1 Quantifiers, Sets, Cartesian Products

We will make frequent use of the quantifiers $\exists$ ("there exists") and $\forall$ ("for all"). Statements containing these quantifiers must be parsed carefully.
For example let $\mathcal{C}$ be the set of all cats, and let $\mathcal{D}$ be the set of all dogs. The two statements

$$\exists\, c \in \mathcal{C} \;:\; \forall\, d \in \mathcal{D} \;\; c \text{ runs faster than } d$$
$$\forall\, c \in \mathcal{C} \;\; \exists\, d \in \mathcal{D} \;:\;\; c \text{ runs faster than } d$$

mean entirely different things. The first statement asserts the existence of a super-cat that runs faster than every dog. The second simply says that given any cat, we can find some poor dog that runs slower than that cat.

We will often deal with a set of objects x that have some property π.
We will write this set as $\{x : \pi\}$ For example, the set of points inside a circle of radius 1 on the xy plane is written as

$$\left\{(x, y) :\; x^2 + y^2 \leq 1\right\}$$

Let X and Y be two sets. The Cartesian product of X and Y, written $X \times Y$ is the set of all ordered pairs consisting of one element from X and and one element from Y. More succinctly,

$$X \times Y = \{(a, b) :\; a \in A, b \in B\}$$

2 Functions, 1-1, Onto

Finally, we will deal with functions. Let X and Y be two sets. A function f from X to Y is written

$$f : X \to Y$$

For every element $x \in X$, this function assigns some element $f(x) \in Y$.
The set X is called in domain, and the set Y is called the co-domain.
The image of f is the set of elements in Y that are assigned to some element of X. We can write

$$\text{Image}(f) = \{y \in Y :\; \exists\, x \in X \text{ such that } f(x) = y\}$$

The function f is called surjective (on-to) if $\text{Image}(f) = Y$, i.e. if

$$\forall\, y \in Y \;\; \exists\, x \in X \;:\; f(x) = y$$

The function f is called injective or one-to-one if distinct elements in X get assigned distinct elements in Y, i.e. $x_1 \neq x_2 \implies f(x_1) \neq f(x_2)$, or (equivalently)

$$f(x_1) = f(x_2) \implies x_1 = x_2$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3 Vectors and Matrices

Concepts such as vector addition, orthogonality, the angle between vectors, dot products, changing co-ordinate systems are intimately familiar in three dimensions. As we shall see, Linear Algebra allows us to generalize these concepts to treat vectors in many dimensions, and to even more abstract notions of vectors.

The set of all $n \times 1$ column vectors of real numbers is written $\mathbb{R}^n$. The i^{th} entry of a column vector v is denoted v_i. Analogous is the set $\mathbb{C}^n$ where the vectors have complex entries. We add vectors component-wise.

The set of all $m \times n$ rectangular matrices of real numbers is written $\mathbb{R}^{m\times n}$. The element in the i^{th} row and j^{th} column of a matrix M is written as m_{ij}. Similarly, we have the set $\mathbb{C}^{m\times n}$. The $n \times n$ matrix whose diagonal elements are 1 and whose other elements are all 0 is the identity matrix and is written I or I_n to explicitly exhibit its size.

We will assume familiarity with elementary matrix concepts such as matrix multiplication, transposes, determinants, and inverses. Matrices do not commute, i.e. in general, for matrices A and B of compatible dimension $AB \neq BA$. So we must be careful to respect the order in expressions with matrices. Indeed, if $v \in \mathbb{R}^n$, $v^T v$ is a real number, while vv^T is an $n \times n$ matrix.

It is useful to think of matrices in terms of their columns (or rows). For example, the product

$$Av = \left[\begin{array}{c|c|c} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{array}\right] \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = v_1 \begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} + v_2 \begin{bmatrix} a_{12} \\ a_{22} \\ a_{32} \end{bmatrix} + v_3 \begin{bmatrix} a_{13} \\ a_{23} \\ a_{33} \end{bmatrix} = \sum_{i=1}^{3} v_i a^{[i]}$$

is the weighted sum of the columns of A, the weights coming from the vector v.

4 Block-partitioned Matrices

It is also very useful to deal with block-partitioned matrices. For example, we could write

$$A = \left[\begin{array}{cc|ccc} a_{11} & a_{12} & a_{13} & b_{12} & b_{13} \\ a_{21} & a_{22} & a_{23} & a_{12} & a_{13} \\ \hline a_{31} & a_{32} & a_{33} & a_{12} & a_{13} \\ a_{31} & a_{32} & a_{33} & a_{12} & a_{13} \end{array}\right] = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]$$

Multiplication of block-partitioned matrices is as one would expect.

$$AB = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \left[\begin{array}{c|c} B_{11} & B_{12} \\ \hline B_{21} & B_{22} \end{array}\right] = \left[\begin{array}{c|c} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ \hline A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{array}\right]$$

This, of course, requires A and B to be partitioned conformably, i.e. so that the various products make sense. For example, in the equation above we would need A_{11} to have as many columns as B_{11} has rows.

Through these notes, we will make assertions about matrices in various Theorems, Propositions, and Lemmas. To make the results more transparent, we will not always explicitly state the dimensions of the matrices involved. The reader should assume that the results hold only when the matrices have appropriately sensible dimensions. For example suppose we write A^{-1} in the statement of some Theorem. It will imply we are assuming that A is square and invertible. Or if we write the expression $I + AB$ in some lemma, it must imply that A is $m \times n$ and B is $n \times m$. But it does not imply that A and B are square matrices.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5 Linear Systems of Equations

Matrices can be used to write a collection of linear equations in several unknowns compactly. For example:

$$\begin{array}{rcl} x_1 + 2x_2 + 7x_3 &=& 4 \\ 3x_2 - 4x_3 &=& 8 \\ 5x_1 - 6x_2 &=& 9 \\ x_3 &=& -1 \end{array} \iff \begin{bmatrix} 1 & 2 & 7 \\ 0 & 3 & -4 \\ 5 & -6 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \\ 9 \\ -1 \end{bmatrix} \iff Ax = b$$

Any collection of linear equations can be written as

$$Ax = b \quad A \in \mathbb{C}^{m \times n}$$

Here x is the vector of n unknowns, and A, b are constructed from the coefficients of the m equations. There are only three possibilities:

(a) *Over-determined.* The equation $Ax = b$ has no solution.

(b) *Under-determined.* The equation $Ax = b$ has several solutions.

(c) The equation $Ax = b$ has only one solution.

6 Transposes and Adjoints

Definition 1. Let $A \in \mathbb{C}^{m \times n}$. The transpose of A written A^T is an $n \times m$ matrix whose i, j^{th} entry is $a_{j,i}$ where $a_{i,j}$ is the i, j^{th} entry of A.

The adjoint of A, written A^* is its complex-conjugate-transpose, i.e. $A^* = (\overline{A})^T$.

Lemma 2. *Properties of Adjoints.*

(a) $(A^*)^* = A$

(b) $(A + B)^* = A^* + B^*$

(c) $(AB)^* = B^* A^*$ □

7 Determinants and Inverses

Definition 3. Let $A \in \mathbb{C}^{n \times n}$. The inverse of A is the unique matrix A^{-1} such that $AA^{-1} = I$.

We will assume you are familiar with minors, co-factors, determinants, and computing inverses.

Lemma 4. *Properties of Inverses.*

(a) $AA^{-1} = A^{-1}A = I$

(b) $(AB)^{-1} = B^{-1}A^{-1}$

(c) $(A^*)^{-1} = (A^{-1})^*$ □

Lemma 5. *Properties of Determinants.*

Let A and B are square matrices of the same size. Then

(a) $\det(AB) = \det(BA) = \det(A)\det(B)$

(b) $\det\left(A^T\right) = \det(A)$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

(c) $\det(A^*) = \overline{\det(A)}$ □

Proposition 6. *Inverses for Block-partitioned Matrices.*

(a) *If X and Z are square and invertible, then*

$$\begin{bmatrix} X & Y \\ 0 & Z \end{bmatrix}^{-1} = \begin{bmatrix} X^{-1} & -X^{-1}YZ^{-1} \\ 0 & Z^{-1} \end{bmatrix}$$

(b) *If A and D are square, and D is invertible, then*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & BD^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ C & D \end{bmatrix}$$

(c) *Suppose A and D are square. If D and $\Delta = A - BD^{-1}C$ are invertible, then*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} \Delta^{-1} & -\Delta^{-1}BD^{-1} \\ -D^{-1}C\Delta^{-1} & D^{-1}C\Delta^{-1}BD^{-1} + D^{-1} \end{bmatrix}$$ □

Corollary 7. *(Determinants for Block-partitioned Matrices.)*

(a) *If X and Z are square, then*

$$\det \begin{bmatrix} X & Y \\ 0 & Z \end{bmatrix} = \det(X)\det(Z)$$

(b) *If A and D are square, and D is invertible, then*

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det\left(A - BD^{-1}C\right)\det(D)$$ □

8 Trace

Definition 8. Let $A \in \mathbb{C}^{n\times n}$. The trace of A, written $\mathrm{Tr}(A)$ is the sum of its diagonal entries,i.e.

$$\mathrm{Tr}(A) = \sum_{1}^{n} a_{ii}$$ □

Trace makes sense only for square matrices.

Lemma 9. *Properties of the Trace.*

(a) $\mathrm{Tr}(A+B) = \mathrm{Tr}(A) + \mathrm{Tr}(B)$

(b) $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$. *Here $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times m}$ so both AB and BA are square.*

(c) $\mathrm{Tr}(A^*A) = \sum_i \sum_j |a_{ij}|^2$ □

Proof:

(a) $\mathrm{Tr}(A+B) = \sum_i (a_{ii} + b_{ii}) = \sum_i a_{ii} \; + \; \sum_i b_{ii} = \mathrm{Tr}(A) + \mathrm{Tr}(B)$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

(b) This is a bit harder. Let $C = AB$. Then, the i, j^{th} element of C can be written as

$$c_{ij} = \sum_k a_{ik} b_{kj}$$

Now, we can compute

$$\begin{aligned} \mathrm{Tr}(AB) &= \sum_i c_{ii} = \sum_i \sum_k a_{ik} b_{ki} \\ &= \sum_k \sum_i b_{ki} a_{ik} = \mathrm{Tr}(BA) \end{aligned}$$

(c) First note that the i, j^{th} element of A^* is $\overline{a_{ji}}$. Next, we have

$$\mathrm{Tr}(A^* A) = \sum_i \sum_k \overline{a_{ki}} a_{ki} = \sum_i \sum_k |a_{ik}|^2$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Transfer Function Models

B. State space Models

C. Nonlinear Models

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Transfer Function Models

1 Linear Time-Invariant (LTI) ODEs

LTI input-output ordinary differential equation (ODE) models relate the input u and output y of a dynamical system through the differential equation

$$y^{[n]} + a_{n-1}y^{[n-1]} + \cdots + a_1\dot{y} + a_0 y = b_m u^{[m]} + b_{m-1}u^{[m-1]} + \cdots + b_1\dot{u} + b_o u \tag{1}$$

subject to the n initial conditions

$$y(0), \dot{y}(0), \cdots, y^{[n-1]}(0)$$

Here, $y^{[k]}$ means $d^k y/dt^k$. All the coefficients a_k, b_k in this model are real constants.

The order of the differential equation model (1) is n, which is the highest derivative of y.

The differential equation model (1) is linear because $y^{[n]}$ is a linear combination of lower derivatives of y and of u. There are no terms like u^2 or $\dot{y}u$ or $\tan\dot{y}$.

This model is time-invariant because the coefficients a_k, b_k are constants, not changing with time.

2 SISO Transfer Functions

We will first introduce some notation that will greatly simplify dealing with these models. Let us introduce the short-hand

$$s = \frac{d}{dt}, \quad s^2 = \frac{d^2}{dt^2}, \quad s^3 = \frac{d^3}{dt^3}, \quad \text{etc.}$$

This lets us compactly write expressions with derivatives. For example, we write

$$\left[3s^4 + 11s^3 - 6s + 10\right] y \quad \text{to mean} \quad 3\frac{d^4y}{dt^4} + 11\frac{d^3y}{dt^3} - 6\frac{dy}{dt} + 10y$$

Using this new notation, we can compactly write (1) as

$$[s^n + a_{n-1}s^{n-1} + \cdots + a_1 s + a_0]y = [b_m s^m + b_{m-1}s^{m-1} + \cdots + b_1 s + b_0]u$$

or even more succinctly as

$$a(s)y = b(s)u \quad \text{or} \quad y = \left[\frac{b(s)}{a(s)}\right] u \quad \text{or} \quad y = H(s)u, \text{ where } H(s) = \frac{b(s)}{a(s)}$$

Here, $H(s)$ is called the transfer function. It is the ratio of two polynomials, which is called a rational function in s.

We will work with transfer functions because they are much easier to manipulate, but it is important to keep in mind that transfer functions are just a convenient short-hand for LTI differential equations. We will refer interchangeably to a transfer function and its associated input-output differential equation. If we write

$$y = \left[\frac{3s+9}{s^2+4s+13}\right] u \tag{2}$$

this means that u and y are related by the differential equation

$$\frac{d^2y}{dt^2} + 4\frac{dy}{dt} + 13y = 3\frac{du}{dt} + 9u$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3 Proper, Strictly Proper Transfer Functions, and Relative Degree

The denominator polynomial $a(s)$ of a transfer function is called the characteristic polynomial. It has degree n, where n is the model order. The numerator polynomial $b(s)$ has degree m.

We call $\delta = n - m$ the relative degree of the transfer function $H(s)$. If $n = m$ ($\delta = 0$), we say the transfer function is proper. If $m < n$ ($\delta > 0$), we say the transfer function is strictly proper.

4 Poles and Zeros

The poles of the transfer function $H(s)$ are the roots of the denominator polynomial $a(s) = 0$. The zeros of the transfer function $H(s)$ are the roots of the numerator polynomial $b(s) = 0$. For example, the transfer function

$$H(s) = \left[\frac{3s+9}{s^2+4s+13}\right]$$

has two complex poles at $-2 \pm 3j$ and one zero at -3.

A transfer function of order n will have n poles, which can be real or complex. When they are complex, they always occur in conjugate pairs. This is because the coefficients of the denominator polynomial $a(s)$ are real. A transfer function has m zeros, where m is the degree of the numerator polynomial $b(s)$. If $b(s)$ is just a constant, the transfer function has no zeros.

5 The Algebra of Transfer Functions

Transfer functions transform differential equations into algebraic ones, which are easier to manipulate. As an example, consider first the cascade or series connection of LTI systems as shown in Figure 1. This connection is a block diagram representation of the differential equations

$$a(s)v = b(s)u, \quad c(s)y = d(s)v$$

Here v is an intermediate signal that we are not too interested in. Eliminating v gives

$$y = \frac{d(s)}{c(s)}v = \frac{d(s)}{c(s)} \cdot \frac{b(s)}{a(s)}u = \frac{d(s)b(s)}{c(s)a(s)}u = G(s) \cdot H(s)u$$

Thus the transfer function from u to y is just the product $P(s) = G(s)H(s)$. This only involves multiplication of polynomials. Of course, we can convert the transfer function $P(s)$ back into a differential equation.

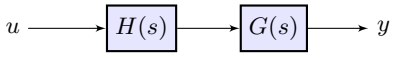

Figure 1: Cascade of two models.

6 MIMO Differential Equation Models

So far, we have mainly looked at single-input single-output systems modeled by linear time-invariant (LTI) ordinary differential equations. When we have multiple inputs and outputs, we will still have differential equations but there will be more of them. For example, we might have a 3 input, 2 output LTI system described by the equations:

$$\begin{aligned} \ddot{y}_1 + 2\dot{y}_1 + 6y_1 &= \ddot{u}_1 + \dot{u}_2 - 3u_3 \\ \dot{y}_2 + 3y_2 &= 2u_2 - \dot{u}_3 \end{aligned}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

7 Transfer Function Matrices

It is easy to convert differential equations to transfer functions in the MIMO case as well. For example, for the system above, we can write

$$\begin{aligned} s^2y_1 + 2sy_1 + 6y_1 &= s^2u_1 + su_2 - 3u_3 \\ sy_2 + 3y_2 &= 2u_2 - su_3 \end{aligned}$$

Re-arranging these equations we get

$$\begin{aligned} y_1 &= \frac{s^2}{s^2+2s+6}u_1 + \frac{s}{s^2+2s+6}u_2 - \frac{3}{s^2+2s+6}u_3 \\ y_2 &= \frac{2}{s+3}u_2 - \frac{s}{s+3}u_3 \end{aligned}$$

We can write this in matrix form as

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \frac{s^2}{s^2+2s+6} & \frac{s}{s^2+2s+6} & \frac{-3}{s^2+2s+6} \\ 0 & \frac{2}{s+3} & \frac{-s}{s+3} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$$

If we have an m input, p output LTI continuous-time system, we will have a $p \times m$ transfer function matrix relating the inputs u to the outputs y.

8 The Algebra of Transfer Function Matrices

We can manipulate MIMO transfer functions like matrices. Consider the interconnected models shown in Figure 2.
For the cascade interconnection (top), the transfer function is $G(s)H(s)$.
For the sum interconnections (middle): the transfer function is $G(s) + H(s)$.
For the feedback interconnection (bottom): we calculate the transfer-function matrix:

$$y = G(u - Hy) \implies (I + GH)\, y = Gu \implies y = (I + GH)^{-1}\, Gu$$

We have to be careful to respect the order in which these transfer-function matrices appear, because matrices do not necessarily commute.

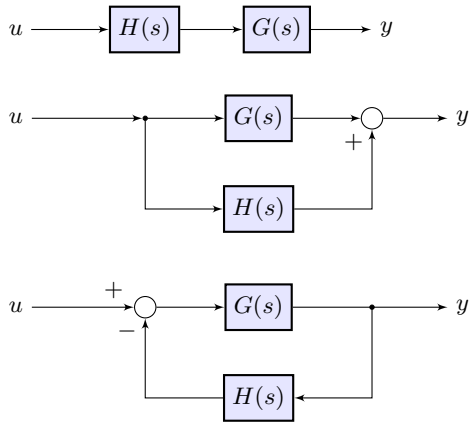

Figure 2: Interconnected models: (a) cascade, (b) sum, (c) feedback

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

B. State Space Models

1 State-space Models

We can rewrite n^{th} order differential equation models as a set of n coupled first-order differential equations. For example, consider the differential equation model

$$y^{[3]} + 3\ddot{y} + 4\dot{y} + 7y = u$$

or equivalently, the transfer function

$$y = \left[\frac{1}{s^3 + 2s^2 + 4s + 7}\right] u = [H(s)]\, u$$

Define the states by

$$x_1 = y, x_2 = \dot{y}, x_3 = \ddot{y}$$

We can then re-write this differential equation model as

$$\begin{aligned}
\dot{x}_1 &= \dot{y} = x_2 \\
\dot{x}_2 &= \ddot{y} = x_3 \\
\dot{x}_3 &= y^{[3]} = -3\ddot{y} - 4\dot{y} - 7y + u = -7x_1 - 4x_2 - 3x_1 + u \\
y &= x_1
\end{aligned}$$

These equations can be written is matrix form as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -7 & -4 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} u$$

$$y = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

or more compactly as

$$\left[\begin{array}{c} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \hline y \end{array}\right] = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -7 & -4 & -3 & 1 \\ \hline 1 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ x_3 \\ \hline u \end{array}\right]$$

More generally we will obtain n coupled first order differential equations that look like:

$$\Sigma \quad \begin{cases} \dot{x}(t) = Ax(t) + Bu(t) \\ y(t) = Cx(t) + Du(t) \end{cases}$$

Here A, B, C, D are matrices and Σ is called a state-space realization. Σ is said to realize the differential equation model or transfer function $H(s)$. We shall reserve the letters m, n, p for the numbers of inputs, states, and outputs respectively. The dimension of a realization Σ is the number of states n.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Note the sizes of the various matrices: $A \in \mathbb{R}^{n\times n}, B \in \mathbb{R}^{n\times m}, C \in \mathbb{R}^{p\times n}, D \in \mathbb{R}^{p\times m}$. The sizes of these matrices allow us to write Σ above more compactly using the packed-matrix notation as

$$\Sigma = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$$

or in text as $\Sigma(A, B, C, D)$.

State-space realizations offer a very convenient data structure to handle LTI ODEs.

State-space realizations are not unique. Many realizations yield the same transfer function.

The D-matrix captures the immediate effect of the input u on the output y. It is called the feed-through term. For most physical systems, the input u does not immediately affect the output y, so we often assume $D = 0$. In this case the realization Σ is called strictly proper.

2 The Concept of State

The state $x(t)$ at time t is the information you need at time t that, together with future values of the input, will let you compute future values of the output y.

For example, suppose we have a point mass subject to a force input. To predict the future motion of the particle at times $t > 0$, we need to know the force, and also the initial condition.The initial condition is the position and velocity of the particle at $t = 0$.

For an unconstrained rigid body moving in three dimensions, the state consists of the position and velocity of its center of mass, its orientation (three angles) and three angular velocities. So the state has 12 dimensions.

