm square of a signal \( y \) of the form \( y = C(t)x \) \((i.e., \ Q(t) = C'(t)C(t))\), then this would mean that for every initial state, the output of the unforced linear system
\[ y = C(t)x \]
\[ \dot{x} = A(t)x \]
tends to zero as \( t \to \infty \).

Justification for (14.16) rests on the following lemma which is useful in its own right.
Lemma 12 (Convergence) Let $\mu : [0, \infty) \to \mathbb{R}$ be a continuous function with a bounded, piecewise-continuous derivative $\dot{\mu}$. If the limit
\[
\lim_{t \to \infty} \int_{0}^{t} \mu(\tau) \, d\tau
\]
exists and is bounded, then
\[
\lim_{t \to \infty} \mu(t) = 0
\]

Proof of Convergence Lemma: It is enough to show that for any positive number $\epsilon$, the set $S_\epsilon \triangleq \{ t : |\mu(t)| \geq \epsilon \}$ is either empty or bounded from above. For if this is so, then for any number $\epsilon > 0$, $|\mu(t)|$ must be smaller than $\epsilon$ for any value of $t$ greater than $S_\epsilon$’s upper bound; and this in turn must imply that (14.18) is true.

To proceed, pick $\epsilon > 0$ and suppose that $S_\epsilon$ is not bounded from above. Let $b$ be any positive number bounding $\dot{\mu}$; that is $|\dot{\mu}(t)| < b$, $\forall t \geq 0$. The Cauchy convergence test together with the hypothesis that the integral in (14.17) converges to a finite limit imply that a time $t_0 > 0$ sufficiently large can be found for which
\[
\left| \int_{t_0}^{t} \mu(s) \, ds \right| < \frac{\epsilon^2}{2b}, \quad \forall t > t_0
\]
Moreover, since $S_\epsilon$ is supposed to be unbounded, $t_0$ can be chosen large enough to be in $S_\epsilon$ as well. Thus $|\mu(t_0)| \geq \epsilon$, so for $t \geq t_0$
\[
\text{sign}(\mu(t_0)) \mu(t) = \text{sign}(\mu(t_0)) \left( \mu(t_0) + \int_{t_0}^{t} \dot{\mu}(s) \, ds \right)
\geq |\mu(t_0)| - \int_{t_0}^{t} |\dot{\mu}(s)| \, ds
\geq \epsilon - \int_{t_0}^{t} b \, ds
= \epsilon - b(t - t_0)
\]
Therefore
\[
\int_{t_0}^{t_0+\frac{\epsilon}{b}} \text{sign}(\mu(t_0)) \mu(s) \, ds \geq \int_{t_0}^{t_0+\frac{\epsilon}{b}} (\epsilon - b(s - t_0)) \, ds = \frac{\epsilon^2}{2b}
\]
Clearly
\[
\left| \int_{t_0}^{t_0+\frac{\epsilon}{b}} \mu(s) \, ds \right| \geq \frac{\epsilon^2}{2b}
\]
which contradicts (14.19).

14.4.2 Time-Invariant Linear Differential Equations

A major problem with the results derived in the last section for time-varying linear systems is that they don’t really provide constructive tests for deciding when a system is uniformly stable or exponentially stable. This is because one has no effective means available for actually computing matrices $P(t)$ and $Q(t)$ with the required properties. Fortunately for time-invariant systems the situation is quite different. In the sequel we will prove the following theorem which provides a constructive test for deciding whether or not a constant matrix $A$ is asymptotically stable.
Theorem 42 (Lyapunov Stability) Let \( A \) and \( Q \) be any real, constant \( n \times n \) matrices.

- If \( A \) is an asymptotic stability matrix, then the linear matrix equation
  \[
  A' P + P A + Q = 0
  \]  
  (14.20)
  has exactly one solution \( P \). Moreover, if \( Q \) is \{symmetric\} positive definite then so is \( P \).

- If \( P \) and \( Q \) are positive definite matrices satisfying (14.20), then \( A \) is an asymptotic stability matrix.

The theorem provides an effective test for deciding whether or not a given matrix \( A \) is an asymptotic stability matrix, without having to compute the matrix’s eigenvalues. The test consists of first picking any convenient positive definite matrix for \( Q \) (say \( Q = I \)). Next one tries to solve (14.20), which is a linear algebraic equation. \( A \) is an asymptotic stability matrix just in case (i) a solution \( P \) exists and (ii) \( P \) is positive definite. There are a variety of ways to test a symmetric matrix for positive definiteness. See the Linear Algebra Notes for this.

Proof of Theorem 42: Let us note right away that if \( Q \) and \( P \) are positive definite matrices satisfying (14.20), then these matrices automatically satisfy the hypotheses of Corollary 10. From this it follows that the statement following the second bullet of Theorem 42 is true.

To prove the statements following the first bullet, let \( A \) and \( Q \) be arbitrary real \( n \times n \) matrices and define

\[
X(t) \overset{\Delta}{=} \int_0^t e^{\mu A} Q e^{\mu A} d\mu
\]  
(14.21)

By changing the variable of integration from \( \mu \) to \( \tau \overset{\Delta}{=} t - \mu \) and then differentiating with respect to \( t \) we obtain the linear matrix differential equation

\[
\dot{X} = A' X + X A + Q
\]  
(14.22)

Thus the matrix on the right side of (14.21) is the unique solution to (14.21) which starts in state \( X(0) = 0 \) at \( t = 0 \).