State-space realizations offer computational advantages over transfer-function representations. Transfer functions involve rational functions. Multiplying high-order rational functions and finding roots of high-degree polynomials can be numerically unstable. By contrast, computations with state-space realizations involve linear algebra: eigenvalue computation, least-squares, singular-value decomposition, etc. These operations are stable and computationally attractive. Indeed state-space methods can easily handle multi-input, multi-output, and high-order systems.

3 From State-space to Transfer Functions

It is easy to compute the transfer function from the realization

$$\Sigma \quad \left\{ \begin{array}{rcl} \dot{x}(t) & = & Ax(t) + Bu(t) \\ y(t) & = & Cx(t) + Du(t) \end{array} \right.$$

Using the s-notation, we can rewrite these equations as

$$sx = Ax + Bu \quad \Longrightarrow \quad (sI - A)x = Bu \quad \Longrightarrow \quad x = (sI - A)^{-1}Bu$$

Substituting this into the equation for y we get the transfer function from u to y realized by Σ:

$$H(s) = D + C(sI - A)^{-1}B$$

We write this as $\Sigma \sim H(s)$.

For example, consider the state space realization

$$\Sigma = \left[\begin{array}{ccc|ccc} 0 & -6 & 0 & -6 & 0 & -3 \\ -2 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & -1 & 0 & 2 & 3 \\ \hline 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & -1 \end{array}\right]$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

After some algebra, you can check that Σ realizes the 2-output 3-input transfer function matrix

$$H(s) = C(sI - A)^{-1}B + D = \begin{bmatrix} \frac{s(s-3)}{(s-4)(s+3)} & \frac{s}{(s-4)(s+3)} & \frac{6}{(s-4)(s+3)} \\ 0 & \frac{2}{s+1} & \frac{3}{s+1} - 1 \end{bmatrix}$$

4 Transfer Function Poles

Consider the realization

$$\Sigma = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right]$$

We have just seen that the associate transfer function is

$$\Sigma \sim H(s) = C(sI - A)^{-1}B + D = \frac{C\text{adj}(sI - A)B}{\det(sI - A)} + D$$

Here adj is the standard adjoint matrix used to calculate matrix inverses. There may be cancellations of terms between the numerator matrix $C\text{adj}(sI - A)B$ and the denominator polynomial $\det(sI - A)$. Also note that the roots of $\det(sI - A) = 0$ are the eigenvalues of A. Thus,

$$\text{poles of } H(s) \subseteq \text{eigenvalue of } A$$

5 Realization Theory

The inverse problem of building internal descriptions from transfer functions is less trivial and is the subject of realization theory. We begin by offering a closed form realization for SISO transfer functions. We will then use this to construct realizations of MIMO transfer function matrices.

Example 10. Consider the SISO transfer function with no zeros (numerator = 1):

$$y = \begin{bmatrix} H(s) \end{bmatrix} u = \left[\frac{1}{s^3 + 9s^2 + 4s + 5} \right] u$$

Define the states as: $x_1 = y, x_2 = \dot{y}, x_3 = \ddot{y}$. Then we have:

$$\dot{x_1} = x_2, \quad \dot{x_2} = x_3, \quad \dot{x_3} = -5x_1 - 4x_2 - 9x_3 + u, \quad y = x_1$$

We get the state-space realization of $H(s)$:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ y \end{bmatrix} = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -5 & -4 & -9 & 1 \\ \hline 1 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ u \end{bmatrix}$$

Example 11. Now let us include a numerator polynomial in $H(s)$:

$$y = \begin{bmatrix} H(s) \end{bmatrix} u = \left[\frac{3s^2 + 7s + 2}{s^3 + 9s + 4s + 5} \right] u = \left[3s^2 + 7s + 2\right] \cdot \left[\frac{1}{s^3 + 9s + 4s + 5} \right] u$$

Rewrite this as:

$$y = \left[3s^2 + 7s + 2\right] q, \quad q = \left[\frac{1}{s^3 + 9s^2 + 4s + 5} \right] u$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Define the states as: $x_1 = q, x_2 = \dot{q}, x_3 = \ddot{q}$. Then we have:

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_3, \quad \dot{x}_3 = -5x_1 - 4x_2 - 9x_3 + u, \quad y = 3x_3 + 7x_2 + 2x_1$$

We get the state-space realization of $H(s)$:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ y \end{bmatrix} = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -5 & -4 & -9 & 1 \\ \hline 2 & 7 & 3 & 0 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ u \end{bmatrix} \quad \text{or} \quad \Sigma = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{ccc|c} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ -5 & -4 & -9 & 1 \\ \hline 2 & 7 & 3 & 0 \end{array}\right]$$

The following observations make it easy to remember the form of the realization Σ:

(a) The denominator coefficients of $H(s)$ appear directly in the bottom row A matrix in reverse order with negative signs. The super-diagonal of the A matrix contains 1's, and the rest of the entries are 0's.

(b) The numerator coefficients of $H(s)$ appear directly in the C matrix in reverse order.

(c) The B matrix is easy to remember, and the D matrix is 0.

State-space realizations are not unique. We could re-order the states or define new states as linear combinations of the old states and get very different realization for the same transfer function.

Our choice of states in Example 10 was $x_1 = y, x_2 = \dot{y}, x_3 = \ddot{y}$. This doesn't always work. Try to figure out what our states are in Example 11 in terms of $u, \dot{u}, \ddot{u}$ and $y, \dot{y}, \text{and} \ddot{y}$.

6 Controllable Canonical Form

Example 11 easily extends to general SISO transfer functions. Consider the single-input single-output transfer function

$$H(s) = \frac{\beta_{n-1}s^{n-1} + \cdots + \beta_1 s + \beta_0}{s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0} + d$$

We can verify that

$$\Sigma = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] = \left[\begin{array}{ccccc|c} 0 & 1 & \cdots & 0 & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-2} & -\alpha_{n-1} & 1 \\ \hline \beta_0 & \beta_1 & \cdots & \beta_{n-2} & \beta_{n-1} & d \end{array}\right] \tag{3}$$

realizes $H(s)$. This realization is called the controllable canonical form. We stress that this form applies only to SISO transfer functions.

7 Observable Canonical Form

Let X^T denote the transpose of the matrix X. Since $H(s)$ is a single-input single-output transfer function, it follows that

$$H(s) = H(s)^T = \left[C(sI - A)^{-1}B + D\right] = B^T(sI - A^T)^{-1}C^T + D^T$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Thus,

$$\Sigma_{\text{dual}} = \left[\begin{array}{c|c} A^T & C^T \\ \hline B^T & D^T \end{array}\right] = \left[\begin{array}{ccccc|c} 0 & 0 & \cdots & 0 & -\alpha_0 & \beta_0 \\ 1 & 0 & \cdots & 0 & -\alpha_1 & \beta_1 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -\alpha_{n-2} & \beta_{n-2} \\ 0 & 0 & \cdots & 1 & -\alpha_{n-1} & \beta_{n-1} \\ \hline 0 & 0 & \cdots & 0 & 1 & d \end{array}\right] \tag{4}$$

also realizes $H(s)$. This is called the observable canonical form.

8 Realizations of sums of Transfer Functions

Suppose

$$G(s) \sim \left[\begin{array}{c|c} A_1 & B_1 \\ \hline C_1 & D_1 \end{array}\right], \quad H(s) \sim \left[\begin{array}{c|c} A_2 & B_2 \\ \hline C_2 & D_2 \end{array}\right]$$

Then, it is easy to check that

$$G(s) + H(s) \sim \left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \\ \hline C_1 & C_2 & D_1 + D_2 \end{array}\right]$$

9 Modal form: real distinct poles

Assume distinct poles

$$H(s) = d + \sum_k \frac{c_k}{s - p_k}$$

Using the discussion in point 8 above, we can write this as

$$H(s) \sim \begin{bmatrix} c_1 & c_2 & \cdots c_r \end{bmatrix} \begin{bmatrix} \frac{1}{s-p_1} & 0 & \cdots & 0 \\ 0 & \frac{1}{s-p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{s-p_r} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

This can be realized as

$$H(s) \sim \left[\begin{array}{cccc|c} p_1 & 0 & \cdots & 0 & 1 \\ 0 & p_2 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & p_r & 1 \\ \hline c_1 & c_2 & \cdots & c_r & d \end{array}\right]$$

10 Modal form: complex distinct poles

Consider the transfer function

$$H(s) = \frac{c_1(s+a) + c_2\omega}{(s+a)^2 + \omega^2}$$

It is easy to check that

$$H(s) \sim \left[\begin{array}{cc|c} -a & -\omega & 1 \\ \omega & -a & 0 \\ \hline c_1 & c_2 & 0 \end{array}\right]$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

11 Modal form: repeated real poles

It is easy to check that

$$\frac{c}{(s-p)^2} \sim \left[\begin{array}{cc|c} p & 1 & 0 \\ 0 & p & 1 \\ \hline c & 0 & 0 \end{array}\right]$$

This generalizes as

$$\frac{c}{(s-p)^m} \sim \left[\begin{array}{ccccc|c} p & 1 & 0 & \cdots & 0 & 0 \\ 0 & p & 1 & \cdots & 0 & 0 \\ 0 & 0 & p & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & p & 1 \\ \hline c & 0 & 0 & \cdots & 0 & 0 \end{array}\right] = \left[\begin{array}{c|c} pI+N & e_m \\ \hline ce_1^T & 0 \end{array}\right]$$

where I is the identity matrix, N is the matrix with 1's on the super-diagonal, $e_1 = [1\ 0\ \cdots\ 0]^T$.

12 Building Realizations: MIMO case

We can build realizations for multi-input multi-output transfer functions by successive use of the following result.

Theorem 12. *Let $\Sigma_1(A_1, B_1, C_1, D_1)$ and $\Sigma_2(A_2, B_2, C_2, D_2)$ be realizations of the transfer functions $H_1(s)$ and $H_2(s)$ respectively. Then,*

(a) $\Sigma_{column\ stack} = \left[\begin{array}{cc|c} A_1 & 0 & B_1 \\ 0 & A_2 & B_2 \\ \hline C_1 & 0 & D_1 \\ 0 & C_2 & D_2 \end{array}\right]$ *realizes the transfer function* $H(s) = \left[\begin{array}{c} H_1(s) \\ H_2(s) \end{array}\right]$

(b) $\Sigma_{row\ stack} = \left[\begin{array}{cc|cc} A_1 & 0 & B_1 & 0 \\ 0 & A_2 & 0 & B_2 \\ \hline C_1 & C_2 & D_1 & D_2 \end{array}\right]$ *realizes the transfer function* $H(s) = \left[\begin{array}{cc} H_1(s) & H_2(s) \end{array}\right]$

When we use this technique the resulting realization will generally have (unnecessarily) high dimension. We will later discuss methods for removing redundant or unnecessary states and finding minimal realizations from these less succinct composite realizations.

13 Building Realizations: Interconnected systems

1 Start with a general block diagram of interconnected transfer functions

2 Identify the inputs u and the outputs y of the block diagram

3 Label the states x_k , inputs u_k, and outputs y_k of transfer function k as

$$\begin{array}{rcl} \dot{x}_k & = & A_k x_k + B_k u_k \\ y_k & = & C x_k + D_k u_k \end{array}$$

4 Write the state-space equations of each TF using the controllable canonical form

5 The state of the interconnection is the union of the states of the component TFs

$$x = \left[\begin{array}{cccc} x_1 & x_2 & \cdots & x_r \end{array}\right]^T$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

6 Combine all equations to get by eliminating intermediate variables u_k, y_k as

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx + Du \end{aligned}$$

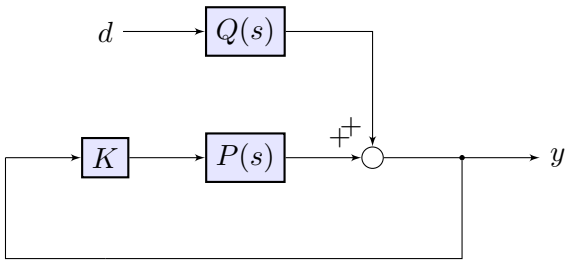

Figure 3: Realizing interconnected systems.

Example 13. Consider the block diagram shown in Figure 3. Here K is a constant, and $P(s)$ and $Q(s)$ have state-space realizations

$$P(s) \sim \left[\begin{array}{c|c} A_p & B_p \\ \hline C_p & 0 \end{array}\right] \qquad Q(s) \sim \left[\begin{array}{c|c} A_q & B_q \\ \hline C_q & 0 \end{array}\right]$$

Let x_p, x_q denote the states of $P(s)$ and $Q(s)$ respectively, and define

$$x = \begin{bmatrix} x_q \\ x_p \end{bmatrix}$$

We could stack the states of the subsystems P, Q in reverse order also to give a different realization. The realization of the interconnected system is

$$\begin{aligned} \dot{x}_q &= A_q x_q + B_q d \\ \dot{x}_p &= A_p x_p + B_p K \left(C_q x_q + C_p x_p\right) = \left(A_p + B_p K C_p\right) x_p + B_p K C_q x_q \\ y &= C_q x_q + C_p x_p \end{aligned}$$

or in matrix form:

$$\begin{bmatrix} \dot{x}_q \\ \dot{x}_p \\ y \end{bmatrix} = \left[\begin{array}{cc|c} A_q & 0 & B_q \\ B_p K C_q & A_p + B_p K C_p & 0 \\ \hline C_q & C_p & 0 \end{array}\right] \begin{bmatrix} x_q \\ x_p \\ d \end{bmatrix}$$

14 Time-varying State-space Models

This is for the case when we have linear ordinary differential equation models that are *not* time-invariant. We shall see later that such models naturally arise when we linearize nonlinear models about trajectories. State space realizations for such models have time-varying system matrices:

$$\Sigma \quad \begin{cases} \dot{x}(t) = A(t)x(t) + B(t)u(t) \\ y(t) = C(t)x(t) + D(t)u(t) \end{cases} \quad \text{or more compactly} \quad \left[\begin{array}{c|c} A(t) & B(t) \\ \hline C(t) & D(t) \end{array}\right]$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

15 **Discrete-time State-space Models** In discrete-time systems, differential equation models are replaced with difference equations, resulting in a state-space realization of the form

$$\Sigma \quad \begin{cases} x_{k+1} &= A_k x_k + B_k u_k \\ y_k &= C_k x_k + D_k u_k \end{cases} \qquad \text{or more compactly} \qquad \left[\begin{array}{c|c} A_k & B_k \\ \hline C_k & D_k \end{array}\right]$$

We will see that many concepts and results for discrete-time systems are analogous to the continuous-time case.

Example 14. The Fibonacci sequence defined by $y[0] = 0,\ y[1] = 1$, and $y[n] = y[n-2] + y[n-1]$ for integers $n > 1$ is an example of a LTI system (with no input) defined by a difference equation.

This can be put into the form of a discrete-time state-space model by taking the states as $x[n] = \begin{bmatrix} x_0[n] & x_1[n] \end{bmatrix}^\top$, $x_0[n] \triangleq y[n]$, and $x_1[n] \triangleq y[n-1]$. Then, with $x[0] = \begin{bmatrix} 0 & 1 \end{bmatrix}^\top$, the following model describes the Fibonacci sequence for integers $n \geq 0$.

$$x[n+1] = \underbrace{\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}}_{A} x[n]$$
$$y[n] = \underbrace{\begin{bmatrix} 1 & 0 \end{bmatrix}}_{C} x[n]$$

The values of the sequence can then be obtained by matrix operations. For example,
$y[2] = Cx[2] = CAx[1] = CA^2x[0] = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 1.$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

C. Nonlinear Models

1 Nonlinear State Space Models

So far, we have focused on linear time invariant (LTI) systems; however, many models of physical processes are nonlinear. Methods for LTI systems are nevertheless useful, as they can be applied to linear approximations of nonlinear systems. Essentially, feedback control algorithms make small adjustments to the inputs based on measured outputs. For small deviations of the input about some nominal input trajectory, the output of a nonlinear system looks like a small deviation around the nominal output. The effects of the small input deviations on the output is well approximated by a linear (possibly time-varying) system.

A nonlinear state space model for a system with m inputs and p outputs has the form

$$\begin{aligned}
\dot{x}_1 &= f_1(x_1, \cdots, x_n, u_1, \cdots, u_m) \\
&\vdots \quad\quad \vdots \\
\dot{x}_n &= f_n(x_1, \cdots, x_n, u_1, \cdots, u_m) \\
y_1 &= h_1(x_1, \cdots, x_n, u_1, \cdots, u_m) \\
&\vdots \quad\quad \vdots \\
y_p &= h_p(x_1, \cdots, x_n, u_1, \cdots, u_m)
\end{aligned}$$

Here each of the possibly nonlinear functions $f_1, \cdots, f_n$ and $h_1, \cdots, h_p$ are scalar-valued. By introducing the notation:

$$x = [x_1, \cdots, x_n]^T, \quad u = [u_1, \cdots, u_m]^T, \quad y = [y_1, \cdots, y_p]^T$$

we can write these equation more compactly in standard state-space form as

$$\begin{aligned}
\dot{x} &= f(x, u) \\
y &= h(x, u)
\end{aligned}$$

Note that in this form, the nonlinear functions f and h are vector-valued. This means that there are actually many scalar functions contained in the symbols f and h.

In the special case of LTI systems, $f(x, u) = Ax + Bu$, $h(x, u) = Cx + Du$.

2 Example: Magnetically Suspended Ball

Consider a ball of mass m suspended by an electromagnet as shown below. Let y be the position of the ball, measured down from the base of the electromagnet. If a current u is injected into the coils of the electromagnet, it will produce a magnetic field pulling the ball with force $F_{up} = -cu^2/y^2$. Note that the force decreases as $1/y^2$ because the effect of the magnet weakens when the ball is further away, and it is proportional to u^2 which is related to the power supplied to the magnet.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

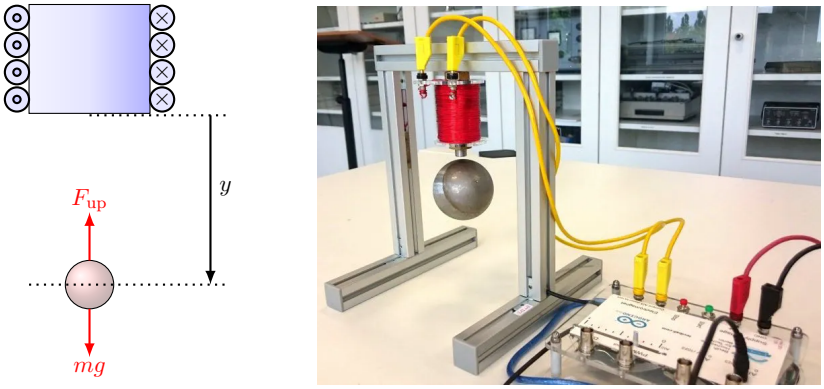

Figure 4: (a) Magnetically suspended ball. (b) Experimental set-up.

We can write a simple model for the motion of the ball from Newton's second law as

$$m\ddot{y} = mg - \frac{cu^2}{y^2} \tag{5}$$

For numerical exercises, we use the following constants:

m	0.15 `Kg`	mass of ball
c	0.01 `Kg-m`^{3}`/Coul`^{2}	magnetic coupling constant

We can convert this nonlinear input-output differential equation model to state-space by defining the states:

$$x_1 = y, x_2 = \dot{y}$$

and obtain the state space equations:

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, x_2, u) = x_2 \\ \dot{x}_2 &= f_2(x_1, x_2, u) = g - \frac{cu^2}{mx_1^2} = 9.8 - \frac{1}{15}\left(\frac{u}{x_1}\right)^2 \\ y &= h(x_1, x_2, u) = x_1 \end{aligned}$$

3 Equilibrium Points

Consider the time-invariant nonlinear model

$$\dot{x} = f(x, u), \quad y = h(x, u) \tag{6}$$

An equilibrium point of (6) is a pair (x^{eq}, u^{eq}) such that $f(x^{eq}, u^{eq}) = 0$.

If we start with initial condition $x(0) = x^{eq}$ and apply the constant input $u(t) = u^{eq}$, then $x(t) = x^{eq}$ is a solution of the differential equation $\dot{x} = f(x, u)$. This is because $\dot{x}(t) = 0$ for $x(t) = x^{eq}$, which is constant. Under mild conditions, such as continuous differentiability of f, the solutions of the differential equation are unique, meaning $x(t) = x^{eq}$ is the only solution. Then we can assert that: if we start the system at x^{eq} and apply the constant input u^{eq}, then $x(t)$ stays at the equilibrium value x^{eq}.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

For a nonlinear system, there can be many equilibrium points. To find all the equilibrium points of a nonlinear system, we simple set

$$f(x^{eq}, u^{eq}) = 0$$

and solve for x^{eq}, u^{eq}.

Note that the output function $h(x, u)$ plays no role in finding equilibrium points, but it allows us to evaluate the equilibrium value of the output from

$$y^{eq} = h(x^{eq}, u^{eq})$$

Example 15. Returning to the magnetically suspended ball example, we calculate the equilibrium:

$$\begin{aligned} f(x^{eq}, u^{eq}) = 0 \quad &\implies \quad f_1(x_1^{eq}, x_2^{eq}, u^{eq}) = 0 \quad \text{and} \quad f_2(x_1^{eq}, x_2^{eq}, u^{eq}) = 0 \\ &\implies \quad x_2^{eq} = 0 \quad \text{and } 10 - \left(\frac{u^{eq}}{3.87 x_1^{eq}} \right)^2 = 0 \\ &\implies \quad x_2^{eq} = 0, \quad x_1^{eq} = 0.082 u^{eq} \end{aligned}$$

The equilibrium points are:

$$x_1^{eq} = 0.082\alpha, x_2^{eq} = 0, u^{eq} = \alpha$$

which are parameterized by the free variable $u^{eq} = \alpha$. The resulting value of output is

$$y^{eq} = h(x^{eq}, u^{eq}) = x_1^{eq} = 0.082\alpha$$

We could also parameterize these equilibrium points by the free variable $y^{eq} = \beta$ as follows:

$$x_1^{eq} = \beta, x_2^{eq} = 0, u^{eq} = 12.12\beta$$

4 Jacobians and the Taylor Expansion

First consider a scalar valued function $\phi(x_1, x_2, \cdots, x_n)$ of n variables. The Jacobian of ϕ is the row vector

$$\frac{\partial \phi}{\partial x} := \begin{bmatrix} \frac{\partial \phi}{\partial x_1} & \cdots & \frac{\partial \phi}{\partial x_n} \end{bmatrix}$$

For example, for the function $\phi(x_1, x_2, x_3)$ of three variables,

$$\phi(x_1, x_2, x_3) = x_1 x_2^2 - 1/x_1 + \sin(x_3)$$

the Jacobian is

$$\frac{\partial \phi}{\partial x} = \begin{bmatrix} \frac{\partial \phi}{\partial x_1} & \frac{\partial \phi}{\partial x_2} & \frac{\partial \phi}{\partial x_3} \end{bmatrix} = \begin{bmatrix} x_2^2 + 1/x_1^2 & 2x_1 x_2 & \cos(x_3) \end{bmatrix}$$

The Jacobian is a function of the variables x_1, x_2, x_3. We can *evaluate* the Jacobian at a point and get actual numbers. In this example, if we evaluate the Jacobian at $(x_1, x_2, x_3) = (1, 2, 0)$:

$$\frac{\partial \phi}{\partial x}(1, 2, 0) = \begin{bmatrix} 5 & 4 & 1 \end{bmatrix}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Now consider a *vector-valued* function $\phi : \mathbb{R}^n \to \mathbb{R}^m$. This means that we have m individual functions in n variables, stacked into an m dimensional vector, ϕ. The variables are collectively represented by the n dimensional vector x. Let us write this function as

$$\phi(x) = \phi(x_1, x_2, \cdots, x_n) = \begin{bmatrix} \phi_1(x_1, \cdots, x_n) \\ \vdots \\ \phi_m(x_1, \cdots, x_n) \end{bmatrix}$$

The Jacobian of ϕ is the $m \times n$ matrix of functions

$$\frac{\partial \phi}{\partial x} = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial \phi_m}{\partial x_1} & \cdots & \frac{\partial \phi_m}{\partial x_n} \end{bmatrix}$$

We can evaluate the Jacobian at a point $\xi \in \mathbb{R}^n$ to get an $m \times n$ matrix of numbers. Our notation for this is

$$\frac{\partial \phi}{\partial x}(\xi) = \begin{bmatrix} \frac{\partial \phi_1}{\partial x_1} & \cdots & \frac{\partial \phi_1}{\partial x_n} \\ \vdots & \cdots & \vdots \\ \frac{\partial \phi_m}{\partial x_1} & \cdots & \frac{\partial \phi_m}{\partial x_n} \end{bmatrix}_{@\xi}$$

For example, consider the function

$$\phi(x_1, x_2, x_3) = \begin{bmatrix} \phi_1(x_1, x_2, x_3) \\ \phi_2(x_1, x_2, x_3) \end{bmatrix} = \begin{bmatrix} x_1 \sin(x_2) \\ x_2^2 \cos(x_3) \end{bmatrix}$$

The Jacobian of $\phi(x)$ evaluated at $(1, \pi/2, 0)$ is

$$\frac{\partial \phi}{\partial x}(1, \pi/2, 0) = \begin{bmatrix} \sin(x_2) & x_1 \cos(x_2) & 0 \\ 0 & 2x_2 \cos(x_3) & -x_2^2 \sin(x_3) \end{bmatrix}_{@(1,\pi/2,0)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \pi & 0 \end{bmatrix}$$

Analogous to the classical Taylor series expansion of a scalar function of one variable, we can write the Taylor series expansion of ϕ around ξ as

$$\phi(x) = \phi(\xi) + \frac{\partial \phi}{\partial x}(\xi)\,(x - \xi) + \text{ higher order terms}$$

The first two terms above form an affine linear approximation of ϕ around ξ, as illustrated geometrically in Figure 5.

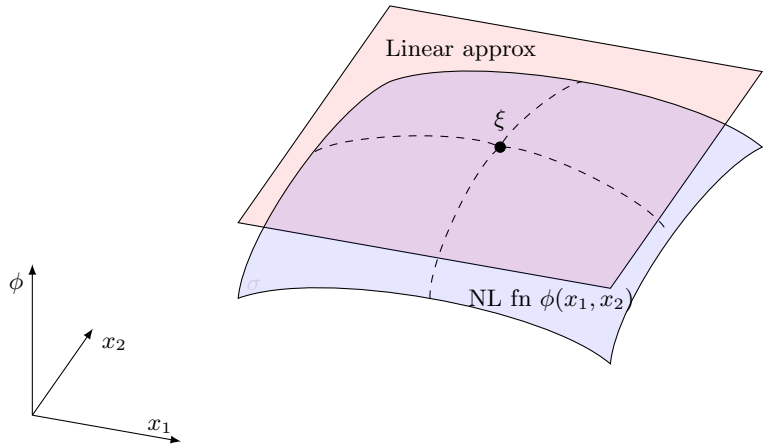

Figure 5: Geometric illustration of linearization.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5 Linearization about Equilibrium Points

We use the Taylor expansion to approximate a nonlinear model by a LTI model. This approxinmation, called Jacobian linearization, is good as long as the applied input $u(t)$ is close to its equilibrium value u_{eq} *and* the resulting state trajectory $x(t)$ is close to the equilibrium value x_{eq}.

Consider a nonlinear time-invariant input-output differential equation model realized as

$$\dot{x} = f(x,u), \quad y = h(x,u)$$

Let $\xi = (x^{eq}, u^{eq})$ be an equilibrium point of this model, and define $y^{eq} = h(x^{eq}, u^{eq})$.
We can approximate f in a neighborhood of ξ as

$$\begin{aligned} f(x,u) &\approx f(x^{eq},u^{eq}) + \frac{\partial f}{\partial x}(\xi)(x-x^{eq}) + \frac{\partial f}{\partial u}(\xi)(u-u^{eq}) \\ &= 0 + \underbrace{\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \dots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}_{@\xi}}_{=:A} (x-x^{eq}) + \underbrace{\begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \dots & \frac{\partial f_1}{\partial u_m} \\ \vdots & \dots & \vdots \\ \frac{\partial f_n}{\partial u_1} & \dots & \frac{\partial f_n}{\partial u_m} \end{bmatrix}_{@\xi}}_{=:B} (u-u^{eq}) \end{aligned}$$

Thus, if we define the deviation variables $\tilde{x} := x - x^{eq}$ and $\tilde{u} := u - u^{eq}$, we can write

$$\dot{\tilde{x}} = \dot{x} = f(x,u) \approx A\tilde{x} + B\tilde{u}$$

Note that linearization is meaningful only around an equilibrium because, otherwise, $f(x^{eq}, u^{eq}) \neq 0$ remains in the equation above and the approximation is no longer linear.

In a similar fashion we can approximate h in a neighborhood of ξ as

$$\begin{aligned} h(x,u) &\approx h(x^{eq},u^{eq}) + \underbrace{\begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \dots & \frac{\partial h_1}{\partial x_n} \\ \vdots & \dots & \vdots \\ \frac{\partial h_p}{\partial x_1} & \dots & \frac{\partial h_p}{\partial x_n} \end{bmatrix}_{@\xi}}_{=:C} (x-x^{eq}) + \underbrace{\begin{bmatrix} \frac{\partial h_1}{\partial u_1} & \dots & \frac{\partial h_1}{\partial u_m} \\ \vdots & \dots & \vdots \\ \frac{\partial h_p}{\partial u_1} & \dots & \frac{\partial h_p}{\partial u_m} \end{bmatrix}_{@\xi}}_{=:D} (u-u^{eq}) \\ &= y^{eq} + C\tilde{x} + D\tilde{u} \end{aligned}$$

Thus, the output deviation variable $\tilde{y} = y - y_{eq}$ can be approximated as

$$\tilde{y} \approx C\tilde{x} + D\tilde{u}$$

To summarize, Jacobian linearization at the equilibrium $\xi = (x^{eq}, u^{eq})$ approximates the nonlinear system by the following LTI model for the deviations of x, u, y from their equilibrium values:

$$\begin{aligned} \dot{\tilde{x}}(t) &= A\tilde{x}(t) + B\tilde{u}(t) \\ \tilde{y}(t) &= C\tilde{x}(t) + D\tilde{u}(t) \end{aligned}$$

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \dots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \dots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}_{@\xi} \qquad B = \begin{bmatrix} \frac{\partial f_1}{\partial u_1} & \dots & \frac{\partial f_1}{\partial u_m} \\ \vdots & \dots & \vdots \\ \frac{\partial f_n}{\partial u_1} & \dots & \frac{\partial f_n}{\partial u_m} \end{bmatrix}_{@\xi}$$

$$C = \begin{bmatrix} \frac{\partial h_1}{\partial x_1} & \dots & \frac{\partial h_1}{\partial x_n} \\ \vdots & \dots & \vdots \\ \frac{\partial h_p}{\partial x_1} & \dots & \frac{\partial h_p}{\partial x_n} \end{bmatrix}_{@\xi} \qquad D = \begin{bmatrix} \frac{\partial h_1}{\partial u_1} & \dots & \frac{\partial h_1}{\partial u_m} \\ \vdots & \dots & \vdots \\ \frac{\partial h_p}{\partial u_1} & \dots & \frac{\partial h_p}{\partial u_m} \end{bmatrix}_{@\xi}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

This approximation is good provided (x, u) remains sufficiently close to the equilibrium point $\xi = (x^{eq}, u^{eq})$; that is, $(\tilde{x}, \tilde{u})$ remains close to zero. This is guaranteed in control applications, where $\tilde{u}$ is designed as a function of $\tilde{x}$ to ensure $\tilde{x}$ and $\tilde{u}$ indeed remain close to zero. This is the subject of stabilization by feedback, which we will study later.

Example 16. Consider again the magnetically suspended ball. To linearize the dynamics about

$$x^{eq} = \begin{bmatrix} \beta \\ 0 \end{bmatrix} \quad u^{eq} = 12.12\beta \quad y^{eq} = \beta$$

we calculate the matrices

$$A = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix}_{@eq} = \begin{bmatrix} 0 & 1 \\ \frac{19.59}{\beta} & 0 \end{bmatrix} \qquad B = \begin{bmatrix} \frac{\partial f_1}{\partial u} \\ \frac{\partial f_2}{\partial u} \end{bmatrix}_{@eq} = \begin{bmatrix} 0 \\ \frac{-1.61}{\beta} \end{bmatrix}$$

$$C = \begin{bmatrix} \frac{\partial h}{\partial x_1} & \frac{\partial h}{\partial x_2} \end{bmatrix}_{@eq} = \begin{bmatrix} 1 & 0 \end{bmatrix} \qquad D = \begin{bmatrix} \frac{\partial h}{\partial u} \end{bmatrix}_{@eq} = 0$$

Note that the A-matrix here has eigenvalues at $\pm 4.43/\sqrt{\beta}$, one of which is in the right half-plane. This means that the equilibrium is unstable: small deviations will send the ball away from the equilibrium point. An appropriate feedback design based on the linear model will, however, stabilize this equilibrium locally.

6 Linearization about Nominal Trajectories

The ideas of the previous item can be easily extended to linearize a nonlinear system about a nominal trajectory. To this end, consider a general nonlinear time-invariant input-output differential equation model realized as

$$\dot{x} = f(x, u), \quad y = h(x, u)$$

Let $\xi(t) = (x_{nom}(t), u_{nom}(t))$ be a state/input trajectory pair that satisfies the system equation

$$\dot{x}_{nom}(t) = f(x_{nom}(t), u_{nom}(t))$$

As before, introduce the deviation variables $\tilde{x} = x - x_{nom}(t)$, $\tilde{u} = u - u_{nom}(t)$, and $\tilde{y} = y - y_{nom}(t)$. We can then write

$$\begin{aligned} \dot{\tilde{x}} &= \dot{x} - \dot{x}_{nom}(t) = f(x, u) - f(x_{nom}(t), u_{nom}(t)) \\ &\approx \frac{\partial f}{\partial x}(\xi(t))\tilde{x} + \frac{\partial f}{\partial u}(\xi(t))\tilde{u} = A(t)\tilde{x} + B(t)\tilde{u} \end{aligned}$$

Similarly, we have

$$\begin{aligned} \tilde{y} &= y - y_{nom}(t) = h(x, u) - h(x_{nom}(t), u_{nom}(t)) \\ &\approx \frac{\partial h}{\partial x}(\xi(t))\tilde{x} + \frac{\partial h}{\partial u}(\xi(t))\tilde{u} = C(t)\tilde{x} + D(t)\tilde{u} \end{aligned}$$

This approximation is good provided the trajectories of the nonlinear system remain sufficiently close to nominal. Note that this approximation is a *linear time-varying* system. This is the most common source of linear time-varying systems in practical applications. The matrices $A(t), B(t), C(t), D(t)$ are obtained the same way as in the previous section, but they are now time varying because the Jacobians are evaluated over a trajectory $\xi(t)$ rather than at equilibrium ξ.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Algebraic Aspects

B. Geometric Aspects

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Algebraic Aspects

1 What is in this Section

We are intimately familiar with the 3-dimensional space $\mathbb{R}^3$ in which we live. Some of the vector spaces we will encounter in this course are less familiar than $\mathbb{R}^3$. They will have more dimensions, or even infinitely many dimensions. Indeed, the vectors may be sequences or functions. In this section we introduce vector spaces and study their algebraic properties.

2 Vector Spaces

A vector space is an abstraction that captures all the properties of $\mathbb{R}^3$. For example, the set of $m \times n$ real matrices $R^{m\times n}$ is also a vector space. Think of writing a matrix A as a long vector in $\mathbb{R}^{mn}$ by stacking its columns. We also have infinite dimensional vector spaces like sequence spaces and function spaces. The following definition captures all of these cases.

Definition 17. A vector space $\mathbb{V}$ is a set of vectors $\mathbb{V}$ together with a set of scalars $\mathbb{F}$ (either $\mathbb{R}$ or $\mathbb{C}$ in this course) and two operations: vector-vector addition $(+)$ and vector-scalar multiplication $(\circ)$ such that for all $\alpha, \beta \in \mathbb{F}$ and all $v_1, v_2, v_3 \in \mathbb{V}$ the following properties hold:

Closure	$v_1 + v_2 \in \mathbb{V},\ \alpha \circ v_1 \in \mathbb{V}$
Commutativity	$v_1 + v_2 = v_2 + v_1$
Associativity	$(v_1 + v_2) + v_3 = v_1 + (v_2 + v_3)$
Distribution	$\alpha \circ (v_1 + v_2) = \alpha \circ v_1 + \alpha \circ v_2$ $(\alpha + \beta) \circ v_1 = \alpha \circ v_1 + \beta \circ v_1$
Identity	There exists a vector $\underline{0} \in \mathbb{V}$ such that $v + \underline{0} = v$
Additive Inverse	For all $v \in \mathbb{V}$, there exists a $(-v) \in \mathbb{V}$ such that $v + (-v) = \underline{0}$

We omit the notation $\circ$ when clear from the context. If the set of scalars is $\mathbb{R}$ we have a real vector space. If the set of scalars is $\mathbb{C}$ we have a complex vector space.

3 Examples

Example 18. The following are examples of vector spaces:

- $\mathbb{R}^n$, $\mathbb{C}^n$ with component-wise addition and scalar multiplication.
- $\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$ with matrix addition and scalar multiplication.
- $\mathbb{R}[s]$ (polynomials in s with real coefficients) over $\mathbb{F} = \mathbb{R}$, with usual addition and scalar multiplication of polynomials.
- $C[a, b] = \{f : [a, b] \to \mathbb{R},\ f \text{ continuous}\}$ over $\mathbb{R}$ with pointwise addition and multiplication.
- The Lebesgue spaces $L_p[a, b],\ 1 \le p < \infty$ defined as

$$L_p[a, b] = \left\{ f : [a, b] \to \mathbb{R} \ : \ \int_a^b |f(t)|^p dt < \infty \right\}$$

  are vector spaces over $\mathbb{R}$ with pointwise addition and multiplication.
- The space ℓ_p of sequences of real numbers as

$$\ell_p = \left\{ v = [v_1, v_2, \cdots, v_k, \cdots] : \quad v_k \in \mathbb{R},\ \sum_{k=1}^{\infty} |v_k|^p < \infty \right\}$$

  over $\mathbb{R}$ with pointwise addition and multiplication.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

4 Linear Dependence, Independence, Basis, Dimension

Definition 19. Let $\mathbb{S} = \{v_i : \ i \in I\}$ be a set (possibly infinite) of vectors from $\mathbb{V}$. The expression

$$\sum_{i \in I} \alpha_i v_i$$

with at most *finitely* many α_i being *nonzero* is called a finite linear combination of vectors from $\mathbb{S}$. If further, *not all* α_i are zero, it is called a nontrivial linear combination. □

For example, consider the vectors in $\mathbb{R}^3$:

$$v_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \ v_2 = \begin{bmatrix} 2 \\ 4 \\ 6 \end{bmatrix}, \ \ v_3 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \ v_4 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

Stack these vectors side-by-side. Then any linear combination of these vectors looks like

$$\alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 v_3 + \alpha_4 v_4 = \left[\begin{array}{c|c|c|c} 1 & 2 & 0 & 1 \\ 2 & 4 & 1 & 0 \\ 3 & 6 & 0 & 1 \end{array}\right] \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \end{bmatrix}$$

which is matrix-vector multiplication.

We only permit *finite* linear combinations. This is because we know how to add a finite number of vectors. To add infinitely many vectors, we need a notion of convergence. We write

$$w = \sum_{k=0}^{\infty} \alpha_k v_k$$

if the sequence of partial sums converges to w:

$$\lim_{N \to \infty} \sum_{k=0}^{N} \alpha_k v_k \ \to \ w$$

This means that this sequence gets closer and closer to w. For this, we need a notion of distance between two vectors. This is called a *norm* and will be introduced later.