Now suppose that \( A \) is asymptotically stable. Then \( e^{\mu A} Q e^{\mu A} \) is a matrix whose entries are all linear combinations of decaying exponentials so the integral on the right side of (14.21) must converge to a finite limit \( X_\infty \) as \( t \to \infty \). This is the same thing as saying that the solution to (14.22) must converge to \( X_\infty \). In the limit, \( \dot{X} \) vanishes and we have

\[
0 = A' X_\infty + X_\infty A + Q
\]

This means that for each real \( n \times n \) matrix \( Q \), there exists a solution to the equation

\[
L(Y) = -Q
\]  
(14.23)

where \( L \) is the linear transformation \( L : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n} \) defined by

\[
L(Y) \overset{\Delta}{=} A' Y + Y A
\]

Evidently \( L \) is an epimorphism \{i.e., a surjective linear transformation\}. In other words, image \( L = \mathbb{R}^{n \times n} \). Moreover \( L \)'s domain is \( \mathbb{R}^{n \times n} \) so

\[
\dim \text{image } L + \dim \ker L = \dim \mathbb{R}^{n \times n}
\]

just like any linear function. It follows that \( \dim \ker L = 0 \) and thus that \( \ker L = 0 \). This means that there is only one \( Y \) which solves (14.23). In summary, if \( A \) is asymptotically stable, then for each matrix \( Q \), the equation \( 0 = A' Y + Y A + Q \) has exactly one solution \( Y \) and \( Y = X_\infty \).
Now suppose $Q$ is positive definite. To complete the theorem’s proof, it is enough to show that $X_{\infty}$ (i.e. $P$) is positive definite as well. But this is clear since for $x \neq 0$,
\[
 x'X_{\infty}x = x' \left( \int_{0}^{\infty} e^{\mu A} Q e^{\mu A} d\mu \right) x = \int_{0}^{\infty} \| \sqrt{Q} e^{\mu A} x \|^2 d\mu > 0
\]
Therefore both statements following the first bullet of Theorem 42 are true.

14.5 Perturbed Systems

In this section we will develop a result which enables us to draw conclusions about the asymptotic stability of matrices of the form $A(t) + B(t)$ where $A(t)$ is an asymptotic stability matrix and $B(t)$ is a matrix whose norm is “small” in a suitably defined sense. The result plays a central role in the analysis of slowly varying time-varying linear systems and adaptive control systems. We begin with the idea of a nondestabilizing signal.

14.5.1 Nondestabilizing Signals

Let us agree to say that a bounded, piecewise-continuous matrix $B : [0, \infty) \rightarrow \mathbb{R}^{n \times m}$ is nondestabilizing with growth rate $\rho_B \geq 0$ if (i) for each real number $\rho > \rho_B$ there is a finite nonnegative constant $b_\rho$ such that
\[
 \int_{\tau}^{t} \| B(s) \| ds \leq b_\rho + \rho (t - \tau), \quad t \geq \tau \geq 0
\]
and (ii) $\rho_B$ is the least nonnegative number with this property. Note that any piecewise-continuous bounded matrix $B : [0, \infty)$ is nondestabilizing with growth rate $\rho_B \leq \sup_{t \in [0, \infty)} \| B(t) \|$

Examples of nondestabilizing signals with zero growth rates include all bounded, piecewise-continuous function $B : [0, \infty) \rightarrow \mathbb{R}^{n \times m}$ which tend to zero or which satisfy
\[
 \int_{0}^{\infty} \| B(\mu) \|^i d\mu < \infty
\]
for either $i = 1$ or $i = 2$. We will encounter other examples in the sequel.

The term “nondestabilizing” is prompted by the following basic result.

**Theorem 43 (Nondestabilization)** Let $A : [0, \infty) \rightarrow \mathbb{R}^{n \times n}$ and $B : [0, \infty) \rightarrow \mathbb{R}^{n \times n}$ be bounded, piecewise-continuous matrices. Suppose that $A$ is exponentially stable and let $a$ and $\lambda$ be positive constants for which $A$’s state transition matrix $\phi_A(t, \tau)$ satisfies
\[
 \| \phi_A(t, \tau) \| \leq ae^{-\lambda (t-\tau)}, \quad \forall t \geq \tau \geq 0
\]
Suppose $B$ is nondestabilizing with growth rate $\rho_B$ and for each number $\rho > \rho_B$, let $b_\rho$ be a constant for which
\[
 \int_{\tau}^{t} \| B(s) \| ds \leq b_\rho + \rho (t - \tau), \quad t \geq \tau \geq 0
\]
(14.25)
Then for $\rho > \rho_B$, the state transition matrix of $A + B$ satisfies
\[
 \| \phi_{A+B}(t, \tau) \| \leq ae^{(ab_\rho - (\lambda - a\rho)(t-\tau))}, \quad t \geq \tau \geq 0
\]
As an immediate consequence of this theorem we see that if $B$’s growth rate is sufficiently small, namely
\[\rho_B < \frac{\lambda}{a}\]
then $A + B$ will be exponentially stable. This of course will always be true if $B$’s growth rate is zero.

**Proof of Theorem 43:** Fix $\tau$ and define $\theta(t) \overset{\Delta}{=} \Phi_{A+B}(t, \tau)$. Then
\[
\dot{\theta} = A\theta + B\theta \\
\theta(\tau) = I
\]
By viewing $B\theta$ as a forcing function in the preceding, one may write the variation of constants formula
\[
\theta(t) = \Phi_A(t, \tau) + \int_\tau^t \Phi_A(t, s)B(s)\theta(s)ds
\]
Therefore
\[
||\theta(t)|| = ||\Phi_A(t, \tau) + \int_\tau^t \Phi_A(t, s)B(s)\theta(s)ds|| \\
\leq ||\Phi_A(t, \tau)|| + \int_\tau^t ||\Phi_A(t, s)B(s)\theta(s)||ds \\
\leq ae^{-\lambda(t-\tau)} + \int_\tau^t ae^{-\lambda(t-\tau)}||B(s)||||\theta(s)||ds
\]
Hence
\[
e^{\lambda(t-\tau)}||\theta(t)|| \leq a + \int_\tau^t a||B(s)||\{e^{\lambda(s-\tau)}||\theta(s)||\}ds
\]
By the Bellman-Gronwall Lemma,
\[
e^{\lambda(t-\tau)}||\theta(t)|| \leq ae^{\int_\tau^t a||B(s)||ds}
\]
Therefore
\[
||\theta(t)|| \leq ae^{(-\lambda(t-\tau)+\int_\tau^t a||B(s)||ds)}
\]
Using (14.25) there follows
\[
||\theta(t)|| \leq ae^{(abu-(\lambda-a\mu)(t-\tau))}
\]
which is the desired result. ■
Consider the $n$-dimensional linear system

\begin{align*}
\dot{x} &= Ax + Bu & (15.1) \\
y &=Cx + Du & (15.2) \\
s &= \bar{C}x & (15.3)
\end{align*}

which is now regarded as an internal model $\mathcal{P}$ of a physical process to be controlled. It is assumed that $A, B, C, \bar{C}$ and $D$ are known approximately and that they can’t be altered. $\mathcal{P}$ is often called the process model, $u \in \mathbb{R}^m$ the open-loop control input, $y \in \mathbb{R}^p$ the controlled output, and $s \in \mathbb{R}^q$ the sensor output\(^1\). It is not unusual for $y$ and $s$ to be one and the same, or more generally for $y$ to be readable for $s$ in the sense that for some matrix $R$, $y = Rs$.

Consider next the ‘closed-loop’ system exhibited diagramatically as follows.

![Figure 15.1: Feedback Control System](image)

Here $r \in \mathbb{R}^n_r$ is a reference input vector which, depending on the particular situation, may or may not be present. The function of $\mathcal{C}$ is to use sensor output $s$ and $r$ to generate a control signal $u_\mathcal{C}$ which, when $u$ is set equal to $u_\mathcal{C}$, causes $y$ to behave in some prescribed manner.

In these notes we consider controllers which can be modelled by linear, time-invariant dynamical systems. Thus a linear controller $\mathcal{C}$ admits the description

\begin{align*}
\dot{x}_\mathcal{C} &= A_\mathcal{C}x_\mathcal{C} + B_\mathcal{C}s + E_\mathcal{C}r & (15.4) \\
u_\mathcal{C} &= C_\mathcal{C}x_\mathcal{C} + D_\mathcal{C}s + H_\mathcal{C}r & (15.5)
\end{align*}

\(^1\)All of the properties of feedback systems discussed in the sequel extend to processes with sensor outputs of the form $s = \bar{C}x + D_\mathcal{u}$. The less general output relation is adopted throughout for simplicity.
where \( A, B, C, D, E, H \) are constant matrices, \( x \in \mathbb{R}^n \) is the state of \( \mathbb{C} \). \textit{Closed-loop control} is achieved by setting

\[
u = \nu_C \tag{15.6}
\]

so \( \nu_C \in \mathbb{R}^m \).

Let us recall that time function differentiation is invariably avoided when implementing a controller because of its adverse effect on noise. The preceding controller involves no differentiation and may be constructed from standard analog or digital components. Relative to this, the preceding describes the most general linear system for controlling a dynamical process.

### 15.1 State-Feedback

By a \textit{state-feedback control} is meant a feedback law of the form

\[
u = Fx + Gr \tag{15.7}
\]

where \( F \) and \( G \) are constant matrices. If this law is applied to \( \mathbb{P} \) the resulting closed-loop system \( \mathbb{P}_{F,G} \) is described by the equations

\[
\begin{align*}
\dot{x} &= (A + BF)x + BGr \tag{15.8} \\
y &= (C + DF)x + DGr \tag{15.9} \\
z &= Cx \tag{15.10}
\end{align*}
\]

Algebraically the effect of state feedback is thus to transform a linear system \( \mathbb{P} \) into another \( \mathbb{P}_{F,G} \) of the same dimension and form. It is mainly because of the simplicity of this transformation that the study of feedback control in state space terms has proved so useful. Although \( x \) usually can’t be measured directly, thereby precluding implementation of a state-feedback law as a controller \( C \) such as described above, in due course we shall justify our consideration of control laws of this type.

Observe that if \( G \) is nonsingular, then \( \text{Im} \ B = B \) and \( \langle A + BF \mid B \rangle \) is the controllable space of \( \mathbb{P}_{F,G} \). This being the case, the controllable space of \( \mathbb{P}_{F,G} \) turns out to be the same as the controllable space of \( \mathbb{P} \).

**Proposition 12** \textit{For all} \( F \),

\[
\langle A + BF \mid B \rangle = \langle A \mid B \rangle
\]

**Proof:** The equation

\[
\sum_{j=1}^{i} (A + BF)^{j-1}B = \sum_{j=1}^{i} A^{j-1}B \tag{15.11}
\]

is valid for \( i = 1 \). Suppose (15.11) holds for \( i = k \); then

\[
\sum_{j=1}^{k+1} (A + BF)^{j-1}B = (A + BF) \sum_{j=1}^{k} (A + BF)^{j-1}B + B = (A + BF) \sum_{j=1}^{k} A^{j-1}B + B = A \sum_{j=1}^{k} A^{j-1}B + B
\]

(by hypothesis)
By induction, (15.11) holds for all \( i \geq 0 \) and in particular for \( i = n \). ■

An immediate consequence of Proposition 12 is that \((A + BF, B)\) is controllable if and only if \((A, B)\) is. A corresponding statement concerning the effect of state feedback on observability cannot be made; i.e., even if \((C, A)\) is observable, \((C + DF, A + BF)\) may not be.

### 15.2 Spectrum Assignment

Consider the system \( P_F, I \) to which the state feedback law \( u = Fx + r \) has been applied. As already noted, the spectrum of \( A + BF \) has a significant effect on the transient behavior of the system; e.g., the system is asymptotically stable if and only if all eigenvalues of \( A + BF \) lie within the open left-half of the complex plane. It is therefore important to understand the extent to which the spectrum of \( A + BF \) can be adjusted with \( F \).

Recall that the spectrum of \( A + BF \) is a symmetric set of complex numbers since the characteristic polynomial of \( A + BF \) has real coefficients. In the sequel it will be shown that the symmetric spectrum of \( A + BF \) can be freely assigned with suitable \( F \) just in case \((A, B)\) is a controllable pair. This property of controllable pairs depends upon the following preliminary results.