Definition 20. The set $\mathbb{S}$ of vectors is called linearly dependent if there exists a nontrivial, finite, linear combination of vectors from $\mathbb{S}$ such that

$$\sum_{i \in I} \alpha_i v_i = 0$$

Otherwise, the set of vectors $\mathbb{S}$ is said to be linearly independent. □

Definition 21. The dimension of a vector space $\mathbb{V}$ is the maximal number of linearly independent vectors in $\mathbb{V}$. □

Definition 22. A set $\mathbb{B}$ of vectors in $\mathbb{V}$ is called a basis for $\mathbb{V}$ if every vector in $\mathbb{V}$ can be *uniquely* expressed as a finite linear combination of vectors in $\mathbb{B}$. □

Example 23. The following examples illustrate the concepts we have covered so far.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

- ◇ In the vector space $\mathbb{R}^2$,

$$v_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix},\ v_2 = \begin{bmatrix} 2 \\ 3 \end{bmatrix},\ v_3 = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$$

  is a set of linearly dependent vectors because $-v_1 + 2v_2 + v_3 = 0$.
- ◇ The set of vectors $\mathbb{S} = \{1, t, t^2, \cdots\}$ are linearly independent in $\mathcal{C}[0, 1]$.
- ◇ The dimension of $\mathbb{R}^n$ is n and the set

$$\mathbb{B} = \left\{ \begin{bmatrix} 1 \\ 0 \\ . \\ . \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ . \\ . \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ 0 \\ . \\ 1 \end{bmatrix} \right\}$$

  qualifies as a basis for this vector space. This is called the standard basis for $\mathbb{R}^n$. The i^{th} standard basis vector is written e_i or $e^{[i]}$.
- ◇ The dimension of $\mathbb{R}^{m \times n}$ is $m \cdot n$. Exhibit a basis for this space.
- ◇ Bases are not unique. For example, both 1 and -1 qualify as basis for $\mathbb{R}$. □

Theorem 24. *Let $\mathbb{V}$ be an n-dimensional vector space. Let $\mathbb{B}$ be a collection of n linearly independent vectors drawn from $\mathbb{V}$. Then $\mathbb{B}$ is a basis.* □

Proof. Let $\mathbb{B} = \{b_1, b_2, \cdots, b_n\}$ be a set of n linearly independent vectors in $\mathbb{V}$.
Fix any vector $v \in \mathbb{V}$. Then the set of vectors $\{v, b_1, b_2, \cdots, b_n\}$ contains $n + 1$ vectors. This set must be linearly *dependent* because $\dim(\mathbb{V}) = n =$ the maximal number of linearly independent vectors in $\mathbb{V}$. Therefore, there exist scalars $\alpha_i, i = 0, \cdots n$ not all zero such that

$$\alpha_0 v + \alpha_1 b_1 + \alpha_2 b_2 + \cdots + \alpha_n b_n = 0$$

Suppose $\alpha_0 = 0$. Then

$$\alpha_1 b_1 + \alpha_2 b_2 + \cdots + \alpha_n b_n = 0$$

with some α_i being nonzero, i.e. the set $\mathbb{B}$ is linearly dependent which contradicts the hypothesis. Thus $\alpha_0 \neq 0$. Then we can write

$$v = -\alpha_0^{-1}\alpha_1 b_1 - \alpha_0^{-1}\alpha_2 b_2 - \cdots - \alpha_0^{-1}\alpha_n b_n$$

Thus v is expressible as a linear combination of vectors in $\mathbb{B}$.
We now establish that this representation is unique. Suppose

$$\begin{aligned} v &= \sum_{i=1}^{n} \alpha_i b_i \\ &= \sum_{i=1}^{n} \beta_i b_i \end{aligned}$$

Subtracting we arrive at

$$\sum_{i=1}^{n} (\alpha_i - \beta_i)\, b_i = 0$$

Since $\mathbb{B}$ is linearly independent, it must be the case that all the coefficients in the linear combination above are zero, i.e. $\alpha_i = \beta_i,\ i = 1, \cdots, n$. This establishes that $\mathbb{B}$ is a basis. □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5 Subspaces

Definition 25. Let $\mathbb{V}$ be a vector space. A subset $\mathcal{S} \subseteq \mathbb{V}$ is called a subspace if $\mathcal{S}$ is itself a vector space. □

In $\mathbb{R}^2$ subspaces look like lines passing through the origin. In $\mathbb{R}^3$ subspaces look like planes or lines passing through the origin. These generalize to hyperplanes in $\mathbb{R}^n$. See Figure 6.

Here is a simple test for determining if $\mathcal{S}$ is a subspace.

Theorem 26. *A set $\mathcal{S} \subseteq \mathbb{V}$ is a subspace if and only if it is closed under vector addition and scalar multiplication, i.e. for all $\alpha \in \mathbb{F},\ w_1, w_2 \in \mathcal{S}$ we have*

$$\alpha w_1 \in \mathcal{S},\ w_1 + w_2 \in \mathcal{S}$$ □

The zero vector must lie in every (nonempty) subspace because we can select $\alpha = 0$.

Definition 27. Let $\mathbb{S} = \{v_i :\ i \in I\}$ be a set of vectors drawn from $\mathbb{V}$. Span$\{\mathbb{S}\}$ is the set of all finite linear combinations of vectors drawn from $\mathbb{S}$. □

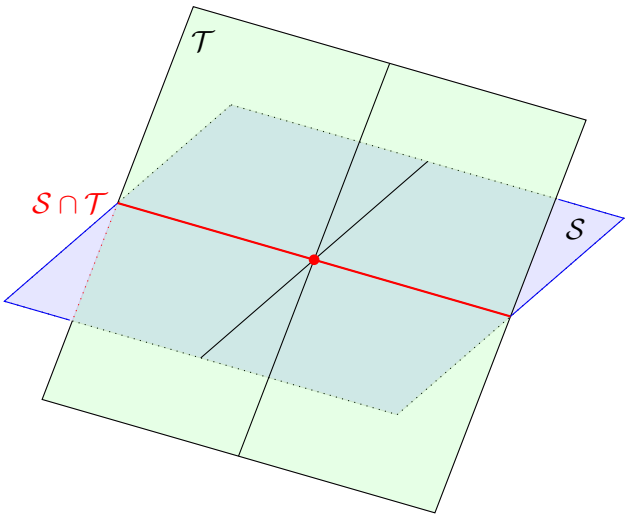

Figure 6: Subspaces in $\mathbb{R}^3$ are planes and lines through the origin: $\mathcal{S}$ and $\mathbb{T}$ and their intersection.

Theorem 26 implies that Span$\{\mathbb{S}\}$ is a subspace.

Definition 28. Let $\mathcal{S}$ and $\mathcal{T}$ be two subspaces of the vector space $\mathbb{V}$. The sum of subspaces is defined as

$$\mathcal{S} + \mathcal{T} = \{x \in \mathbb{V} :\ x = s + t,\ s \in \mathcal{S} \text{ and } \ t \in \mathcal{T}\}$$ □

The intersection $\mathcal{S} \cap \mathcal{T}$ and sum $\mathcal{S} + \mathcal{T}$ are both subspaces, but the union $\mathcal{S} \cup \mathcal{T}$ is *not* a subspace.

6 Direct-sum Decomposition

Definition 29. Let $\mathbb{V}$ be a vector space. If

$$\mathbb{V} = \mathcal{S}_1 + \mathcal{S}_2$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

we say that $\mathbb{V}$ is a (subspace) decomposition of S_1 and S_2.

If further $\mathcal{S}_1 \cap \mathcal{S}_2 = \{0\}$, we say that $\mathbb{V}$ is a direct sum decomposition of S_1 and S_2. This is written as

$$\mathbb{V} = \mathcal{S}_1 \oplus \mathcal{S}_2$$

Lemma 30. *Let* $\mathbb{V} = \mathcal{S}_1 \oplus \mathcal{S}_2$. *Then any vector* $v \in \mathbb{V}$ *can be* uniquely *expressed as*

$$v = y_1 + y_2 : \quad y_i \in \mathcal{S}_i$$

Proof. Since $\mathbb{V} = \mathcal{S}_1 + \mathcal{S}_2$, we know that any vector $v \in \mathbb{V}$ can be expressed as

$$v = y_1 + y_2 : \quad y_i \in \mathcal{S}_i$$

We only have to show uniqueness. So suppose we have another decomposition

$$v = z_1 + z_2 : \quad z_i \in \mathcal{S}_i$$

Subtracting, we get

$$0 = (y_1 - z_1) + (y_2 - z_2) \quad \Longrightarrow \quad \mathcal{S}_1 \ni (y_1 - z_1) = -(y_2 - z_2) \in \mathcal{S}_2$$

This implies

$$(y_1 - z_1) \text{ and } (y_2 - z_2) \in \mathcal{S}_1 \cap \mathcal{S}_2 = 0$$

establishing that $y_1 = z_1$ and $y_2 = z_2$ completing the proof. □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

B. Geometric Aspects

1 What is in this Section

So far we have dealt with the algebra of vector spaces. We now need a "ruler" to measure the length of a vector. This ruler is called the *norm* and is defined precisely below. Norms play a key role in analysis. For example, when we say that a sequence of vectors converges to v, i.e.

$$v_1, v_2, v_3, \cdots \to v$$

we mean that v_k gets closer and closer to v. We use the norm to make sense of the phrase *closer*.

We also need a notion of "angle" between vectors. For example, we need to be able to say when two vectors are perpendicular to each other. For this we introduce the notion of an *inner product*.

Norms and inner products equip vector spaces with a rich geometrical structure. They allow us to generalize our native Euclidean geometry in $\mathbb{R}^3$. We can speak of convergence, describe geometric objects, pose optimization problems, etc.

2 Norms

Definition 31. Let $\mathbb{V}$ be a vector space. A norm is a function $\|\cdot\| : \mathbb{V} \to \mathbb{R}$ such that for all $\alpha \in \mathbb{F},\ v, w \in \mathbb{V}$,

(a) (non-negative) $\|v\| \geq 0$ and $\|v\| = 0 \iff v = 0$.

(b) (scaling) $\|\alpha v\| = |\alpha|\ \|v\|$

(c) (triangle inequality) $\|v + w\| \leq \|v\| + \|w\|$.

A vector space on which a norm has been defined is called a normed space.

For a candidate norm, the only axiom that may be nontrivial to verify is the triangle inequality. This is illustrated in Figure 7. It asserts that in a triangle, the sum of the lengths of two sides is less than the third.

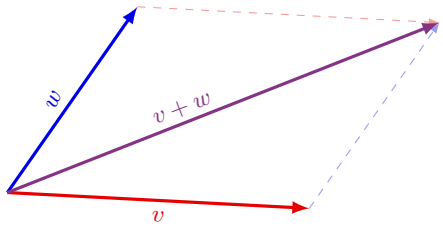

Figure 7: Triangle inequality.

3 Examples

On $\mathbb{R}^n$ or $\mathbb{C}^n$:

- $\|v\|_1 = \sum_{i=1}^n |v_i|$
- $\|v\|_2 = \left(\sum_{i=1}^n |v_i|^2\right)^{\frac{1}{2}} = \sqrt{v^* v}$ (2-norm)
- $\|v\|_\infty = \max_i\ |v_i|$ (sup norm)

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

where $v = \begin{bmatrix} v_1 & v_2 & \cdot & \cdot & v_n \end{bmatrix}^T$.

On the space of matrices $\mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$:

- ◇ The Frobenius norm

$$\|M\|_F = \left[\sum_{i=1}^{m}\sum_{j=1}^{n}|m_{i,j}|^2\right]^{\frac{1}{2}} = [\text{trace}(M^*M)]^{\frac{1}{2}}$$

  where $m_{i,j}$ is the i,j-th element of the matrix M. We can regard the matrix M as a long vector $\text{vec}(M)$ in $\mathbb{R}^{mn}$ by stacking its columns. Then the Frobenius norm of M is the same as the 2-norm of $\text{vec}(M)$.

- ◇ The induced 2-norm

$$\|M\|_2 = \sup_{x\neq 0}\frac{\|Mx\|_2}{\|x\|_2}$$

  Think of a matrix $M \in \mathbb{R}^{m\times n}$ as a black box. It takes in vectors in $x \in \mathbb{R}^n$ and produces vectors in $Mx \in \mathbb{R}^m$. Think of the 2-norm as a measure of energy. Then, the induced 2-norm is the maximum energy amplification the box can produce.

On the sequence spaces ℓ_p:

- ◇ $\|v\|_p = \left[\sum_{k=1}^{\infty}|v_k|^p\right]^{\frac{1}{p}}$ for $1 \leq p < \infty$
- ◇ $\|v\|_\infty = \sup_k |v_k|$ for $p = \infty$

On the function spaces $L_2[a,b]$, $L_p[a,b]$:

- ◇ $\|f\|_2 = \left[\int_a^b |f(t)|^2 dt\right]^{\frac{1}{2}}$
- ◇ $\|f\|_p = \left[\int_a^b |f(t)|^p dt\right]^{\frac{1}{p}}$, for $1 \leq p < \infty$

4 Equivalent Norms

Definition 32. Let $\mathbb{V}$ be a vector space, and let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two norms defined on $\mathbb{V}$. These norms are called equivalent if there exist constants m, M with $m \neq 0$ such that

$$m \leq \frac{\|v\|_a}{\|v\|_b} \leq M \quad \text{for all } v \in \mathbb{V}\setminus\{0\}$$

Lemma 33. *For finite-dimensional vector spaces, all norms are equivalent.*

5 Inner Products

Inner products generalize the familiar notion of dot products in $\mathbb{R}^3$ (where we multiply the vectors component-wise and add the results) to general vector spaces. The allow us to compute the angle θ between vectors v, w as

$$v \cdot w = \|v\|\,\|w\|\cos(\theta)$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Definition 34. Let $\mathbb{V}$ be a vector space. An inner product on $\mathbb{V}$ is a function $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \to \mathbb{C}$ such that

(a) $\langle v, w \rangle = \overline{\langle w, v \rangle}$ where $\overline{\langle w, v \rangle}$ denotes the conjugate of $\langle w, v \rangle$
(b) $\langle v, \alpha w \rangle = \alpha \langle v, w \rangle$
(c) $\langle v, w_1 + w_2 \rangle = \langle v, w_1 \rangle + \langle v, w_2 \rangle$
(d) $\langle v, v \rangle \geq 0, \quad \langle v, v \rangle = 0 \Leftrightarrow v = 0$

Inner products are linear in the second argument and conjugate-linear in the first argument:

$$\langle v + \alpha w, v + \alpha w \rangle = \langle v, v \rangle + \overline{\alpha} \langle w, v \rangle + \alpha \langle v, w \rangle + \alpha \overline{\alpha} \langle w, w \rangle$$

A vector space on which an inner product has been defined is called a inner product space.

6 Examples

- On $\mathbb{R}^n$, $\langle v, w \rangle = v^T w$
- On $\mathbb{C}^n$, $\langle v, w \rangle = v^* w$ where v^* denotes the conjugate transpose of v
- On $\mathbb{R}^{m \times n}$, $\langle A, B \rangle = \text{Trace}(A^T B)$
- On ℓ_2, $\langle v, w \rangle = \sum v_k w_k$
- On $L_2[a, b]$, $\langle f(t), g(t) \rangle = \displaystyle\int_a^b \overline{f(t)} g(t) \, dt$

7 Cauchy-Schwartz Inequality

Theorem 35. *(*Cauchy-Schwartz) *Let $\mathbb{V}$ be an inner product space. Then*

$$|\langle v, w \rangle| \leq \langle v, v \rangle^{\frac{1}{2}} \langle w, w \rangle^{\frac{1}{2}} \tag{7}$$

Proof. The result clearly holds if $w = 0$. Assume $w \neq 0$ and observe that

$$0 \leq \langle v + \alpha w, v + \alpha w \rangle = \langle v, v \rangle + \overline{\alpha} \langle w, v \rangle + \alpha \langle v, w \rangle + \alpha \overline{\alpha} \langle w, w \rangle$$

Let us choose $\alpha = -\langle w, v \rangle / \langle w, w \rangle$. With this choice, we obtain

$$\begin{aligned} 0 &\leq \langle v, v \rangle - \frac{\langle v, w \rangle \langle w, v \rangle}{\langle w, w \rangle} - \frac{\langle w, v \rangle \langle v, w \rangle}{\langle w, w \rangle} + \frac{\langle v, w \rangle \langle w, v \rangle}{\langle w, w \rangle} \\ &= \langle v, v \rangle - \frac{\langle v, w \rangle \langle w, v \rangle}{\langle w, w \rangle} \end{aligned}$$

This can be re-arranged as

$$|\langle v, w \rangle|^2 = \langle v, w \rangle \langle w, v \rangle \leq \langle v, v \rangle \langle w, w \rangle$$

proving the claim. □

An immediate consequence of the Cauchy-Schwartz inequality applied to the inner product space $L_2[a, b]$ is

$$\int_a^b f(t) g(t) \, dt \leq \left[\int_a^b f^2(t) \, dt \right]^{\frac{1}{2}} \left[\int_a^b g^2(t) \, dt \right]^{\frac{1}{2}}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Theorem 36. *Let* $\mathbb{V}$ *be an inner product space. Then*

$$\|v\| := \langle v, v\rangle^{\frac{1}{2}}$$

qualifies as a norm on $\mathbb{V}$.

Proof. We simply need to verify that $\|v\| = \langle v, v\rangle^{\frac{1}{2}}$ satisfies the defining properties of a norm. Clearly, $\|v\| \geq 0$. Also $\|v\| = 0$ if and only if $v = 0$.
Next, $\|\alpha v\| = \langle \alpha v, \alpha v\rangle^{\frac{1}{2}} = |\alpha| \langle v, v\rangle^{\frac{1}{2}} = |\alpha|\|v\|$.
To check the triangle inequality, we need the Cauchy-Schwartz inequality. Note that

$$\begin{aligned}
\|v+w\|^2 &= \langle v+w, v+w\rangle = \langle v, v\rangle + \langle w, w\rangle + \langle v, w\rangle + \langle w, v\rangle \\
&= \|v\|^2 + \|w\|^2 + \langle v, w\rangle + \langle w, v\rangle \\
&\leq \|v\|^2 + \|w\|^2 + |\langle v, w\rangle| + |\langle w, v\rangle| \\
&\leq \|v\|^2 + \|w\|^2 + 2\langle v, v\rangle^{\frac{1}{2}} \langle w, w\rangle^{\frac{1}{2}} \\
&= \|v\|^2 + \|w\|^2 + 2\|v\|\|w\| = (\|v\| + \|w\|)^2
\end{aligned}$$

proving the claim. □

The Cauchy-Schwartz inequality states that the magnitude of the dot product of two vectors is smaller than the product of the vector lengths:

$$|v \cdot w| \leq \|v\| \, \|w\|$$

The norm defined above is said to be induced by the inner product. In an inner product space this is the natural norm to use. Figure 8 illustrates the relationship between vector spaces, normed spaces, and inner product spaces.

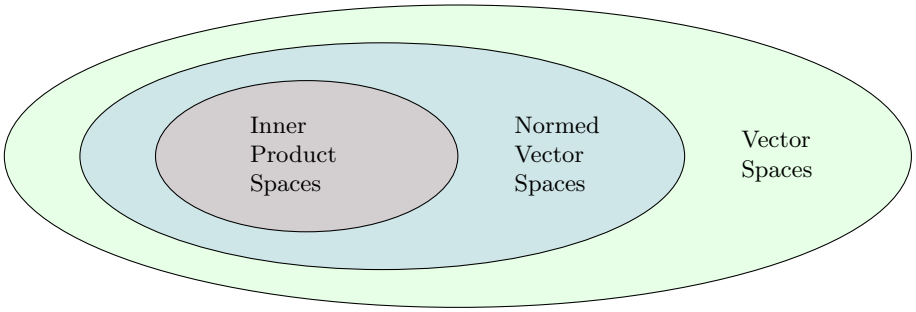

Figure 8: Vector spaces, normed vector spaces, inner product spaces.

8 Orthogonality

Definition 37. In an inner product space $\mathbb{V}$ two vectors v, w are said to be orthogonal if $\langle v, w\rangle = 0$. This is written as $v \perp w$.

Further, v is orthogonal to the set of vectors $\mathcal{S}$ if $v \perp w$ for all $w \in \mathcal{S}$. This is written as $v \perp \mathcal{S}$.

A set of vectors $\mathcal{S}$ is called orthogonal if

$$v \perp w \text{ for all } \; v \neq w, v, w \in \mathcal{S}$$

and is called orthonormal if in addition $\|v\| = 1$ for all $v \in \mathcal{S}$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Theorem 38. *(Pythagoras) If* $v \perp w$, *then* $\|v+w\|^2 = \|v\|^2 + \|w\|^2$

Proof. Observe that

$$\begin{aligned} \|v+w\|^2 &= \langle v+w, v+w\rangle = \langle v,v\rangle + \langle v,w\rangle + \langle w,v\rangle + \langle w,w\rangle \\ &= \langle v,v\rangle + \langle w,w\rangle = \|v\|^2 + \|w\|^2 \end{aligned}$$

9 Projection

Let $v \in \mathbb{R}^3$ be an arbitrary vector and let $b \in \mathbb{R}^3$ be a unit vector (i.e. a vector of length $\|b\| = 1$). Then the dot product $b \cdot v = \langle b, v\rangle$ is the length of the projection of v along the direction b. The projection itself is $\hat{s} = \langle b, v\rangle\, b$.

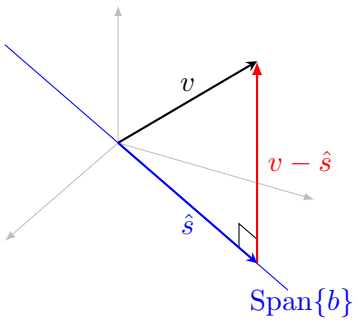

Figure 9: Projection onto Span$\{b\}$.

An important property of the projection $\hat{s}$ is that it is the closest point in Span$\{b\}$ to v and the difference $v - \hat{s}$ is perpendicular to Span$\{b\}$; see Figure 9. We now generalize these observations to an arbitrary inner product space:

Lemma 39. *Let* $\mathbb{V}$ *be an inner product space and fix* $v \in \mathbb{V}$. *Let* $b \in \mathbb{V}$ *be a unit vector (*$\|b\| = 1$*) and consider the subspace* $\mathcal{S} = Span\{b\}$. *We wish to find the closest point to* v *in the subspace* $\mathcal{S}$*; more precisely, we wish to solve the problem:*

$$\min_{s \in \mathcal{S}} \ \|v - s\|$$

(a) The solution is unique and given by

$$\hat{s} = \langle b, v\rangle\, b$$

(b) $v - \hat{s} \perp \mathcal{S}$

Proof. (b) When $\hat{s} = \langle b, v\rangle\, b$, we have

$$\langle b, v - \hat{s}\rangle = \langle b, v - \langle b, v\rangle\, b\rangle = \langle b, v\rangle - \ \langle b, v\rangle\, \langle b, b\rangle = \langle b, v\rangle - \ \langle b, v\rangle = 0$$

Since any $s \in \mathcal{S}$ can be written as $s = \alpha b$, we have $\langle s, v - \hat{s}\rangle = 0$, thus $(v - \hat{s}) \perp \mathcal{S}$.