**Proposition 13** Let \((A, b)\) be a single-input, \( n \)-dimensional, real controllable pair. For each symmetric set of \( n \) complex numbers \( \Lambda \) there exists a real matrix \( F \) such that

\[
\text{spectrum } (A + bf) = \Lambda
\]

**Proof:** From the symmetric set \( \Lambda = \{\lambda_i : i \in \mathbb{N}\} \) construct the polynomial

\[
\tilde{\alpha}(s) = \prod_{i=1}^{n} (s - \lambda_i) = s^n + \tilde{a}_ns^{n-1} + \tilde{a}_{n-1}s^{n-2} + \ldots + \tilde{a}_2s + \tilde{a}_1
\]

Let

\[
\alpha(s) = s^n + a_ns^{n-1} + a_{n-1}s^{n-2} + \ldots + a_2s + a_1
\]

denote the characteristic polynomial of \( A \) and write \( T \) for the matrix which transforms \((A, b)\) into control canonical form. That is \( TAT^{-1} = A_C \) and \( Tb = b_C \) where

\[
A_C = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
-a_1 & -a_2 & -a_3 & \cdots & -a_n
\end{bmatrix}_{n \times n}, \quad b_C = \begin{bmatrix} 0 \\
0 \\
\vdots \\
0 \\
1 \end{bmatrix}_{n \times 1}
\]

Note that if we define the row vector

\[
f_C \triangleq [a_1 - \tilde{a}_1 \quad a_2 - \tilde{a}_2 \quad \cdots \quad a_n - \tilde{a}_n]
\]

then

\[
A_C + b_C f_C = \begin{bmatrix}
0 & 1 & 0 & \cdots & 0 \\
0 & 0 & 1 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 1 \\
-a_1 & -a_2 & -a_3 & \cdots & -a_n
\end{bmatrix}
\]
Thus \( \alpha(s) \) is the characteristic polynomial of \( A_C + b_C f_C \) which means that \( \Lambda \) is \( (A_C + b_C f_C) \)'s spectrum. Define
\[
f \triangleq f_C T
\]
and note that \( T(A + bf)T^{-1} = A_C + b_C f_C \). Since \( A + bf \) and \( A_C + b_C f_C \) are similar, both matrices must have the same spectrum. Hence (15.12) is true. ■

Consider next the multi-input controllable pair \( (A, B) \). Spectrum assignment in this case reduces immediately to the single-input case whenever there exists a vector \( b \in \text{image } B \) such that \( (A, b) \) is a controllable pair. For when this is so, Proposition 13 can be used to construct a matrix \( f \) which assigns to \( A + bf \) a prespecified symmetric spectrum. Since \( b \in \text{image } B \), the equation \( b = Bg \) has a solution \( g \); thus with \( F = gf \), there follows \( A + BF = A + bf \) so \( (A + BF) \)'s spectrum is as prescribed.

Unfortunately a vector \( b \in \text{image } B \) with the required property may not exist. For example, if \( A = 0 \) and \( B \) is the \( n \times n \) identity then \( (A, B) \) is a controllable pair but there's no \( b \in \text{image } I = \{0\} \) for which \( (0, b) \) is controllable. Indeed from our prior discussion, we know that such a vector \( b \) cannot exist if \( A \) is not cyclic.

To treat the general multi-input case, a slightly different approach will be taken. It will be shown that if \( (A, B) \) is controllable, then there exists a vector \( b \in \text{image } B \) and a matrix \( F \) such that \( (A + BF, b) \) is controllable. Thanks to M. Heymann, even more is true:

**Lemma 13 (Heymann).** Let \( (A, B) \) be an \( n \)-dimensional, \( m \)-input controllable pair and let \( b \) be any given nonzero vector in \( \text{image } B \). There exists a matrix \( F \) such that \( (A + BF, b) \) is a controllable pair.

**Proof:** Define \( B \triangleq \text{image } B \). Since \( b \neq 0 \) there is a largest integer \( n_1 \), such that \( \{b, Ab, \ldots, A^{n_1-1}b\} \) is an independent set. If \( n_1 = n \), set \( F = 0 \) since in this case \( (A, b) \) is controllable. If \( n_1 < n \), there exists a vector \( b_2 \in B \) such that \( b_2 \notin \text{span } \{b, Ab, \ldots, A^{n_1-1}b\} \). For if this were not so \( B \) would be a subspace of \( \text{span } \{b, Ab, \ldots, A^{n_1-1}b\} \). Since \( A \mathcal{R}_1 \subset \mathcal{R}_1 \), it would then follow that \( <A|B> = \mathcal{R}_1(\neq \mathbb{R}^n) \) which contradicts the controllability hypothesis. Thus a vector \( b_2 \) with the required properties exists. Let \( n_2 \) be the largest integer such that \( \{b, Ab, \ldots, A^{n_1-1}b, b_2, Ab_2, \ldots, A^{n_2-1}b_2\} \) is an independent set. Since \( (A, B) \) is controllable, this process can be continued until for some integer \( k \leq \dim B \),

\[
\{b, Ab, \ldots, A^{n_1-1}b, b_2, Ab_2, \ldots, A^{n_2-1}b_2, \ldots, b_k, \ldots, A^{n_k-1}b_k\}
\]
is a basis for \( \mathbb{R}^n \).

Define
\[
x_1 = b \tag{15.13}
\]
and
\[
x_i = Ax_{i-1} + b_{i-1} \quad i = 2, 3, \ldots, n \tag{15.14}
\]
where
\[
\bar{b}_{(n_1+n_2+\ldots+n_j)+1} = b_{j+1} \quad j = 1, 2, \ldots, k - 1
\]
and all other \( \bar{b}_i = 0 \). It is easy to verify that \( \{x_1, x_2, \ldots, x_n\} \) is a basis for \( \mathbb{R}^n \). Compute vectors \( u_i \) so that
\[
Bu_i = \bar{b}_i \quad i = 1, 2, \ldots, n \tag{15.15}
\]
Finally, define
\[
F \triangleq [u_1 \quad u_2 \ldots \quad u_n] [x_1 \quad x_2 \ldots \quad x_n]^{-1}
\]
This means that
\[
Fx_i = u_i \quad i \in \{1, 2, \ldots, n\} \tag{15.16}
\]
It follows from (15.13)-(15.16) that
\[ x_i = Ax_{i-1} + \bar{b}_{i-1} = Ax_{i-1} + Bu_{i-1} = (A + BF)x_{i-1} \quad i = 2, 3, \ldots, n \]

Using (15.13), \( x_i \) can therefore be written as
\[ x_i = (A + BF)^{i-1}b, \quad i \in \{1, 2, \ldots, \} \]
since the \( x_i \) span \( \mathbb{R}^n \), the pair \( (A + BF, b) \) is controllable. ■

The main theorem on spectrum assignment is as follows.

**Theorem 44** An \( n \)-dimensional matrix pair \((A, B)\) is controllable if and only if for each symmetric set of \( n \) complex numbers \( \Lambda \), there exists a matrix \( F \) such that
\[ \text{spectrum} \ (A + BF) = \Lambda \]

We will only prove that controllability implies spectrum assignability. We leave the proof of the converse implication to the reader as an exercise.

**Proof:** If \((A, B)\) is controllable, there is \{by Lemma 13\} for any fixed nonzero vector \( b \in B \), a matrix \( F_b \) such that \((A + BF_b, b)\) is controllable. Proposition 15.12 thus provides a row vector \( f \) such that spectrum \((A + BF_b + bf) = \Lambda\). Since \( b \in \text{image} \ B \), the equation \( b = Bg \) has a solution \( g \). With \( F \Delta F_b + gf \), there follows \( A + BF = A + BF_b + bf \) so \( F \) has the required property. ■

### 15.3 Observers

In the last section a procedure was developed for adjusting the closed-loop spectrum of a linear system with a state-feedback law of the form
\[ u = Fx + Gv \] (15.17)

In practice direct measurement of system state is often not possible and so controllers such as the preceding cannot typically be implemented. There is however an alternative controller which can be implemented and which achieves similar results. The architecture of the control consists of a linear sub-system called an “observer” whose role is to “asymptotically estimate” \( Fx \) from measurement \( y \); this estimate is then used in place of \( Fx \) in (15.17) to define the actual control signal \( u \) to be feedback to the process. In the sequel we define and characterize observers.

#### 15.3.1 Definition

Consider the unforced \( n \)-dimensional linear system
\[ y = Cx, \quad \dot{x} = Ax \] (15.18)
together with an additional output signal of the form
\[ w = Lx \] (15.19)
where \( L \) is any given matrix with \( n \) columns. Our aim is to define a linear system of the general form

\[
\begin{align*}
\dot{z} &= Hz + Ky \quad (15.20) \\
\hat{w} &= Mz + Ny \quad (15.21)
\end{align*}
\]

in such a way so that along any trajectory of the combined linear system defined by (15.18)- (15.21), \( \hat{w} \) converges to \( w \) in the limit as \( t \to \infty \). We call \( \hat{w} \) an asymptotic estimate of \( w \). Since (15.21) must hold even when the estimate is exact, \( \{\text{i.e., when } \hat{w} = w\} \), we require the algebraic equations

\[
Lx = Mz_x + Ny \quad y = Cx
\]

to have a solution \( z_x \) for each possible \( x \in \mathbb{R}^n \). In other words, for each such \( x \) we require there to be a vector \( z_x \) such that

\[
(L - NC)x = Mz_x
\]

Since for any given \( x \) the preceding is a linear algebraic equation, such a \( z_x \) will exist if and only if

\[
(L - MC)x \in \text{image } M
\]

To ensure a solution no matter what the value of \( x \) we must therefore require that

\[
\text{image } (L - NC) \subset \text{image } M
\]

But this is exactly the condition for there to exist a solution \( V \) to the linear matrix equation

\[
L - NC = MV \quad (15.22)
\]

Suppose that \( V \) is such a matrix and define the error signal

\[
e = z - Vx \quad (15.23)
\]

where \( z \) and \( x \) are now the state vectors of (15.18) and (15.20) respectively. Then using (15.22), (15.18), and (15.21) we see that the estimation error \( \hat{w} - w \) can be written as

\[
\begin{align*}
\hat{w} - w &= Mz + Ny - w \\
&= Mz + (NC - L)x \\
&= Mz - MVx \\
&= Me
\end{align*}
\]

For \( \hat{w} \) to asymptotically estimate \( w \) it is therefore enough to make sure that \( e \) converges to zero as \( t \to \infty \).

Let us stipulate that the convergence of \( e \) is to be exponential by requiring \( e \) to satisfy a differential equation of the form

\[
\dot{e} = \bar{H}e \quad (15.24)
\]

where \( \bar{H} \) a suitably defined constant, exponentially stable matrix. In order to determine what’s required for (15.24) to hold, let us use the definition of \( e \) in (15.23) together with (15.22) and the (15.19)-(15.21) to write

\[
\begin{align*}
\dot{e} &= \dot{z} - V\dot{x} \\
&= Hz + Ky - VAx \\
&= Hz + (KC - VA)x \\
&= He + (HV + KC - VA)x
\end{align*}
\]

Therefore if we set \( H = \bar{H} \) and require \( HV + KC - VA = 0 \) to hold, then we will have satisfied the requirement that (15.24) hold.
In the light of the preceding we now call the linear system \( \{H, K, M, N\} \) defined by (15.21) an \( L \)-observer if for some matrix \( V \) the observer design equations

\[
\begin{align*}
L &= MV + NC \\
VA &= HV + KC
\end{align*}
\]

both hold. Thus if \( \{H, K, M, N\} \) is an \( L \)-observer, then the estimation error \( \hat{w} = w \) will satisfy

\[
\hat{w}(t) - w(t) = Me^{Ht}(z(0) - Vx(0))
\]

for all initial states \( x(0) \) and \( z(0) \). There remains the problem of determining matrices \( H, K, M, N \) and \( V \) for given \( C, A \) and \( L \), in such a way that the observer design equations hold.