(a) Next, we show that no other vector s in $\mathcal{S}$ is closer to v than $\hat{s}$ is. This is because

$$\|v - s\|^2 = \|v - \hat{s} - (s - \hat{s})\|^2 = \|v - \hat{s}\|^2 + \|s - \hat{s}\|^2 \geq \|v - \hat{s}\|^2$$

where the second equality follows from Pythagoras' theorem, because $s - \hat{s} \in \mathcal{S}$ and, from part (b) that we already proved, $v - \hat{s} \perp s - \hat{s}$. Note that the final inequality is strict if $s \neq \hat{s}$. Thus, any other point $s \in \mathcal{S}$ is farther from v than $\hat{s}$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

10 The Projection Theorem

We now extend the ideas above to find the projection of a vector v onto a subspace $\mathcal{S}$ that is more general than the span of a single vector.

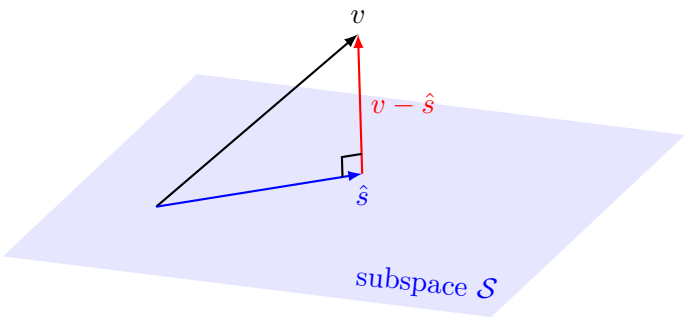

Figure 10: Projection onto a subspace $\mathcal{S}$.

Theorem 40. *(Projection Theorem) Let $\mathbb{V}$ be an inner-product space and fix $v \in \mathbb{V}$. Consider the subspace*

$$\mathcal{S} = Span\{b_1, b_2, \cdots, b_m\}$$

where the vectors $\{b_1, \cdots, b_m\}$ are linearly independent. We wish to solve the problem:

$$\min_{s \in \mathcal{S}} \|v - s\|$$

(a) The (unique) solution is

$$\hat{s} = \alpha_1 b_1 + \alpha_2 b_2 + \cdots + \alpha_m b_m$$

where the coefficients α_k solve the Gram equations:

$$\begin{bmatrix} \langle b_1, b_1 \rangle & \cdots & \langle b_1, b_m \rangle \\ \vdots & \ddots & \cdots \\ \langle b_m, b_1 \rangle & \cdots & \langle b_m, b_m \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \langle b_1, v \rangle \\ \vdots \\ \langle b_m, v \rangle \end{bmatrix}$$

(b) $v - \hat{s} \perp \mathcal{S}$

Before proving the theorem we remark that the $m \times m$ matrix in the Gram equations is invertible because the vectors $\{b_1, b_2, \cdots, b_m\}$ are linearly independent. (This will be proven in a homework problem.) Thus, the Gram equations admit the unique solution

$$\begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = \begin{bmatrix} \langle b_1, b_1 \rangle & \cdots & \langle b_1, b_m \rangle \\ \vdots & \ddots & \cdots \\ \langle b_m, b_1 \rangle & \cdots & \langle b_m, b_m \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle b_1, v \rangle \\ \vdots \\ \langle b_m, v \rangle \end{bmatrix}$$

.

We also note that, once part (b) is proven, the proof of part (a) is identical to that in the theorem of the previous section. Thus, we will only prove part (b).

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Proof. (b) Note that, for any $k \in \{1, \ldots, m\}$,

$$\begin{aligned} \langle b_k, v - \hat{s} \rangle &= \langle b_k, v \rangle - \langle b_k, \hat{s} \rangle = \langle b_k, v \rangle - \sum_i \langle b_k, \alpha_i b_i \rangle = \langle b_k, v \rangle - \sum_i \langle b_k, b_i \rangle \alpha_i \\ &= \langle b_k, v \rangle - \begin{bmatrix} \langle b_k, b_1 \rangle & \cdots & \langle b_k, b_m \rangle \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_m \end{bmatrix} = 0 \end{aligned}$$

where the last equality follows from the Gram equations. Thus $v - \hat{s} \perp b_k$ for all k, which implies $v - \hat{s} \perp \mathcal{S}$.

11 The Projection Theorem in $\mathbb{R}^n$

If we are working with the vector space $\mathbb{R}^n$ under the usual inner product, the Projection Theorem simplifies. Fix $v \in \mathbb{R}^n$ and consider the subspace

$$\mathcal{S} = \text{Span } \{b_1, b_2, \cdots, b_m\} \subseteq \mathbb{R}^n$$

Assume the vectors $\{b_1, \cdots, b_m\}$ are linearly independent. Stack these vectors side-by-side to form the matrix

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

Then, the solution to the optimization problem:

$$\min_{s \in \mathcal{S}} \ \|v - s\|$$

can be written succinctly as

$$\hat{s} = B(B^T B)^{-1} B^T v$$

This is known as the Least Squares approximation of v in the subspace $\mathcal{S}$, which is commonly used in sciences and engineering.

12 Gram-Schmidt Orthonormalization

Given a collection of vectors

$$\mathcal{B} = \{b_1, b_2, \cdots, b_n\}$$

from an inner product space $\mathbb{V}$, suppose we wish to construct a set of vectors

$$\mathcal{C} = \{c_1, c_2, \cdots, c_r\}$$

that forms an orthonormal basis for Span$\{\mathcal{B}\}$.

The Gram-Schmidt procedure does this iteratively using the Projection Theorem. When $b_1, b_2, \cdots, b_n$ are linearly independent, the Gram-Schmidt procedure consists of the following steps:

- ◇ Normalize b_1 to get a unit vector c_1 and put it in $\mathcal{C}$.
- ◇ Now take b_2 and finds its projection $\hat{s}$ onto the subspace $\mathcal{S}_1 = \text{Span } \{c_1\}$. From the Projection Theorem, $y_2 := b_2 - \hat{s}$ is orthogonal to $\mathcal{S}_1$. Normalize y_2 to get a unit vector c_2 and add it to $\mathcal{C}$. Note that $c_2 \perp c_1$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

- ◇ Then take b_3 and find its projection $\hat{s}$ onto the subspace $\mathcal{S}_2 = \text{Span}\,\{c_1, c_2\}$. From the projection theorem, $y_3 = b_2 - \hat{s}$ is orthogonal to $\mathcal{S}_2$. Normalize y_3 to get a unit vector c_3 and add it to $\mathcal{C}$. Note that $c_3 \perp c_1$ and $c_3 \perp c_2$.
- ◇ Continue this until $\mathcal{B}$ is exhausted.

If $b_1, b_2, \cdots, b_n$ are not linearly independent, some of the vectors $b_1, y_2, y_3, \ldots$ in the procedure above will be zero and, thus, can't be normalized. In this case we skip the corresponding step and continue to the next element of $\mathcal{B}$.

The resulting Gram-Schmidt algorithm is summarized below.

1	Initialize $k = 1$
2	If $b_1 \neq 0$, compute $c_1 = \frac{b_1}{\|b_1\|}$
3	While $k < n$
	$k = k + 1$
	$y_k = b_k - \sum_{i=1}^{k-1} \langle c_i, b_k \rangle \, c_i$
	If $y_k \neq 0$, compute $c_k = \frac{y_k}{\|y_k\|}$
3	end

Figure 11: Gram-Schmidt Procedure

Example 41. Consider the following set in $\mathbb{R}^3$:

$$\mathcal{B} = \left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \\ 5 \end{bmatrix} \right\}$$

We use the Gram-Schmidt procedure to find an orthonormal basis for Span $\{\mathcal{B}\}$.

$$y_1 = b_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \text{ and } c_1 = \frac{1}{\sqrt{14}} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$y_2 = b_2 - \langle c_1, b_2 \rangle \, c_1 = \frac{1}{7} \begin{bmatrix} 4 \\ 1 \\ -2 \end{bmatrix}, \text{ and } c_2 = \frac{7}{\sqrt{21}} \begin{bmatrix} 4 \\ 1 \\ -2 \end{bmatrix}$$

$$y_3 = b_3 - \langle c_1, b_3 \rangle \, c_1 - \langle c_2, b_3 \rangle \, c_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

Then $\mathcal{C} = \{c_1, c_2\}$ forms the desired orthonormal basis.

For $\mathbb{R}^n$ there are much more efficient methods to find orthonormal basis using the Singular Value Decomposition. We will learn about this later.

13 Orthogonal Complements

Definition 42. Let $\mathcal{S}$ be a subspace of $\mathbb{V}$. The orthogonal complement of $\mathcal{S}$ is the set

$$\mathcal{S}^\perp = \{v \in \mathbb{V} : \; v \perp \mathcal{S}\}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Lemma 43. *(a)* $\mathcal{S}^\perp$ *is a subspace.*

(b) $\mathcal{S} \cap \mathcal{S}^\perp = \{0\}$.

(c) If $\mathcal{S} \subset \mathcal{T}$*, then* $\mathcal{T}^\perp \subset \mathcal{S}^\perp$

Proof.

(a) Let $v, w \in \mathcal{S}^\perp$. Then,

$$\langle v, s\rangle = \langle w, s\rangle = 0 \text{ for all } s \in \mathcal{S}$$

Also for all $s \in \mathcal{S}$,

$$\langle v + w, s\rangle = \langle v, s\rangle + \langle w, s\rangle = 0 \quad \text{and} \quad \langle \alpha v, s\rangle = \alpha \langle v, s\rangle = 0$$

We have argued that αv and $v + w$ are in $\mathcal{S}^\perp$, proving that this is a subspace.

(b) Let $s \in \mathcal{S} \cap \mathcal{S}^\perp$. Then, $s \in \mathcal{S}$ and $s \in \mathcal{S}^\perp$. So

$$s \perp s \quad \Longrightarrow \quad 0 = \langle s, s\rangle = \|s\|^2 \quad \Longrightarrow \quad s = 0$$

(c) Suppose $x \in \mathcal{T}^\perp$. Then $\langle x, t\rangle = 0 \ \forall t \in \mathcal{T}$, which implies $\langle x, s\rangle = 0 \ \forall s \in \mathcal{S}$ since $\mathcal{S} \subset \mathcal{T}$. Thus $x \in \mathcal{S}^\perp$. We have shown $x \in \mathcal{T}^\perp \ \Rightarrow \ x \in \mathcal{S}^\perp$, which means $\mathcal{T}^\perp \subset \mathcal{S}^\perp$. □

The next lemma holds when the subspace $\mathcal{S}$ is finite dimensional.

Lemma 44. *Suppose* $\mathcal{S}$ *is finite dimensional. Then*

(a) $\mathbb{V} = \mathcal{S} \oplus \mathcal{S}^\perp$

(b) $(\mathcal{S}^\perp)^\perp = \mathcal{S}$

Proof. The proof uses the Projection Theorem, which is applicable to finite dimensional subspaces.

(a) Let v be any vector in $\mathbb{V}$ and let s be the projection of v onto $\mathcal{S}$. Write

$$v = s + (v - s) = s + t$$

From the Projection Theorem,

$$t = (v - s) \perp \mathcal{S} \quad \Longrightarrow \quad t \in \mathcal{S}^\perp$$

Thus, $\mathbb{V} = \mathcal{S} + \mathcal{S}^\perp$. Moreover $\mathcal{S} \cap \mathcal{S}^\perp = \{0\}$ from the previous lemma. Therefore, the sum is a direct sum: $\mathbb{V} = \mathcal{S} \oplus \mathcal{S}^\perp$.

(b) Let $s \in \mathcal{S}$. Then, for all $t \in \mathcal{S}^\perp$, we have $\langle s, t\rangle = 0$. So, $s \in (\mathcal{S}^\perp)^\perp$, proving the containment $\mathcal{S} \subseteq (\mathcal{S}^\perp)^\perp$. Next, suppose $v \in (\mathcal{S}^\perp)^\perp$ or $v \perp \mathcal{S}^\perp$. From (c), we can write

$$v = s + t, \ s \in \mathcal{S}, t \in \mathcal{S}^\perp$$

Notice that $v \perp \mathcal{S}^\perp$ and $s \perp \mathcal{S}^\perp$. Then we have

$$\langle t, t\rangle = \langle v - s, t\rangle = \langle v, t\rangle - \langle s, t\rangle = 0 \quad \Longrightarrow \quad \|t\|^2 = 0, \text{ or } t = 0$$

Thus $v = s \in \mathcal{S}$, proving the containment $(\mathcal{S}^\perp)^\perp \subseteq \mathcal{S}$. □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Orthogonal complements in $\mathbb{R}^3$ are illustrated in Figure 12. $\mathcal{S}^\perp$ is the subspace perpendicular to $\mathcal{S}$. Every vector in $\mathbb{R}^3$ can be expressed uniquely as a linear combination of $\mathcal{S}$ and $\mathcal{S}^\perp$.

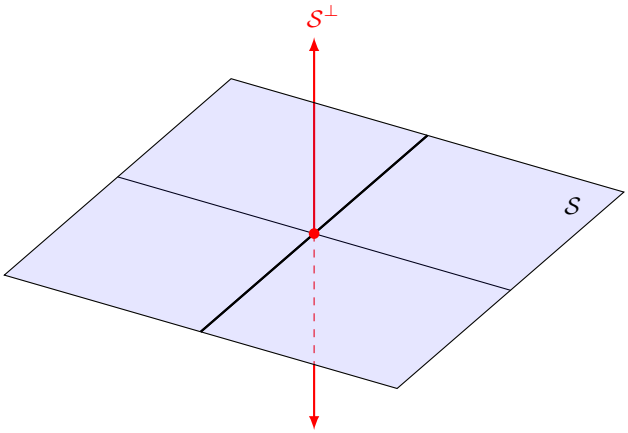

Figure 12: Orthogonal complements.

14 Completeness

Consider a normed vector space $\mathbb{V}$. Suppose the sequence of vectors

$$v_1,\ v_2,\ v_3,\ v_4,\ \cdots \quad \text{converges to } w$$

This means that v_k gets closer and closer to w as k increases, i.e.

$$\lim_{k\to\infty} \|v_k - w\| = 0$$

We can check if the sequence of vectors in $\mathbb{V}$ converges by testing convergence of the sequence of real numbers $\|v_k - w\|$. But this requires we know the candidate limit w.

Can we identify convergent sequences without explicitly identifying their limits?

For a wide range of normed vector spaces (which are called complete normed vector spaces) a convergent sequence is synonymous with a *Cauchy sequence*. As you will see in the definition below, checking if a sequence is Cauchy doesn't require knowledge of a limit.

Definition 45. Let $\mathbb{V}$ be a normed vector space and let $\{v_k\}$ be a sequence of vectors from $\mathbb{V}$. The sequence is called a Cauchy sequence if $\|v_m - v_n\| \to 0$ as $n, m \to \infty$. In other words, given any $\epsilon > 0$ there exists N_ϵ such that for all $m, n > N_\epsilon$,

$$\|v_m - v_n\| < \epsilon$$

Every convergent sequence is Cauchy. To see this suppose $v_k \to w$. Then

$$\|v_m - v_n\| = \|v_m - w + w - v_n\| \leq \|v_m - w\| + \|v_n - w\| \to 0.$$

A normed vector space in which the converse is also true (every Cauchy sequence is convergent) is called complete. Thus, in complete spaces, convergent and Cauchy sequences are synonymous and we can identify convergent sequences without knowing their limits.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Every finite dimensional vector space is complete. Therefore, to find a Cauchy sequence that does not converge, we need to consider infinite dimensional spaces. Here is one: consider the vector space $\mathbb{P}[0,1]$ of polynomials in the variable $t \in [0,1]$. The sequence of polynomials

$$1,\ 1+t,\ 1+t+\frac{t^2}{2!},\ 1+t+\frac{t^2}{2!}+\frac{t^3}{3!},\cdots$$

converges to

$$e^t = 1+t+\frac{t^2}{2!}+\frac{t^3}{3!}+\frac{t^4}{4!}+\cdots$$

But the limit is an exponential function which is not a polynomial. It is not back in the space $\mathbb{P}[0,1]$. This means that the sequence above is not convergent. On the other hand it is a Cauchy sequence (you can show this by using an appropriate norm, such as $\|p\| = \max_{t\in[0,1]} |p(t)|$). Therefore, $\mathbb{P}$ is not complete.

A complete normed vector space is called a Banach space.
A complete inner product space is called a Hilbert space.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Notation

$\mathcal{R}(A)$	range space of the operator A
$\mathcal{N}(A)$	null space of the operator A
A^T	the transpose of the matrix A
A^*	the adjoint of the operator A, or the complex-conjugate-transpose of the matrix A
$A > 0$	a positive-definite matrix
$A \geq 0$	a positive-semi-definite matrix
$\lambda_i(A)$	i^{th} eigenvalue of A
$\text{Spec}(A)$	the set of eigenvalues (or "spectrum") of A
$\rho(A)$	spectral radius of $A \quad = \max_i \lvert\lambda_i(A)\rvert$
$\sigma_i(A)$	i^{th} singular value of A (in descending order)
$\overline{\sigma}$	largest singular value
$\underline{\sigma}$	smallest nonzero singular value

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

B. Linear Operators

1 Linear Operators

Let $\mathbb{V}$ and $\mathbb{W}$ be vector spaces over the same field $\mathbb{F}$.

Definition 46. A linear operator is a mapping

$$\mathcal{L} : \mathbb{V} \longrightarrow \mathbb{W}$$

such that for all $v_1, v_2 \in \mathbb{V}$ and all scalars $\alpha \in \mathbb{F}$

(a) $\mathcal{L}(v_1 + v_2) = \mathcal{L}(v_1) + \mathcal{L}(v_2)$ (additivity)
(b) $\mathcal{L}(\alpha v_1) = \alpha \mathcal{L}(v_1)$ (homogeneity) □

Example 47. The following operators are linear:

(a) $\mathbb{V} = \mathbb{R}^n$, $\mathbb{W} = \mathbb{R}^m$, $\mathcal{L}(x) = Ax$, where $A \in \mathbb{R}^{m \times n}$
(b) $\mathcal{L} : \mathcal{C}(-\infty, \infty) \to \mathbb{R} : f(t) \to f(0)$.
Recall that $\mathcal{L} : \mathcal{C}(-\infty, \infty)$ is the vector space of continuous functions on $(-\infty, \infty)$.
(c) Suppose $h(t) \in L_2[a, b]$ and consider

$$\mathcal{L} : L_2[a, b] \to \mathbb{R} : f(t) \to \int_a^b h(t) f(t) dt$$

This operator is well-defined, i.e. $\mathcal{L}(f)$ is always a real number, because $f \in L_2[a, b]$. To see this, we use the Cauchy-Schwartz inequality:

$$|\mathcal{L}(f)| = \left| \int_a^b h(t) f(t) dt \right| = |\langle h, f \rangle| \leq \|h\| \|f\| < \infty$$

(d) Let $\mathbb{P}^N$ denote the vector space of polynomials of degree $\leq N$ with real coefficients:

$$\mathbb{P}^N = \{p : p(t) = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + \cdots + \alpha_N t^N, \alpha_0, \alpha_1, \cdots, \alpha_N \in \mathbb{R}\}$$

and consider the differentiation operator $\mathcal{L} : \mathbb{P}^N \to \mathbb{P}^{N-1}$:

$$\mathcal{L}(p)(t) = \frac{dp(t)}{dt}$$

(e) Let $A, B, X \in \mathbb{R}^{n \times n}$ and consider the Sylvester operator

$$\mathcal{L} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n} : X \to AX + XB$$ □

Example 48. The following operators are not linear:

(a) $\mathbb{V} = \mathbb{R}^n$, $\mathbb{W} = \mathbb{R}^m$, $\mathcal{L}(x) = Ax$, where $b \neq 0$
(b) Let $A, X \in \mathbb{R}^{n \times n}$ and consider

$$\mathcal{L} : \mathbb{R}^{n \times n} \to \mathbb{R}^{n \times n} : X \to XAX + X$$ □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

2 Matrix Representation of Linear Operators

Let $\mathbb{V}$ and $\mathbb{W}$ be finite dimensional vector spaces over a field $\mathbb{F}$, with $\dim(\mathbb{V}) = n$, $\dim(\mathbb{W}) = m$, and suppose $\mathcal{L} : \mathbb{V} \to \mathbb{W}$ is linear. Given a basis $\{v_1, v_2, \cdots, v_n\}$ for $\mathbb{V}$ and $\{w_1, w_2, \cdots, w_n\}$ for $\mathbb{W}$, we can construct an $m \times n$ matrix L (with entries in $\mathbb{F}$) such that

$$v = \sum_{j=1}^{n} \alpha_j v_j \quad \Longrightarrow \quad \mathcal{L}(v) = \sum_{i=1}^{m} \beta_i w_i$$

where

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = L \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

This means that we can turn the action of the operator into a familiar matrix-vector product.