### 15.3.2 Full-State Observers

Just about the easiest solution to the observer design problem that one can think of, is the one for which

\[
M = L, \quad N = 0, \quad V = I_{n \times n} \quad \text{and} \quad H = A - KC.
\]

Any observer of this type is called a full-state observer because in this case \( z \) is an asymptotic estimate of \( x \); i.e.

\[
z(t) - x(t) = e^{(A-KC)t}(z(0) - x(0))
\]

Of course we need to make sure that \( K \) is chosen so that \( A - KC \) is an asymptotically stable matrix. One way to accomplish this, assuming \((C, A)\) is an observable pair, is to exploit duality and use spectrum assignment. In particular, we know that the spectrum of \( A-KC \) is the same as the spectrum of \( A' - C'K' \). We also know that observability of \((C, A)\) is equivalent to controllability of \((A', C')\). Thus with an observability assumption, the problem of choosing \( K \) to asymptotically stabilize \( A - KC \) is the same as the problem of choosing \( K' \) to asymptotically stabilize the \( A' - C'K' \) with \((A', C')\) controllable.

No matter how one goes about defining \( K \), the definitions of \( H, M, N, \) and \( V \) given above show that a full-state observer is a system of equations of the form

\[
\dot{z} = (A - KC)z + K y \quad \hat{w} = z
\]

These are also the equations of a time-invariant Kalman filter. In the time-invariant case, the main difference between a full-state observer and a Kalman filter is the way in which \( K \) is defined. In the case of an observer, one typically appeals to spectrum assignment as we’ve just pointed out. In the Kalman filter case, \( K \) is chosen in systematic manner to deal with the trade-off between fast estimation, and reducing the effects of measurement noise on the estimate if \( x \). For the Kalman filtering problem to make sense, one must assume that the sensor output to be processed is of the form \( y + n \) where \( n \) is an additive noise process with known statistics. The way \( K \) is ultimately computed in this case is by solving what’s sometimes called a “matrix Riccati” equation. Under suitable assumptions such as observability of \((C, A)\), the \( K \) which results turns out to stabilize \( A - KC \). Kalman filtering is not restricted to time-invariant systems or even to continuous time systems. Kalman filtering has had a large impact in practice. The topic is typically addressed in detail in a course on estimation theory.

Before concluding this section we point out that a full-state observer always has a dimension equal to that of the process whose output \( w \) is to be estimated. Depending on \( A, C \) and \( L \) it is typically possible to construct observers of lower dimension with the same capabilities. For example if \( L = C \) or if \( C = I \) one need not not construct an observer at all. For in the first case \( w = y \) so we could define \( \hat{w} \overset{\Delta}{=} y \) while in the second we could define \( \hat{w} \overset{\Delta}{=} Ly \) since \( w = Ly \). Even under less drastic assumptions, it is always possible to generate an asymptotically correct estimate of \( w \) with an observer of dimension less than \( n \).
15.3.3 Forced Systems

So far we have been concerned with the problem of asymptotically estimating \( w = Lx \) from measurement of the output \( y = Cx \) of an unforced linear system with state equation \( \dot{x} = Ax \). In the case of a force \( n \)-dimensional linear system such as

\[
y = Cx, \quad \dot{x} = Ax + Bu \tag{15.27}
\]

\( w = Lx \) can still be estimated asymptotically provided \( u \) can be sensed. Thus \( u \) cannot be an unmeasurable input such as a “disturbance.” The way to go about estimating \( w \) in this case, is to first define observer matrices \( H, K, M, N, \) and \( V \) so that the observer design equations

\[
L = MV + NC \tag{15.28}
\]

\[
VA = HV + KC \tag{15.29}
\]

hold, just as before, second to generate the estimate \( \hat{w} \) using a modified observer of the form

\[
\dot{z} = Hz + Ky + VBu \tag{15.30}
\]

\[
\hat{w} = Mz + Ny \tag{15.31}
\]

We leave it to the reader to verify that if this is done, then with \( e \triangleq z - vx \),

\[
\dot{e} = He \]

and

\[
\hat{w}(t) - w(t) = Me^{Ht}(z(0) - Vx(0))
\]

just as before. These equations hold for all \( u \). In essence what’s happening is that the influence of \( u \) on \( w \) is being taken into account in just the right way by adding the term \( vBu \) to the differential equation for \( z \).

15.4 Observer-Based Controllers

Consider the \( n \)-dimensional linear system

\[
\dot{x} = Ax + Bu \tag{15.32}
\]

\[
y = Cx + Du \tag{15.33}
\]

\[
s = \bar{C}x \tag{15.34}
\]

which we again regard as an internal model \( \mathbb{P} \) of a physical process to be controlled. For any fixed \( F \) and \( G \), the state-feedback law

\[
u = Fx + Gr \tag{15.35}
\]

transforms \( \mathbb{P} \) into a system of the form

\[
\dot{x} = (A + BF)x + BGr \tag{15.36}
\]

\[
y = (C + DF)x + DGr \tag{15.37}
\]

\[
s = \bar{C}x \tag{15.38}
\]

But this control cannot typically be implemented because \( Fx \) may not be available as a sensor output. To deal with this situation, suppose, instead of (15.35) we implement a control of the form

\[
u = \hat{w} + Gr \tag{15.39}
\]
\[ \dot{z} = Hz + Ks + VBu \quad (15.40) \]
\[ \hat{w} = Mz + Ns \quad (15.41) \]

where (15.40) and (15.41) is an \( F \)-observer designed on the basis on the equations
\[ F = MV + N\hat{C} \quad (15.42) \]
\[ VA = HV + K\hat{C} \quad (15.43) \]

Note that (15.39) - (15.41) define a controller \( C \) of the form
\[ \dot{z} = (H + VBF)z + Ks + VBGr \quad (15.44) \]
\[ u = Mz + Ns + Gr \quad (15.45) \]