How do we construct L? For $j = 1, \ldots, n$, pick the basis vector v_j, write $\mathcal{L}(v_j)$ as a linear combination of the basis vectors $w_1, w_2, \cdots, w_n$, and store the coefficient of w_i as ℓ_{ij}. Then $L = [\ell_{ij}]$. Thus, the entries of L are selected such that

$$\mathcal{L}(v_j) = \sum_{i=1}^{m} \ell_{ij} w_i, \quad j = 1, \cdots, n$$

To see why this construction works, note that

$$v = \sum_{j=1}^{n} \alpha_j v_j \quad \Longrightarrow \quad \mathcal{L}(v) = \sum_{j=1}^{n} \alpha_j \mathcal{L}(v_j) \quad \Longrightarrow \quad \mathcal{L}(v) = \sum_{j=1}^{n} \alpha_j \sum_{i=1}^{m} \ell_{ij} w_i$$

where the first implication is due to the linearity of $\mathcal{L}$. Rearranging the double summation on the right, we get

$$\mathcal{L}(v) = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \ell_{ij} \alpha_j \right) w_i$$

That is, $\mathcal{L}(v) = \sum_{i=1}^{m} \beta_i w_i$, where $\beta_i = \sum_{j=1}^{n} \ell_{ij} \alpha_j$, which means

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_m \end{bmatrix} = L \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix}$$

Note that the matrix L depends on the choice of bases for $\mathbb{V}$ and $\mathbb{W}$. Therefore, the matrix representing the same operator $\mathcal{L}$ may look different depending on what bases we choose.

Example 49. (Rotation operators)

Consider the vector space $\mathbb{R}^2$ and let $\mathcal{L}$ be the operator that rotates a given vector by θ counter-clockwise. Choose the standard basis vectors e_1, e_2 for $\mathbb{V} = \mathbb{W} = \mathbb{R}^2$, and follow the construction above. First consider the action of $\mathcal{L}$ on e_1: we get a new vector that we can write as

$$\mathcal{L}(e_1) = \underbrace{\cos\theta}_{\ell_{11}} e_1 + \underbrace{\sin\theta}_{\ell_{21}} e_2$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Similarly,

$$\mathcal{L}(e_2) = \underbrace{-\sin\theta}_{\ell_{12}} e_1 + \underbrace{\cos\theta}_{\ell_{22}} e_2$$

Thus,

$$L = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}.$$

Repeat this construction using a different basis for $\mathbb{R}^2$. You should get a different L.

Example 50. What is a matrix representation for the differentiation operator $\mathcal{L} : \mathbb{P}^N \to \mathbb{P}^{N-1}$:

$$\mathcal{L}(p)(t) = \frac{dp(t)}{dt}$$

discussed above? Choose the basis $\{1, t, t^2, \cdots, t^N\}$ for $\mathbb{P}^N$ and $\{1, t, t^2, \cdots, t^{N-1}\}$ for $\mathbb{P}^{N-1}$. That is, $v_i(t) = w_i(t) = t^{i-1}$. Then,

$$\begin{aligned}
\mathcal{L}(v_1)(t) = \frac{d1}{dt} = 0 &= 0w_1(t) + 0w_2(t) + 0w_3(t) + \cdots + 0w_N(t) \\
\mathcal{L}(v_2)(t) = \frac{dt}{dt} = t &= 1w_1(t) + 0w_2(t) + 0w_3(t) + \cdots + 0w_N(t) \\
\mathcal{L}(v_3)(t) = \frac{dt^2}{dt} = 2t &= 0w_1(t) + 2w_2(t) + 0w_3(t) + \cdots + 0w_N(t) \\
&\vdots \\
\mathcal{L}(v_{N+1})(t) = \frac{dt^N}{dt} = Nt^{N-1} &= 0w_1(t) + 0w_2(t) + 0w_3(t) + \cdots + Nw_N(t)
\end{aligned}$$

Collecting the coefficients above, we get

$$L = \begin{bmatrix}
0 & 1 & 0 & \cdots & \cdots & 0 \\
0 & 0 & 2 & 0 & \cdots & 0 \\
\vdots & & \ddots & \ddots & \ddots & \vdots \\
0 & \cdots & \cdots & 0 & N-1 & 0 \\
0 & \cdots & \cdots & \cdots & 0 & N
\end{bmatrix}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

C. Range and Null Spaces

1 Range and Null Spaces

Definition 51. The range space $\mathcal{R}(\mathcal{L})$ of a linear operator $\mathcal{L}$ is the set

$$\mathcal{R}(\mathcal{L}) = \{\mathcal{L}(v) : v \in \mathbb{V}\}$$

that is, the set of all w such that $w = \mathcal{L}(v)$ for some $v \in \mathbb{V}$.

The null space $\mathcal{N}(\mathcal{L})$ is the set

$$\mathcal{N}(\mathcal{L}) = \{v \in \mathbb{V} : \ \mathcal{L}(v) = 0\}$$

Show that $\mathcal{R}(\mathcal{L})$ and $\mathcal{N}(\mathcal{L})$ are subspaces of $\mathbb{W}$ and $\mathbb{V}$ respectively.

For a matrix $A \in \mathbb{C}^{m\times n}$, $\mathcal{R}(\mathcal{L})$ and $\mathcal{N}(\mathcal{L})$ refer implicitly to the linear transformation $\mathcal{L}(x) = Ax$ from $\mathbb{C}^n$ to $\mathbb{C}^m$. In this case, $\mathcal{R}(A)$ is simply the set of all linear combinations of the columns of A and $\mathcal{N}(A)$ is the set of vectors $x \in \mathbb{C}^n$ such that $Ax = 0$.

In the following theorem we collate some very important properties of range and null spaces. We will make very frequent use of these properties.

Theorem 52. *Let $A \in \mathbb{C}^{m\times n}$. Then*

(a) $\mathcal{R}^{\perp}(A) = \mathcal{N}(A^*)$
(b) $\mathbb{C}^m = \mathcal{R}(A) \oplus \mathcal{N}(A^*)$
(c) $\mathcal{N}(A^*A) = \mathcal{N}(A)$
(d) $\mathcal{R}(AA^*) = \mathcal{R}(A)$

Proof.

(a) $v \in \mathcal{N}(A^*) \iff A^*v = 0 \iff v^*A = 0 \iff v^*Ax = 0$ for all x
$\iff v \perp Ax$ for all $x \iff v \perp \mathcal{R}(A) \iff v \in \mathcal{R}(A)^{\perp}$

(b) For any subspace $\mathcal{S}$ of $\mathbb{C}^m$, we can write

$$\mathbb{C}^m = \mathcal{S} \oplus \mathcal{S}^{\perp}$$

Notice that $\mathcal{R}(A)$ is a subspace of $\mathbb{C}^m$. Then, we have (using (a))

$$\mathbb{C}^m = \mathcal{R}(A) \oplus \mathcal{R}^{\perp}(A) = \mathcal{R}(A) \oplus \mathcal{N}(A^*)$$

(c) We first show the containment $\mathcal{N}(A) \subseteq \mathcal{N}(A^*A)$. For this, we have

$$v \in \mathcal{N}(A) \implies Av = 0 \implies A^*Av = 0 \implies v \in \mathcal{N}(A^*A)$$

We next show the reverse containment $\mathcal{N}(A^*A) \subseteq \mathcal{N}(A)$ as

$$\begin{aligned} v \in \mathcal{N}(A^*A) &\implies A^*Av = 0 \implies v^*A^*Av = 0 \\ &\implies \|Av\| = 0 \implies Av = 0 \implies v \in \mathcal{N}(A) \end{aligned}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

(d) It follows from part (c) that $\mathcal{N}(AA^*) = \mathcal{N}(A^*)$ (just replace A with A^* in part (c)). Then, $\mathcal{N}(A^*)^\perp = \mathcal{N}(AA^*)^\perp$ and, from part (a),

$$\mathcal{R}(\mathcal{A}) = \mathcal{R}((\mathcal{A}\mathcal{A}^*)^*) = \mathcal{R}(\mathcal{A}\mathcal{A}^*)$$

□

2 Rank and Nullity

Definition 53. The rank of a matrix A is the dimension of $\mathcal{R}(A)$.

The nullity of a matrix A is the dimension of $\mathcal{N}(A)$.

Theorem 54. *Let $A \in \mathbb{C}^{m\times n}$ and $B \in \mathbb{C}^{n\times r}$. Then*

(a) $rank\ (A) = rank\ (A^*)$
(b) $rank\ (A) \leq \min\{m, n\}$
(c) $rank\ (A) + nullity\ (A^*) = m$
(d) $rank\ (A^*) + nullity\ (A) = n$
(e) $rank\ (A) = rank\ (AA^*) = rank\ (A^*A)$
(f) (Sylvester's inequality)

$$rank\ (A) + rank\ (B) - n \leq rank\ (AB) \leq \min\{rank\ (A), rank\ (B)\}$$

Multiplication by invertible matrices preserves rank. More precisely, we have the following:

Theorem 55. *Let $A \in \mathbb{C}^{m\times n}$ and $C \in \mathbb{C}^{n\times n}$ and suppose C is invertible. Then,*

$$\mathcal{R}(A) = \mathcal{R}(AC) \text{ and thus } rank\ (A) = rank\ (AC)$$

Proof. We first prove that $\mathcal{R}(A) \subseteq \mathcal{R}(AC)$. For this,

$$v \in \mathcal{R}(A) \implies v = Ax \implies v = ACC^{-1}x \implies v \in \mathcal{R}(AC)$$

To show the reverse containment, we write

$$v \in \mathcal{R}(AC) \implies v = ACx \implies v = Ay \text{ where } y = Cx \implies v \in \mathcal{R}(A)$$

proving the claim. □

3 Invertible Matrices

Definition 56. Let $A \in \mathbb{C}^{n\times n}$. The matrix A is called invertible or nonsingular if there exists a matrix $B \in \mathbb{C}^{n\times n}$ such that $AB = I$ □

Theorem 57. *Let $A \in \mathbb{R}^{n\times n}$. The following are equivalent*

(a) *A is invertible*
(b) *$\mathcal{R}(A) = \mathbb{R}^n$ i.e. $rank\ (A) = n$*
(c) *$\mathcal{N}(A) = \{0\}$ i.e. $nullity(A) = 0$*

4 Linear Systems of Equations

Theorem 58. *Let $A \in \mathbb{R}^{m\times n}, b \in \mathbb{R}^m$. Then,*

(a) *$Ax = b$ is solvable for $x \in \mathbb{R}^n$ if and only if $b \in \mathcal{R}(A)$*
(b) *Suppose $Ax = b$ is solvable for some b. This solution is unique if and only if $\mathcal{N}(A) = \{0\}$*
(c) *$Ax = b$ has a unique solution for every b if and only if A is square and invertible.*

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

D. Eigenvalues and Eigenvectors

1 Eigenvectors and Eigenvalues

Definition 59. Let $A \in \mathbb{C}^{n\times n}$. The characteristic polynomial of A is written as

$$\chi_A(s) = \det(sI - A) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s^1 + \alpha_0 s^0$$

The roots of $\chi_A(s)$, of which there are n, are called the eigenvalues of A.
The set of eigenvalues of A including multiplicity is called the Spectrum of A written Spec A.

Lemma 60. *The eigenvalues of $A \in \mathbb{C}^{n\times n}$ are continuous functions of the entries of A.*

Lemma 61. *Let $A \in \mathbb{C}^{n\times n}$ and λ be an eigenvalue of A. Then, there exists a vector $v \neq 0$ such that $Av = \lambda v$. This vector v is called an* eigenvector *of A with associated eigenvalue λ.*

Proof. Since λ is an eigenvalue of A, we have

$$\det(\lambda I - A) = \chi_A(\lambda) = 0$$

Thus the matrix $(\lambda I - A)$ is not invertible. Consequently, its null space is not trivial, i.e.

$$\exists v \neq 0 \;\text{ such that }\; (\lambda I - A)v = 0$$

proving the claim. □

It is evident that any nonzero multiple of v is also an eigenvector corresponding to the same eigenvalue. We regard all (nonzero) scalar multiples of an eigenvector as being the same eigenvector.

Let $A \in \mathbb{C}^{n\times n}$. If all the eigenvalues of A are distinct, we can clearly find one eigenvector for each eigenvalue. it turns out (we prove this below) that these eigenvectors are linearly independent. If, however, A has repeated eigenvalues it may or may not have n linearly independent eigenvectors.

Definition 62. A matrix $A \in \mathbb{C}^{n\times n}$ is called simple if it has distinct eigenvalues. A matrix A is called semi-simple if it has n linearly independent eigenvectors. □

2 The Simple Case

In the special case where A is simple, we have the following key result.

Theorem 63. *Let $A \in \mathbb{C}^{n\times n}$ be simple with eigenvalues λ_i and corresponding eigenvectors v_i, for $i = 1, \cdots, n$. Then*

(a) The eigenvectors form a basis for $\mathbb{C}^n$.

(b) Define the (nonsingular) matrix T and the matrix Λ by

$$T = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \quad , \quad \Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Then, $T^{-1}AT = \Lambda$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Proof. (a) We will use contradiction. Suppose $v_1, \cdots, v_n$ are linearly dependent. Then one of these vectors can be expressed as a linear combination of the remaining vectors. Without loss of generality assume the first vector is a linear combination of the others:

$$v_1 = \sum_{k=2}^{n} \alpha_k v_k$$

Define the matrix

$$M = (\lambda_2 I - A)(\lambda_3 I - A) \cdots (\lambda_n I - A)$$

Note that for any pair (λ, v) satisfying $Av = \lambda v$,

$$\begin{aligned} Mv &= (\lambda_2 I - A)(\lambda_3 I - A) \cdots (\lambda_{n-1} I - A)(\lambda_n - \lambda)v \\ &= (\lambda_2 I - A)(\lambda_3 I - A) \cdots (\lambda_{n-2} I - A)(\lambda_{n-1} - \lambda)(\lambda_n - \lambda)v \\ &= (\lambda_2 - \lambda)(\lambda_3 - \lambda) \cdots (\lambda_{n-2} - \lambda)v \end{aligned}$$

Using the fact that A has distinct eigenvalues, we have

$$Mv_1 = (\lambda_2 - \lambda_1)(\lambda_3 - \lambda_1) \cdots (\lambda_{n-2} - \lambda_1)v_1 \neq 0$$

For $k = 2, 3, \cdots, n$, we have

$$Mv_k = (\lambda_2 - \lambda_k)(\lambda_3 - \lambda_k) \cdots (\lambda_{n-2} - \lambda_k)v_k = 0$$

It now follows from (**??**) that

$$0 \neq Mv_1 = M\left(\sum_{k=2}^{n} \alpha_k v_k\right) = \sum_{k=2}^{n} \alpha_k M v_k = 0$$

This is a contradiction, proving that the eigenvectors $v_1, \cdots, v_n$ are linearly independent. Equivalently, the matrix

$$T = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \in \mathbb{C}^{n \times n}$$

is invertible.

(b) Observe that

$$AT = \begin{bmatrix} Av_1 & Av_2 & \cdots & Av_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1 & \lambda_2 v_2 & \cdots & \lambda_n v_n \end{bmatrix} = T\Lambda$$

Pre-multiplying by T^{-1} yields $T^{-1}AT = \Lambda$, completing the proof. □

The matrix $T^{-1}AT$ is the representation of the operator $\mathcal{A}(x) = Ax$ in the new basis $\{v_1, v_2, \ldots, v_n\}$ for C^n. The columns of the matrix T are these basis vectors and the transformation $T^{-1}AT$ is called a similarity transformation.

Theorem 63 states that simple matrices can be diagonalized using similarity transformations. Observe that the order in which the eigenvalues appear in Λ corresponds to the order in which its eigenvectors are placed in the matrix T.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Example 64. Let

$$A = \begin{bmatrix} 2 & -3 \\ 1 & -2 \end{bmatrix}$$

This matrix is simple and has eigenvalues at $1, -1$ with associated eigenvectors $\begin{bmatrix} 3 & 1 \end{bmatrix}'$ and $\begin{bmatrix} 1 & 1 \end{bmatrix}'$. It then follows that

$$T^{-1}AT = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \text{where} \quad T = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}$$

3 Jordan Form

The similarity transformation discussed above brings any semi-simple matrix A to a diagonal form. When A is not semi-simple, we can't bring A to a diagonal form, but we can bring it to what is called a Jordan form.

Let $\lambda_1, \cdots, \lambda_k$ be the eigenvalues of A with associated multiplicities $m_1, \cdots, m_k$. Observe that $\sum_i m_i = n$.

We have the following result:

Theorem 65. *There exists a nonsingular matrix $T \in \mathbb{C}^{n\times n}$ such that*

$$T^{-1}AT = \begin{bmatrix} J_1 & 0 & \cdots & 0 \\ 0 & J_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & J_k \end{bmatrix}$$

where the Jordan block $J_\ell \in \mathbb{C}^{\ell\times\ell}$ corresponding to the eigenvalue λ_ℓ has the structure

$$J_\ell = \begin{bmatrix} J_{1,\ell} & 0 & \cdots & 0 \\ 0 & J_{2,\ell} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & J_{r_\ell,\ell} \end{bmatrix}$$

and the Jordan sub-blocks are as

$$J_{i,\ell} = \begin{bmatrix} \lambda_\ell & 1 & \cdots & 0 & 0 \\ 0 & \lambda_\ell & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_\ell & 1 \\ 0 & 0 & \cdots & 0 & \lambda_\ell \end{bmatrix}$$

To determine the Jordan form of A we require the sizes and number of the various sub-blocks $J_{i,\ell}$ as well as the associated transformation matrix T. We will not delve into the details of this task, but we point out the following salient points:

(a) The eigenvalues of A, including multiplicity, appear on the diagonal of the Jordan form.

(b) If A is semi-simple, its Jordan form is diagonal.

(c) In the general case, the Jordan form may have $1's$ along portions of the super-diagonal.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The reason we do not emphasize Jordan forms is the numerical instability inherent in the computation of Jordan forms, illustrated in the example below. Instead we will later see Schur forms that serve a similar purpose and are computationally reliable.

Example 66. The Jordan form of a matrix A is *not* a *continuous* function of the entries of A. To see this consider:

$$A_\epsilon = \begin{bmatrix} 1+\epsilon & 1 \\ 0 & 1 \end{bmatrix}$$

The Jordan form of A_ϵ for $\epsilon \neq 0$ is

$$\begin{bmatrix} 1+\epsilon & 0 \\ 0 & 1 \end{bmatrix}$$

while the Jordan form for A_0 is

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

As a result one should steer clear of algorithms that require computation of the Jordan form of a matrix.

4 Determinant and Trace from Eigenvalues

Although we have advised against numerical computation of the Jordan form, it is conceptually useful to know that any matrix can be brought to Jordan form. For example, the Jordan form helps us derive the following useful formulas:

$$\det(A) = \lambda_1\lambda_2\cdots\lambda_n \qquad \mathrm{Tr}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

These follow because we know that an invertible matrix T brings $T^{-1}AT$ to Jordan form, which is upper triangular with eigenvalues on the diagonal. Thus,

$$\det(T^{-1}AT) = \lambda_1\lambda_2\cdots\lambda_n \qquad \mathrm{Tr}(T^{-1}AT) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$$

Since $\det(\mathrm{T}^{-1}\mathrm{AT}) = \det(\mathrm{TT}^{-1}\mathrm{A}) = \det(\mathrm{A})$ and $\mathrm{Tr}(\mathrm{T}^{-1}\mathrm{AT}) = \mathrm{Tr}(\mathrm{TT}^{-1}\mathrm{A}) = \mathrm{Tr}(\mathrm{A})$, we conclude $\det(A) = \lambda_1\lambda_2\cdots\lambda_n$ and $\mathrm{Tr}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$.

The formula for the determinant is particularly useful because the computation of the determinant suggested by the standard definition is awkward. Instead, we can simply multiply the eigenvalues (multiplicities included).

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

E. Functions of a Square Matrix

1 Functions of a Matrix

Let $A \in \mathbb{C}^{n\times n}$ and let $p(s) = \sum_{i=0}^{k} \alpha_i s^i$ be a polynomial. Then, we *define*

$$p(A) = \sum_{i=0}^{k} \alpha_i A^i \in \mathbb{C}^{n\times n}$$

where $A^0 = I$.

We can generalize this notion to arbitrary (analytic) functions as follows. Consider the Taylor series

$$f(s) = \sum_{i=0}^{\infty} \alpha_i s^i$$

and *assume* that this Taylor series converges on $Spec(A)$. Then, we *define*

$$f(A) = \sum_{i=0}^{\infty} \alpha_i A^i \in \mathbb{C}^{n\times n}$$

(it will happen that this defining Taylor series also converges).