We call \( C \) an observer-based controller. The closed-loop system with input \( r \) and output \( y \) consisting of \( P \) and \( C \) is thus described by the system of equations
\[ y = [C + D\hat{C} \quad DM] [x] + DGr \quad [\dot{x}] = [A + BN\hat{C} \quad BM] [x] + [BG] r \quad (15.46) \]

To understand the effect of \( C \) on \( P \), let us first recall that
\[ \hat{w} = Fe = Me \quad (15.47) \]
and
\[ \dot{e} = He \quad (15.47) \]

where \( e \triangleq z - Vx \). Using the first of these equations and (15.39) it is possible to write
\[ u = Fx + Gr + Me \]

Application of this formula for \( u \) to the original open-loop equations defining \( P \) thus yields
\[ \dot{x} = (A + BF)x + BMe + BGr \quad (15.48) \]
\[ y = (C + DF)x + DMe + DGr \quad (15.49) \]

Combined with (15.47) we thus obtain the system
\[ y = [C + DF \quad DM] [x] + [\dot{e}] = [A + BF \quad BM] [x] + [BG] r \quad (15.50) \]

It is easy to verify that we could have gotten to these equations directly from (15.46) by simply making the change of variables
\[ \begin{bmatrix} x \\ e \end{bmatrix} = \begin{bmatrix} I & 0 \\ -V & I \end{bmatrix} \begin{bmatrix} x \\ z \end{bmatrix} \]

In other words, the systems defined by (15.46) and (15.50) are similar and thus have the same spectrum and the same transfer matrix. From this and the structure of (15.50) we can draw the following important conclusions:

- **Spectral Separation Property**: The spectrum of the closed-loop system consisting of \( P \) and the observer-based controller \( C \) defined by (15.44) and (15.45) separates into two disjoint subsets, one the spectrum of the observer matrix \( H \) and the other the spectrum of the system matrix \( A + BF \) which would have resulted had we been able to apply the state-feedback law \( u = Fx + Gr \) directly to \( P \).
• **Transfer Matrix Property:** The transfer matrix from \( r \) to \( y \) of the closed-loop system consisting of \( P \) and the observer-based controller \( C \) defined by (15.44) and (15.45) is exactly the same as the transfer matrix from \( r \) to \( y \) which would have resulted had we been able to apply the state-feedback law \( u = Fx + Gr \) directly to \( P \).

The spectral is a direct consequence of the block triangular structure of the system matrix appearing in (15.50).

The preceding provides a procedure for stabilizing any process model \( P \) which is controllable and which is observable through its sensor output \( s \). The spectrum can be freely assigned with \( F \) because of controllability of \( (A, B) \) and, at least in the case when a full-state observer is used, the spectrum of \( H \triangleq A - K \bar{C} \) with \( K \) because of observability of \( (\bar{C}, A) \). As you might suspect, there is much, much more to the story than this.
Bibliography


Index

A-invariant subspace, 109
L-observer, 231
$\mathbb{C}^n$, 32
$\mathbb{R}^n$, 32
\text{gcd}, 5
$\mathbb{C}$, 3
$\mathbb{C}[s]$, 4
$i$th unit $n$-vector, 12
$n$-vector, 8
$p$-norm, 52
$\mathbb{K}$, 3
$\mathbb{K}[s]$, 4
$\mathbb{R}$, 3
$\mathbb{R}[s]$, 4
lcm, 5

Fibonacci Sequence, 60
invariant, 80

Abel, Jacobi, Liouville Formula, 96
Abelian group, 143
adjoint, 145
algebraic multiplicity, 106, 124
algebraic system, 3
asymptotic estimate, 230
asymptotic stability matrices, 129, 214
attractor, 79
autonomous, 80, 87

Banach Spaces, 56
basis, 37
basis matrices, 110
Bessel’s inequality, 139
block diagonal matrix, 111
block triangular matrix, 111
bounded, 52, 67