Lemma 67. *Let $f(s), g(s)$ be arbitrary functions and let $h(s) = f(s)g(s)$. Then*

(a) $f(A)g(A) = g(A)f(A) = h(A)$

(b) $f(T^{-1}AT) = T^{-1}f(A)T$

(c) $f\left(\begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right) = \begin{bmatrix} f(A) & 0 \\ 0 & f(B) \end{bmatrix}$

2 Computing functions of a matrix

We begin with the case where $A \in \mathbb{C}^{n\times n}$ is simple. In this case, the Jordan form $J = T^{-1}AT$ of A is diagonal and may be readily computed. We can then employ properties (b) and (c) above to write

$$f(A) = Tf(J)T^{-1} = T \begin{bmatrix} f(\lambda_1) & 0 & \cdots & 0 \\ 0 & f(\lambda_2) & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & f(\lambda_n) \end{bmatrix} T^{-1}$$

Observe that $f(\lambda_i)$ are well-defined because $f(s)$ converges on $Spec(A)$. The reason we impose this requirement in defining $f(A)$ is now transparent.

Example 68. Let

$$A = \begin{bmatrix} 2 & -3 \\ 1 & -2 \end{bmatrix}$$

We compute A^{300}:

$$A^{300} = TJ^{300}T^{-1} = T \begin{bmatrix} 1^{300} & 0 \\ 0 & -1^{300} \end{bmatrix} T^{-1} = I$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

We now turn our attention to the general case. Again, we shall proceed via the Jordan canonical form. Using the same idea, we see that we need to be able to compute $f(J)$ for a general Jordan block

$$J = \begin{bmatrix} \lambda & 1 & \cdots & 0 & 0 \\ 0 & \lambda & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda & 1 \\ 0 & 0 & \cdots & 0 & \lambda \end{bmatrix}$$

It turns out that

$$f(J) = \begin{bmatrix} f(\lambda) & \frac{1}{1!}f^{(1)}(\lambda) & \cdots & \frac{1}{(k-2)!}f^{(k-2)}(\lambda) & \frac{1}{(k-1)!}f^{(k-1)}(\lambda) \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & f(\lambda) & \frac{1}{1!}f^{(1)}(\lambda) \\ 0 & 0 & \cdots & 0 & f(\lambda) \end{bmatrix}$$

Observe that the derivatives $f^{(i)}(\lambda)$ above exist. We are therefore in a position to compute (analytic) functions of an arbitrary matrix.

Example 69. Consider the matrix

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 2 \end{bmatrix}$$

Using the above result, we obtain

$$\cos(A) = \begin{bmatrix} \cos(2) & -\sin(2) \\ 0 & \cos(2) \end{bmatrix}$$

3 The Spectral Mapping Theorem

Theorem 70. *Let $A \in \mathbb{C}^{n\times n}$ and let $f(s)$ be an arbitrary analytic function.*

(a) Suppose A has eigenvalues $\{\lambda_1, \cdots, \lambda_n\}$. Then, the eigenvalues of $f(A)$ are $\{f(\lambda_1), \cdots, f(\lambda_n)\}$.

(b) Let v be an eigenvector of A with associated eigenvalue λ. Then v is also an eigenvector of $f(A)$ with associated eigenvalue $f(\lambda)$.

4 The Cayley-Hamilton Theorem.

We will make use of the following result.

Theorem 71. *Let $A \in \mathbb{C}^{n\times n}$ and let*

$$\chi(s) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s^1 + \alpha_0$$

be the characteristic polynomial of A. Then,

$$\chi(A) = 0$$

As a consequence of this theorem, it is evident that A^n is a linear combination of the set of matrices $\mathbb{S} = \left\{I, A, \cdots, A^{n-1}\right\}$. By an inductive argument it follows that any power of A, and therefore *any function* of A is expressible as a linear combination in $\mathbb{S}$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5 Matrix Exponentials

Matrix exponentials are particularly important and arise in connection with systems of coupled linear ordinary differential equations. Since the Taylor series

$$e^{st} = 1 + st + \frac{s^2t^2}{2!} + \frac{s^3t^3}{3!} + \cdots$$

converges everywhere, we can define the exponential of any matrix $A \in \mathbb{C}^{n\times n}$ by

$$e^{At} = I + At + \frac{A^2t^2}{2!} + \frac{A^3t^3}{3!} + \cdots$$

6 Properties of Matrix Exponentials

Theorem 72. *(a)* $e^0 = I$

(b) If $AB = BA$ *then* $e^{(A+B)t} = e^{At}e^{Bt} = e^{Bt}e^{At}$

(c) $\det\left[\ e^{At}\ \right] = e^{\text{trace}At}$

(d) e^{At} *is always invertible and* $\left[\ e^{At}\ \right]^{-1} = e^{-At}$

7 Computing Matrix Exponentials

How could we compute matrix exponentials? Using the Taylor series definition is hopeless. A much easier way is to go through the Jordan form.

We will focus on the simple case, i.e. all the eigenvalues of A are distinct. We can diagonalize A as

$$A = T\Lambda T^{-1}, \quad \text{where } \Lambda = \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix}$$

As we have shown,

$$e^{At} = Te^{\Lambda t}T^{-1}, \quad \text{where } e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix}$$

Example 73. Let A be as in item Example 64. We compute e^{At}:

$$e^{At} = Te^{Jt}T^{-1} = T\begin{bmatrix} e^t & 0 \\ 0 & e^{-t} \end{bmatrix}T^{-1} = \begin{bmatrix} 1.5e^t - 0.5e^{-t} & -1.5e^t + 1.5e^{-t} \\ 0.5e^t - 0.5e^{-t} & -0.5e^t + 1.5e^{-t} \end{bmatrix} \qquad \square$$

Example 74. Consider the matrix

$$A = \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}$$

You should be able to check that

$$e^{At} = e^{\sigma t}\begin{bmatrix} \cos(\omega t) & \sin(\omega t) \\ -\sin(\omega t) & \cos(\omega t) \end{bmatrix} \qquad \square$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

F. Hermitian and Positive Definite Matrices

1 What is in this Section?

We first introduce Unitary matrices. These are square matrices whose columns form an orthonormal basis.
We then study Hermitian matrices. These are also square matrices. Hermitian matrices generalize the concept of symmetric matrices. They are extremely important in Linear Algebra. Hermitian matrices enjoy very nice numerical properties. They have real eigenvalues, and these can be computed very reliably. Hermitian matrices appear in many applications. In Quantum Mechanics, Hermitian matrices represent physical quantities like energy, linear momentum, and angular momentum.

2 Unitary Matrices

Definition 75. A matrix $U \in \mathbb{C}^{n\times n}$ is called unitary if $U^*U = I = UU^*$.

A real unitary matrix is called an orthogonal matrix.

Example 76. The matrix

$$U = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

is unitary. Since all the entries of U are real, we say that U is orthogonal. It is easy to verify that U is unitary:

$$U^*U = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

We can also think of the columns of U as being an orthonormal basis for $\mathbb{R}^2$. This basis is just the standard basis rotated counter-clockwise by θ degrees.

Theorem 77. *Properties of Unitary Matrices Let $U \in \mathbb{C}^{n\times n}$ be unitary. Then,*

(a) $\|Ux\| = \|x\|$
(b) $\langle Ux, Uy\rangle = \langle x, y\rangle$
(c) $U^{-1} = U^*$
(d) The columns of U form an orthonormal basis of $\mathbb{C}^n$. □

Proof.

(a) $\|Ux\|^2 = x^*U^*Ux = x^*x = \|x\|^2$
(b) $\langle Ux, Uy\rangle = x^*U^*Uy = x^*y = \langle x, y\rangle$
(c) This is immediate.
(d) This follows from (a) and (b) on observing that the i^{th} column of U can be written as Ue_i where e_i is the i^{th} standard basis vector in $\mathbb{C}^n$. □

3 The Schur Form

Unlike the Jordan form, Schur form is extremely reliable computationally. It transforms any square matrix into an upper-triangular matrix T using unitary matrices. The eigenvalues of A appear on the diagonal of T.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Theorem 78. *Let $A \in \mathbb{C}^{n\times n}$. There exists U unitary such that*

$$A = UTU^* \qquad T = \begin{bmatrix} \lambda_1 & * & \cdots & * & * \\ 0 & \lambda_2 & \cdots & * & * \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \lambda_{n-1} & * \\ 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

Proof. We use induction on n. The result clearly holds for $n = 1$. Suppose the result holds for $k = 1, \cdots, n-1$. Let λ_1 be any eigenvalue of A. Normalize its eigenvector:

$$Av_1 = \lambda_1 v_1, \ \|v_1\| = 1$$

Extend v_1 by $\{v_2, v_3, \cdots, v_n\}$ to form an orthonormal basis for $\mathbb{C}^n$. Notice that

$$V = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

is unitary. Then

$$\begin{aligned} AV &= A\begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1 & * & \cdots & * \end{bmatrix} \\ &= \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & * & \cdots & * \\ 0 & * & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ 0 & * & \cdots & \lambda_n \end{bmatrix} = V \begin{bmatrix} \lambda_1 & * \\ 0 & A_{22} \end{bmatrix} \end{aligned}$$

This implies

$$A = V \begin{bmatrix} \lambda_1 & * \\ 0 & A_{22} \end{bmatrix} V^*$$

Since A_{22} is $(n-1)\times(n-1)$, the result holds for A_{22}. We can then write

$$A_{22} = WT_{22}W^*$$

where W is unitary and T_{22} is upper-triangular. It follows that

$$A = \underbrace{V \begin{bmatrix} 1 & 0 \\ 0 & W \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \lambda_1 & * \\ 0 & T_{22} \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} 1 & 0 \\ 0 & W^* \end{bmatrix} V^*}_{U^*}$$

Notice that U is unitary, and T is upper-triangular, proving the claim.

4 Hermitian Matrices

Definition 79. A matrix $H \in \mathbb{C}^{n\times n}$ is called Hermitian if $H = H^*$.

Real symmetric matrices are, in particular, Hermitian. The matrix

$$A = \begin{bmatrix} 1 & j \\ j & 1 \end{bmatrix} \neq \begin{bmatrix} 1 & -j \\ -j & 1 \end{bmatrix} = A^*$$

is symmetric but not Hermitian.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5 Main Results on Hermitian Matrices

We will now prove several results regarding Hermitian matrices. These results also hold almost *verbatim* for real symmetric matrices. The central result is that all the eigenvalues of a Hermitian matrix are real, and the eigenvectors form an orthonormal set.

Theorem 80. *Let $H \in \mathbb{C}^{n\times n}$ be Hermitian.*

(a) The eigenvalues of H are real.

(b) H has a full set of eigenvectors, and these eigenvectors form an orthogonal set.

(c) Hermitian matrices can be diagonalized by unitary transformations, i.e. there exists a unitary matrix U such that

$$H = UDU^*, \quad \textit{where} \quad D = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

Here $\lambda_1, \cdots, \lambda_n$ are the (real) eigenvalues of H.

Proof.

(a) Let λ be any eigenvalue of H. We can find an eigenvector corresponding to λ and write

$$Hv = \lambda v, \text{ with } v \neq 0$$

Taking the complex-conjugate transpose and recognizing that H is Hermitian, we get

$$v^*H^* = v^*H = \overline{\lambda}v^*$$

It follows that

$$v^*Hv = \lambda v^*v = \overline{\lambda}v^*v \implies (\lambda - \overline{\lambda})v^*v = 0$$

Since $v^*v = \|v\|^2 \neq 0$, we have $\lambda = \overline{\lambda}$ proving that λ is real.

(b) and (c). We will prove this by induction on n. The result is clearly true is H is 1×1. Suppose it is true for all Hermitian matrices that are $N \times N$ or smaller. Consider an $(N+1)\times(N+1)$ Hermitian matrix H. Select any eigenvalue λ of H. Let v be the corresponding eigenvector. Normalize v to have unit length. We can write

$$Hv = \lambda v, \|v\| = 1$$

Extend v by $v_1, \cdots v_N$ to form an orthonormal basis for $\mathbb{C}^{N+1}$. This can be done, for example, using the Gram-Schmidt procedure. As a result, the matrix

$$V = \begin{bmatrix} v & v_1 & v_2 & \cdots & v_N \end{bmatrix} \in \mathbb{C}^{(N+1)\times(N+1)} \text{ is unitary.}$$

Next observe that

$$HV = \begin{bmatrix} Hv & Hv_1 & \cdots & Hv_N \end{bmatrix} = V \begin{bmatrix} \lambda & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix} = V \begin{bmatrix} \lambda & w^* \\ 0 & G \end{bmatrix}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Therefore

$$V^*HV = \begin{bmatrix} \lambda & w^* \\ 0 & G \end{bmatrix} \quad \text{or} \quad H = V \begin{bmatrix} \lambda & w^* \\ 0 & G \end{bmatrix} V^*$$

Since H is Hermitian, we get

$$\begin{bmatrix} \lambda & 0 \\ w & G^* \end{bmatrix} = (V^*HV)^* = V^*HV = \begin{bmatrix} \lambda & w^* \\ 0 & G \end{bmatrix}$$

This forces $w = 0$ and $G = G^*$. Since $G \in \mathbb{C}^{N\times N}$, the induction hypothesis lets us write

$$G = ZD_GZ^*$$

where Z is unitary and D_G is a diagonal matrix of the (real) eigenvalues of G. Putting all this together, we get

$$H = V \begin{bmatrix} \lambda & 0 \\ 0 & G \end{bmatrix} V^* = V \begin{bmatrix} 1 & 0 \\ 0 & Z \end{bmatrix} \begin{bmatrix} \lambda & 0 \\ 0 & D_G \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & Z^* \end{bmatrix} V^* = UDU^*$$

where

$$D = \begin{bmatrix} \lambda & 0 \\ 0 & D_G \end{bmatrix} \text{ is diagonal, and } U = V \begin{bmatrix} 1 & 0 \\ 0 & Z \end{bmatrix}$$

Observe that U is unitary because

$$U^*U = \begin{bmatrix} 1 & 0 \\ 0 & Z^* \end{bmatrix} V^*V \begin{bmatrix} 1 & 0 \\ 0 & Z \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & Z^*Z \end{bmatrix} = I$$

completing the induction.

6 Positive Definite Matrices

Definition 81. A matrix $P \in \mathbb{C}^{n\times n}$ is called positive-definite, written $P \succ 0$, if P is Hermitian and further,

$$v^*Pv > 0\ , \quad \text{for all } \ 0 \neq v \in \mathbb{C}^n$$

A matrix $P \in \mathbb{C}^{n\times n}$ is called positive-semi-definite, written $P \succeq 0$, if P is Hermitian and further,

$$v^*Pv \geq 0\ , \quad \text{for all } \ v \in \mathbb{C}^n$$ □

Analogous are the notions of negative- and negative-semi- definite matrices.
Suppose A and B are Hermitian. We will write $A \succ B$ to mean $A - B \succ 0$.

Example 82. A positive-definite matrix can have negative entries. The following matrix is positive-definite:

$$P = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}$$

Clearly P is Hermitian. So see that $P \succ 0$, observe that for all x, y

$$v^*Pv = \begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 2x^2 + 2y^2 - 2xy = x^2 + y^2 + (x+y)^2 \geq 0$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Also, $v^*Pv = 0 \iff x = y = 0$.
Conversely, just because a matrix A has all positive entries, it does not mean that $A \succ 0$. For example, the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix}$$

is not positive-definite. This is apparent when we calculate

$$v^*Av = \begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ -1 \end{bmatrix} = -2$$ □

Theorem 83. *Let $P \in \mathbb{C}^{n\times n}$ be Hermitian.*

(a) $P \succeq 0 \iff$ *all the eigenvalues of* P *are* ≥ 0.
(b) $P \succ 0 \iff$ *all the eigenvalues of* P *are* > 0. □

Proof. We prove (a). Let $P \succeq 0$. Since $P = P^*$, we can write

$$P = U^*\Lambda U, \quad \Lambda = \text{diag}\,\{\lambda_1, \cdots, \lambda_n\}$$

If Spec $(P) \geq 0$, then for any $v \in \mathbb{C}^n$ we have

$$v^*Pv = v^*U^*\Lambda Uv = w^*\Lambda w = \sum \lambda_k |w_k|^2 \geq 0$$

Conversely suppose $P \succeq 0$. Let $w = U^*e_k$ where e_k is the k^{th} standard basis vector. Then

$$0 \leq w^*Pw = e_kUU^*\Lambda UU^*e_k = \lambda_k$$

proving the claim. The proof of (b) is very similar.

Theorem 84. *Let $0 \prec P \in \mathbb{C}^{n\times n}$. Then*

(a) $\|x\|^2 = x^*Px$ *qualifies as a norm*
(b) $\langle x, y\rangle = x^*Py$ *qualifies as an inner-product*

Proof. To prove (b) we just check that x^*Py satisfies all the axioms required of an inner-product. Since $P \succ 0$, we have $x^*Px \geq 0$ and $x^*Px = 0$ if and only if $x = 0$. Linearity in y is easy to verify. Part (a) follows immediately on setting $x = y$.

7 Square-roots

Definition 85. Let $0 \preceq P \in \mathbb{C}^{n\times n}$. We can then write $P = UDU^*$ where U is unitary and D is a diagonal matrix. Define the square-root of P written $P^{\frac{1}{2}}$ by

$$P^{\frac{1}{2}} = UD^{\frac{1}{2}}U^*$$ □

where $D^{\frac{1}{2}}$ is the entry-wise square-root of the diagonal matrix D.

It follows that

$$P^{\frac{1}{2}}P^{\frac{1}{2}} = UD^{\frac{1}{2}}U^*UD^{\frac{1}{2}}U^* = UD^{\frac{1}{2}}D^{\frac{1}{2}}U^* = UDU^* = P$$

which is why we call $P^{\frac{1}{2}}$ the square-root of P.

It is evident that $P^{\frac{1}{2}}$ as defined above is Hermitian, and moreover $P^{\frac{1}{2}} \succeq 0$. Further, if $P \succ 0$, then $P^{\frac{1}{2}} \succ 0$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

G. The Singular Value Decomposition

1 The SVD

Theorem 86. *Let $M \in \mathbb{C}^{m\times n}$ with rank $(M) = r$. Then we can find unitary matrices $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ such that*

$$M = U\Sigma V^* = U \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix} V^*$$

where

$$\Sigma_1 = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix}$$

The real numbers $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are called the singular values of M and the representation above is called a singular-value decomposition (SVD) of M.

Although Σ is uniquely determined, U and V are not: *e.g.*, if we define $\tilde{U} = -U$ and $\tilde{V} = -V$, then $\tilde{U}$ and $\tilde{V}$ are also unitary and $M = \tilde{U}\Sigma\tilde{V}^*$.

When M is real we can choose U and V to be real, orthogonal matrices.

2 Four Subspaces

Theorem 87. *Let $M \in \mathbb{C}^{m\times n}$ with rank $(M) = r$ and let $M = U\Sigma V^*$ be a singular-value decomposition of M. Partition U and V as*

$$U = \begin{bmatrix} U_1 & U_2 \end{bmatrix}, \quad V = \begin{bmatrix} V_1 & V_2 \end{bmatrix}$$

where $U_1 \in \mathbb{C}^{m\times r}$ and $V_1 \in \mathbb{C}^{n\times r}$. Then

(a) The columns of U_1 and U_2 form orthonormal bases for $\mathcal{R}(M)$ and $\mathcal{N}(M^*)$ respectively.

(b) The columns of V_1 and V_2 form orthonormal bases for $\mathcal{R}(M^*)$ and $\mathcal{N}(M)$ respectively.

3 Alternative Forms of SVD

Using the partition above, we get the alternative form

$$M = U_1\Sigma_1 V_1^*.$$

If we denote by $u_1, \ldots, u_r$ the columns of U_1 and by $v_1, \ldots, v_r$ the columns of V_1, then

$$M = \sigma_1 u_1 v_1^* + \cdots + \sigma_r u_r v_r^*,$$

which is yet another representation of SVD.

4 Computing the singular-value decomposition

While definitely not the method of choice *vis-a-vis* numerical aspects, the following result provides an adequate method for determining the singular-value decomposition of a matrix.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Theorem 88. Let $M \in \mathbb{C}^{m\times n}$ with rank $(M) = r$. Let $\lambda_1, \cdots, \lambda_r$ be the nonzero eigenvalues of MM^*. These will be non-negative because $MM^* \succeq 0$. Also, from Theorem 80 it follows that there exists a unitary matrix $U \in \mathbb{C}^{m\times m}$ such that

$$MM^* = U \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} U^*$$

where $\Lambda = diag\,(\lambda_1, \cdots, \lambda_r)$. Then the singular values of M are $\lambda_1^{\frac{1}{2}}, \cdots, \lambda_r^{\frac{1}{2}}$ and a singular-value decomposition of M is

$$M = U \begin{bmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} V^*$$

where $V = [V_1, V_2]$ with V_1 defined as $V_1 = M^*U_1\Lambda^{-1/2}$ and V_2 selected as a matrix whose columns constitute an orthonormal basis for the null space of M.