Cartesian Product, 29
Cauchy Sequence, 55
Cauchy-Schwartz Inequality, 135
characteristic equation, 106
characteristic polynomial, 106
closed, 32
closed interval, 53
closed orbit, 79
closed set, 53
closed set, 55
Closed-loop control, 226
codomain, 40
cofactor, 9
column span, 46
commutative, 143
companion form, 119
complete, 56, 141
completion, 110
complex conjugate, 145
composition, 41
composition rule, 78, 94
congruence canonical form, 155
congruence transformation, 153
congruent, 153
connected, 53
constant coefficient, 87
contained, 32
continuous-time linear dynamical system, 165
continuously differentiable, 53, 216
control canonical form, 186
controllability Gramian, 174
controllability index, 181
controllability matrix, 179
controllability reduction, 176
controllable on $[t_0, T]$, 175
controllable space, 178
controlled output, 225
converge, 55
converge in the mean, 140
convergent sequence, 55
convolution integral, 167
convolution integrals, 97
convolution sum, 168
coordinates, 37
coprime, 5
cyclic, 117
cyclic matrix, 118
dependent variables, 58
derivative, 54
diagonal matrix, 8
diamond notation, 59
difference equation, 58
dimension, 40
dimension, 62
direct sum, 36
discrete dynamical system, 62
discrete-time asymptotic stability matrices, 130, 215
discrete-time linear dynamical system, 165
discrete-time Markov sequence, 169
discrete-time realization, 170
discrete-time stable matrices, 130, 216
discrete-time state transition matrix, 168
discrete-time transfer matrix, 169
discrete-time unstable matrices, 130, 216
discrete-time weighting pattern, 170
divides, 4
domain, 40
dot notation, 59
dual, 196
dynamical system, 62, 77
echelon form, 21
eigenspace, 105
eigenvalue, 105
eigenvector, 105
elementary column operations, 27
elementary divisors, 126
elementary row operation, 20
elementary row operations, 20
elements, 7
dendomorphism, 48
entries, 7
epimorphism, 46
equilibrium point, 78
equivalence class, 29
equivalence relation, 29
equivalent matrices, 28, 30
estimation error, 230
Euclidean metric, 131
Euclidean Norm, 52
Euclidean space, 131
Euler's method, 82
exponentially stable, 210
field, 3
finite dimensional, 36
first integral, 80
forcing function, 87
Fourier coefficients, 141
Fourier series, 141
full rank, 10
full-state observer, 231
function, 40
fundamental matrix, 91
Gauss Elimination, 22
general linear group, 144
generator, 117
geometric multiplicity, 106, 124
globally Lipschitz continuous, 67
golden mean, 61
Gramian, 135
greatest common divisor, 5
group, 143
Hermitian, 147
Hermitian metric, 131
Hilbert spaces, 141
homogeneous, 87
ideal, 4
identity, 143
identity matrix, 9
if and only if, 15
iff, 15
image, 45, 46
impulse matrix realization, 171
impulse response matrix, 167
indefinite matrices, 156
independent, 36
independent variables, 58
indefinite, 4
indistinguishable, 187
inf, 76
infnitum, 76
infinite dimensional, 36
infinity norm, 52
initial value problem, 66
injective, 47
inner product, 131
input, 87, 165
input, 165
input-output formula, 166
integral manifold, 80
intersection, 34
invariant, 112
invariant factors, 120
inverse, 143
invertible, 4
irreducible, 4
isometric, 145
isomorphic, 48
isomorphic systems, 206
isomorphism, 48
Jordon Block, 125
Jordon Normal Form, 126
kernel, 47
Laplace Expansion, 9
least common multiple, 5
left-half open interval, 53
left-right equivalence canonical form, 29
left-right equivalence transformation, 30
left-right equivalent matrices, 30
limit, 55
linear algebraic equations, 25
linear combination, 36
linear controller, 225
linear matrix equations, 27
linear state differential equation, 87
linear transformation, 41
linear vector spaces, 31
linearization, 87, 88
linearly dependent, 36
linearly independent, 36
Lipschitz Constant, 67
locally Lipschitz Continuous, 67
lower triangular matrix, 8
Lyapunov Transformation, 99
main diagonal, 8
Markov sequence, 167
Mathieu Equation, 98
matrix, 7
matrix addition, 11
matrix exponential, 92, 166
matrix inverses, 15
matrix inversion, 24
matrix multiplication, 12
matrix polynomial, 17
matrix representation, 43
maximal interval of existence, 76
maximal solution, 76
metric space, 131
minimal, 192
minimal polynomial, 114, 115
minimal realization, 205
minimum polynomial, 114
minor, 10
modular distributative rule, 35
monic, 4, 106
monomorphism, 47
natural numbers, 165
necessary and sufficient, 15
negative definite matrix, 156
negative semidefinite matrix, 156
neighborhood, 53
nominal solution, 88
nondestabilizing, 222
nonsingular, 15
norm, 51, 132
normal, 146
normalize, 134
normalized, 133, 134
normed vector space, 51
observability Gramian, 188
observability index, 197
observability matrix, 194
observability reduction, 190
observable on $[t_0, T]$, 189
observer design equations, 231
observer-based, 233
one-norm, 52
one-to-one, 47
onto, 46
open ball, 53
open interval, 53
open set, 53
open-loop control input, 225
order, 10, 58
ordinary differential equations, 58
ordinary recursion equation, 58
orthogonal, 133, 144
orthogonal complement, 145
orthogonal projection, 137
orthogonal set, 134
orthonormal, 134
orthonormal basis, 143
outer product, 132
output, 165
pairwise orthogonal, 133
partitioned matrix, 7
period, 79
periodic, 79
permutation group, 144
phase variable form., 63
Picard Iteration, 69
piecewise-continuous function, 54
positive definite, 156
positive definite matrix, 156
positive semi-definite function, 51
positive semidefinite, 156
positive semidefinite matrix, 156
positive-definite functions, 51
prime, 4
principal axes, 154
principal leading minors, 157
principal subarray, 157
process model, 225
proper, 32, 167
pulse response matrix, 168
quotient, 4
quotient set, 29
radially unbounded, 216
rank, 10
rational canonical form, 120
reachable, 173
reachable space, 173
readable, 225
readout equation, 166
real bilinear form, 152
real quadratic form, 152
realization, 170, 171, 201
reduced echelon form, 22
reducible, 100
reference input, 225
reflexivity, 29
relation, 29
remainder, 4
representation, 37
resolvent, 116
restriction, 65, 110
right half-open interval, 53
right inverse, 15
rings, 4
Runge-Kutta method, 83
sampling time, 169
scalar multiplication, 11, 31
scalars, 31
semisimple, 107
sensor output, 225
signature, 155
similar, 49, 103
similar linear systems, 183
similar systems, 206
similarity invariant, 112
similarity transformation, 49, 103
singular, 15
size, 7
span, 36
spans, 36
spectrum, 106
square matrix, 8
square root, 157
stable matrices, 129, 215
stable systems, 212
state, 77, 165
state equation, 166
state space, 77
state space system, 62
state transition matrix, 91, 129, 215
state variables, 62
state-feedback control, 226
stationary, 87, 165
step size, 81
strictly proper, 167
sub-multiplicative property, 52
subarray, 9
subset, 32
subspace, 32
sum, 33, 42
sup, 52
superposition rule, 41
supremum, 52, 67
surjective, 46
symmetric, 106, 147, 153
symmetric matrix, 8
symmetry, 29
time invariant, 87
time-invariant, 165
time-varying, 165
trace, 95
trajectory, 78
transfer matrix, 167
transfer matrix realization, 171
transitivity, 29
transpose, 10
triangle inequality, 51
uncontrollable spectrum, 184
unforced, 87
uniform convergence, 56
uniformly asymptotically stable, 210
uniformly stable, 209
union, 33
unit, 143
unit n-vectors, 36
unit vector, 8
unitary group, 145
unitary matrix, 145
unitary space, 131
unobservable space, 194
unobservable space, 188
unobservable spectrum, 200
unstable linear systems, 212
unstable matrices, 129, 215
upper triangular matrix, 8
value, 40
variation of constants formula, 97
vector addition, 31
vectors, 31
weighting pattern, 170
zero matrix, 8
zero vector, 8, 31