Here U is already selected to be unitary. Convince yourself that the proposed V is also unitary and that the product

$$U \begin{bmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} V^* = U_1\Lambda^{\frac{1}{2}}V_1^*$$

indeed recovers M.

The following variant of the theorem above makes use of M^*M instead of MM^*:

Theorem 89. Let $M \in \mathbb{C}^{m\times n}$ with rank $(M) = r$. Let $\lambda_1, \cdots, \lambda_r$ be the nonzero eigenvalues of M^*M, which are non-negative because $M^*M \succeq 0$. From Theorem 80 there exists a unitary matrix $V \in \mathbb{C}^{n\times m}$ such that

$$M^*M = V \begin{bmatrix} \Lambda & 0 \\ 0 & 0 \end{bmatrix} V^*$$

where $\Lambda = diag\,(\lambda_1, \cdots, \lambda_r)$. Then the singular values of M are $\lambda_1^{\frac{1}{2}}, \cdots, \lambda_r^{\frac{1}{2}}$ and a singular-value decomposition of M is

$$M = U \begin{bmatrix} \Lambda^{\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix} V^*$$

where $U = [U_1, U_2]$ with $U_1 = MV_1\Lambda^{-1/2}$ and U_2 selected as a matrix whose columns constitute an orthonormal basis for the null space of M^*.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

H. Adjoints

1 Adjoints

We said earlier that we should think of matrices as representations of linear operators rather than just an array of numbers. Similarly, the transpose of a real matrix or the adjoint (conjugate transpose) of a complex matrix can be viewed as a matrix representation of what is called an adjoint operator.

Definition 90. Let $\mathcal{A} : \mathbb{V} \to \mathbb{W}$ be a linear operator, where $\mathbb{V}$ and $\mathbb{W}$ are inner-product spaces. The adjoint of $\mathcal{A}$, written $\mathcal{A}^*$, is the operator

$$\mathcal{A}^* : \mathbb{W} \to \mathbb{V}$$

defined by

$$\langle w, \mathcal{A}(v)\rangle = \langle \mathcal{A}^*(w), v\rangle \quad \text{for all} \quad v \in \mathbb{V},\ w \in \mathbb{W}$$

The inner product on the left is for the vector space $\mathbb{W}$ and the one on the right is for $\mathbb{V}$.

To see the connection to the adjoint of a complex matrix, consider the case where $\mathcal{A} : \mathbb{C}^n \to \mathbb{C}^m$ is defined by $\mathcal{A}(v) = Av$, $A \in \mathbb{C}^{m\times n}$. Then the adjoint operator $\mathcal{A}^*$ can be represented with the matrix A^*; that is $\mathcal{A}^*(w) = A^*w$. This follows by applying the definition above with the usual inner product $\langle x, y\rangle = x^*y$ for complex vectors:

$$\langle w, \mathcal{A}(v)\rangle = w^*Av = (A^*w)^*v = \langle \mathcal{A}^*(w), v\rangle$$

2 The Lyapunov operator

Let $A \in \mathbb{R}^{n\times n}$. and consider the Lyapunov operator $\mathcal{L}(X) = A^TX + XA$. We show that $\mathcal{L}^*(Y) = AY + YA^T$.

Note that

$$\mathcal{L} : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n} \quad \text{and} \quad \mathcal{L}^* : \mathbb{R}^{n\times n} \to \mathbb{R}^{n\times n}$$

We use the standard trace inner-product on $\mathbb{R}^{n\times n}$.

$$\langle Y, X\rangle = \text{Trace}(Y^TX)$$

Set $\mathcal{L}^*(Y) = Z$. Using the definition of adjoint,

$$\langle Y, \mathcal{L}(X)\rangle = \langle \mathcal{L}^*(Y), X\rangle \quad \Longrightarrow \quad \text{Trace}(Y^T(A^TX + XA)) = \text{Trace}(Z^TX)$$

for all $X, Y \in \mathbb{R}^{n\times n}$. Note that

$$\begin{aligned}
\text{Trace}(Y^T(A^TX + XA)) &= \text{Trace}(Y^TA^TX) + \text{Trace}(Y^TXA) \\
&= \text{Trace}(Y^TA^TX) + \text{Trace}(AY^TX) \\
&= \text{Trace}((Y^TA^T + AY^T)X) = \text{Trace}((AY + YA^T)^TX)
\end{aligned}$$

Thus, Z must be such that

$$\text{Trace}((AY + YA^T)^TX) = \text{Trace}(Z^TX)$$

for all X, Y. This is indeed the case with

$$Z = \mathcal{L}^*(Y) = AY + YA^T$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3 Operator from function space to Euclidean space

Let $\mathcal{A} : L_2^m[0,T] \to \mathbb{R}^n$ be defined as

$$\mathcal{A}(v) = \int_0^T G(t)v(t)dt$$

where $G(t) \in \mathbb{R}^{n\times m}$, $t \in [0,T]$. To find $\mathcal{A}^* : \mathbb{R}^n \to L_2^m[0,T]$ we use the definition

$$\langle w, \mathcal{A}(v)\rangle = \langle \mathcal{A}^*(w), v\rangle \quad \text{for all} \quad w \in \mathbb{R}^n,\ v \in L_2^m[0,T]$$

The inner product on the left is the inner product for $\mathbb{R}^n$, thus

$$\langle w, \mathcal{A}(v)\rangle = w^T\mathcal{A}(v) = w^T\int_0^T G(t)v(t)dt = \int_0^T w^TG(t)v(t)dt$$

The inner product on the right is for $L_2^m[0,T]$. If we denote $\mathcal{A}^*(w) = u$, it has the form

$$\langle \mathcal{A}^*(w), v\rangle = \langle u, v\rangle = \int_0^T u(t)^Tv(t)dt$$

Matching the two inner products, we get

$$\int_0^T w^TG(t)v(t)dt = \int_0^T u(t)^Tv(t)dt$$

For this to hold for all v, w, we need

$$u(t) = G(t)^Tw$$

Thus, for $w \in \mathbb{R}^n$, $\mathcal{A}^*(w)$ is a function whose value at t is

$$\mathcal{A}^*(w)(t) = G(t)^Tw$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Solution of state space equations

B. Stability

C. LTV Systems

D. Controllability

E. Observability

F. Modal Observability and Controllability Tests

G. Kalman decomposition

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Solution of State-Space Equations

1 Solution of State Space equations

Theorem 91. *Consider the LTI System*

$$\Sigma: \quad \begin{cases} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{cases} \qquad \textit{with initial condition} \quad x(0) = \xi$$

The closed-form solution of these differential equations is

$$\begin{aligned} x(t) &= e^{At}\xi + \int_0^t e^{A(t-\tau)}Bu(\tau)d\tau \\ y(t) &= Ce^{At}\xi + \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t) \end{aligned} \tag{8}$$

Proof. All we have to do is verify that the proposed solution satisfies the initial conditions and solves the state space differential equation. We check:

$$x(0) = e^{A0}\xi + \int_0^0 e^{A(t-\tau)}Bu(\tau)d\tau = \xi$$

Next using the Leibniz differentiation rule[1].

$$\begin{aligned} \frac{dx}{dt} &= Ae^{At}\xi + \frac{d}{dt}\left[\int_0^t e^{A(t-\tau)}Bu(\tau)d\tau\right] \\ &= Ae^{At}\xi + A\int_0^t e^{A(t-\tau)}Bu(\tau)d\tau + Bu(t) \\ &= Ax(t) + Bu(t) \end{aligned}$$

This general solution has two parts: the free response and the forced response.

The free response is the part of the solution due to initial conditions only with input $u = 0$:

$$\text{free response:} \quad y^{\text{free}}(t) = Ce^{At}\xi \tag{9}$$

The forced response is the part due to the input alone with initial conditions $\xi = 0$:

$$\text{forced response:} \quad y^{\text{forced}}(t) = \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t) \tag{10}$$

Remark 92. The D-term captures the immediate effect of the input u on the output y. Most often, $D = 0$ because physical systems do not respond immediately to the input.

[1] The Leibniz differentiation formula tells us how to take the derivative of a definite integral:

$$\frac{d}{dt}\int_{a(t)}^{b(t)} f(t,\tau)d\tau = \frac{db(t)}{dt}\cdot f(t,b(t)) - \frac{da(t)}{dt}\cdot f(t,a(t)) + \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}(t,\tau)d\tau$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

2 Free Response

The free response is the part of the solution due to initial conditions only with input $u = 0$:

$$y^{\text{free}}(t) = Ce^{At}\xi$$

In the case where A is semisimple, the matrix exponential e^{At} has the form

$$e^{At} = Te^{\Lambda t}T^{-1}, \quad \text{where} \quad e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix}$$

We see that the free response is a linear combination of $e^{\lambda_i t}$ where $\lambda_i \in \text{Spec } A$. We can understand the free response qualitatively by plotting $e^{\lambda t}$ for various eigenvalue locations λ.

We begin with real eigenvalues. The free response term $e^{\lambda t}$ is shown in Figure 13. Notice that

(a) there are no oscillations

(b) the response decays exponentially to zero if and only if Real $\lambda < 0$ with faster decay when λ is more negative

(c) the response grows exponentially when Real $\lambda > 0$ with faster growth when λ is more positive

(d) the response is constant when $\lambda = 0$

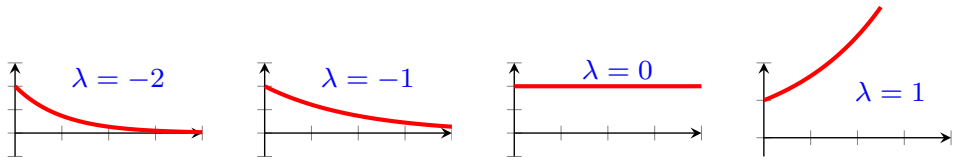

Figure 13: Free response terms for real eigenvalues.

We now turn to the case of complex eigenvalues. When $\lambda = \sigma \pm j\omega$, the Euler identity gives

$$e^{\lambda t} = e^{\sigma t}e^{j\omega t} = e^{\sigma t}\left[\cos(\omega t) + j\sin(\omega t)\right]$$

This looks like a sinusoid of frequency ω modulated by an exponential envelop $e^{\sigma t}$. The real part of $e^{\lambda t}$ is shown in Figure 14. Note that

(a) there are oscillations

(b) the response decays exponentially to zero if and only if Real $\lambda < 0$ with faster decay when $\sigma = \text{Real } \lambda$ is more negative

(c) the response grows exponentially when Real $\lambda > 0$ with faster growth when $\sigma = \text{Real } \lambda$ is more positive

(d) the frequency of the oscillations increases with ω

(e) the response is a pure sinusoid when $\sigma = 0$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

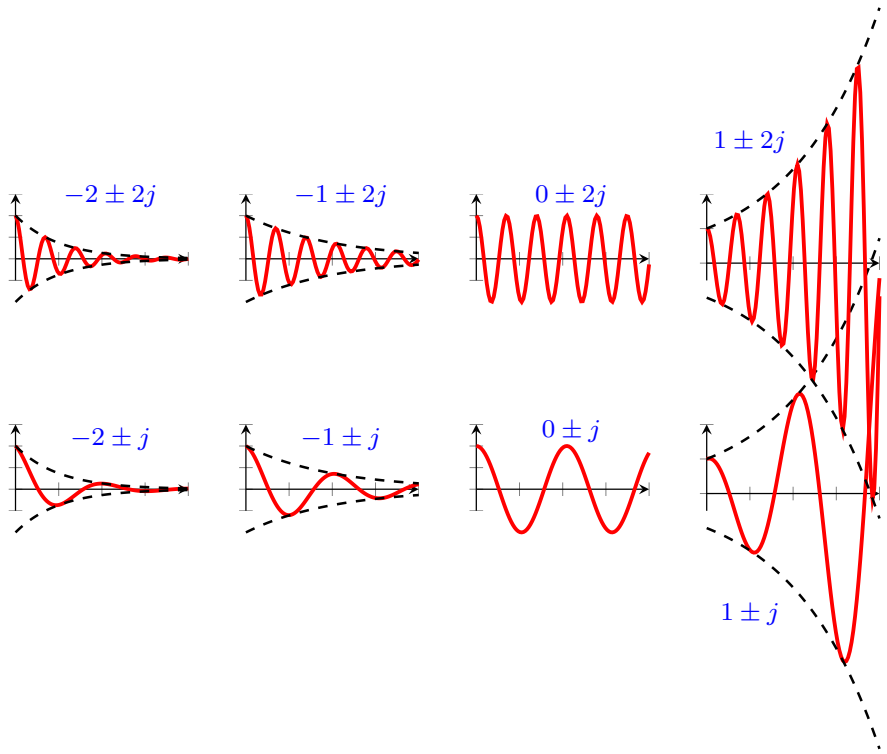

Figure 14: Real part of $e^{\lambda t}$ for various complex values of λ.

When A is not semisimple, the free response has terms like

$$e^{\lambda t}, te^{\lambda t}, t^2 e^{\lambda t}, \cdots, t^{q-1} e^{\lambda t}$$

where q is the size of the largest Jordan block associated with the eigenvalue λ.

3 Forced Response as a Convolution

Note that, when $D = 0$, the forced response has the form

$$y(t) = \int_0^t h(t-\tau) u(\tau) d\tau$$

where

$$h(t) = Ce^{At}B$$

You may recognize the integral above as the convolution of the functions h and u, denoted $h * u$, if you have studied convolution before. The function h above is called the impulse response, as it is the output when we apply the Dirac delta function as the input, i.e. $u = \delta$, with zero initial conditions. Indeed, applying the forced response formula with $u = \delta$, we get

$$\int_0^t Ce^{A(t-\tau)} B\delta(\tau) d\tau = \begin{cases} Ce^{At}B & t \geq 0 \\ 0 & t < 0 \end{cases}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

4 Discrete-time LTI solution

Consider the discrete-time LTI realization

$$\begin{bmatrix} x_{k+1} \\ y_k \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} x_k \\ u_k \end{bmatrix}, \quad \text{with initial condition } x_0 = \xi$$

The solution of these equations is

$$\begin{aligned} x_k &= A^k \xi + \sum_{\ell=0}^{k-1} A^\ell B u_{k-1-\ell} \\ y_k &= CA^k \xi + \sum_{\ell=0}^{k-1} CA^\ell B u_{k-1-\ell} + Du_k \end{aligned}$$

We can rewrite these equations as

$$\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix} = \begin{bmatrix} I \\ A \\ A^2 \\ \vdots \\ A^k \end{bmatrix} \xi + \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 \\ B & 0 & 0 & \cdots & 0 \\ AB & B & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ A^{k-1}B & A^{k-2}B & A^{k-3}B & \cdots & 0 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{k-1} \end{bmatrix}$$

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix} = \begin{bmatrix} I \\ A \\ A^2 \\ \vdots \\ A^k \end{bmatrix} \xi + \begin{bmatrix} D & 0 & 0 & \cdots & 0 \\ CB & D & 0 & \cdots & 0 \\ CAB & CB & D & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ CA^{k-1}B & CA^{k-2}B & CA^{k-3}B & \cdots & D \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{k-1} \end{bmatrix}$$

The first term is the free response, the second term is the forced response. Define the Markov parameters:

$$H_0 = D, \quad H_k = CA^{k-1}B$$

The forced response can now be written more clearly as:

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_k \end{bmatrix}_{\text{forced}} = \begin{bmatrix} H_0 & 0 & 0 & \cdots & 0 \\ H_1 & H_0 & 0 & \cdots & 0 \\ H_2 & H_1 & H_0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ H_{k-1} & H_{k-2} & H_{k-3} & \cdots & H_0 \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \\ u_{k-1} \end{bmatrix}$$

or more compactly, as

$$y = Tu$$

The matrix T is block lower-triangular It has a special structure – blocks repeat down all diagonals. Such a matrix is called Toeplitz.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

5 MatLab Commands

State-space realizations offer a very convenient data structure to handle LTI ODEs. In MATLAB, we can use the `ss` command to create state space realization objects. For example these simple commands create the object `plant` as a state-space realization.

```
>> A = [0  1 0; 0 0 1; -2 -3 -4];
>> B = [0; 0; 1];
>> C = [1 0 0];
>> D = 0;
>> plant = ss(A,B,C,D);
```

We can simulate the response of a state-space realization to inputs `u` and initial conditions `x0` using the `lsim` command.

```
% lsim(sys,u,t)
% lsim(sys,u,t,x0)
>> t = linspace(0,10, 1000);
>> u = sin(0.8*t) + randn(1,1000);
>> y = lsim(plant, u, t);
>> plot(t,y);
% to specify the initial conditions x0 use
>> y = lsim(plant, u, t, x0)
```

For discrete-time state-space models, you can specify the sampling time `ts`.

```
% to leave sampling time unspecified, use ts = -1
>> plant = ss(A,B,C,D, ts);
% simulation works as before
>> y = lsim(plant, u, t, x0);
>> plot(t,y);
```

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

B. Stability

1 Internal Stability

There are many notions of stability for state-space models. For LTI systems these diverse notions collapse and are equivalent. We begin with the following, which only concerns the free response of the system:

Definition 93. The system Σ is called asymptotically stable if the solution of

$$\dot{x} = Ax, \quad x(0) = \xi$$

with *any* initial condition $\xi \in \mathbb{R}^n$ satisfies

$$\lim_{t \to \infty} \quad x(t) = 0$$

Theorem 94. *The linear time-invariant system $\Sigma(A, *, *, *)$ is asymptotically stable if and only if all the eigenvalues of A have negative real parts:*

$$Real(Spec(A)) < 0$$

Proof: We assume A has distinct eigenvalues say $\lambda_1, \cdots, \lambda_n$ with associated eigenvectors $v_1, \cdots, v_n$. Then, we can write

$$A = T\Lambda T^{-1} = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_n \end{bmatrix} \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix}^{-1}$$

Using this decomposition, we can calculate

$$e^{At} = Te^{\Lambda t}T^{-1} = T \begin{bmatrix} e^{\lambda_1 t} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & e^{\lambda_n t} \end{bmatrix} T^{-1}$$

Clearly $e^{At} \to 0$ if and only if $e^{\lambda_k t} \to 0$ for each eigenvalue λ_k. That will happen if and only if $\text{Real}\{\lambda\} < 0$. □

A matrix whose eigenvalues have strictly negative real parts is sometimes called Hurwitz.

2 The Lyapunov equation

Theorem 95. *Consider the matrix equation*

$$A^*P + PA + Q = 0 \tag{11}$$

The following are equivalent

(a) The Lyapunov equation has a unique solution P for some Q

(b) The Lyapunov equation has a solution P for every Q

(c) The Lyapunov equation has a unique solution P for every Q

(d) For all $\lambda_i, \lambda_j \in Spec(A)$, $\quad \lambda_i + \overline{\lambda_j} \neq 0$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Proposition 96. *If A is Hurwitz, the unique solution of the Lyapunov equation (11) is*

$$P = \int_0^\infty e^{A^*t} Q e^{At} dt$$

3 Stability Test using the Lyapunov Equation

Asymptotic stability can also be characterized in terms of the Lyapunov equation:

Theorem 97. *If there exists $P \succ 0$ such that*

$$Q = -A^*P - PA \succ 0$$

then the system Σ is asymptotically stable.

Algebraic proof. We show that the condition above implies A is Hurwitz, this Σ is asymptotically stable. Let $\lambda \in \text{Spec } (A)$. Then

$$Av = \lambda v, v \neq 0$$

Notice that $v^*A^* = \overline{\lambda} v^*$. Multiplying the Lyapunov equation by v on the right and v^* on the left we get

$$\begin{aligned} v^*Qv &= -v^*A^*Pv - v^*PAv = -\overline{\lambda} v^*Pv - \lambda v^*Pv + v^*Qv \\ &= -(\lambda + \overline{\lambda}) v^*Pv = -2 \text{ Real } \lambda v^*Pv \end{aligned}$$

Since $P, Q \succ 0$ and $v \neq 0$ we conclude that

$$\text{Real } \lambda = -\frac{v^*Qv}{2v^*Pv} < 0$$

proving that A is Hurwitz.

Trajectory-based proof. This proof bypasses the eigenvalue characterization of asymptotic stability, used in the algebraic proof. It shows directly that $x(t) \to 0$. Define the function

$$V(x) = x^*Px$$

which satisfies

$$V(x) \geq 0 \quad \text{and} \quad V(x) = 0 \iff x = 0$$

Next along trajectories of $\dot{x} = Ax$, we have

$$\frac{d}{dt} V(x(t)) = \dot{x}(t)^*Px(t) + x(t)^*P\dot{x}(t) = x(t)^*A^*Px(t) + x(t)^*PAx(t) = -x(t)^*Qx(t)$$

Since $Q \succ 0$, we have $\frac{d}{dt} V(x(t)) \leq 0$ so $V(x(t))$ is a decreasing function of time. Since $P \succ 0$, $V(x(t)) \geq 0$. These two facts together imply that $V(x(t))$ converges to a constant $c \geq 0$. It can be argued[2] that $c = 0$, which means that $V(x(t)) \to 0$. Since $x^*Px = 0$ implies $x = 0$ by $P \succ 0$, we conclude $x(t) \to 0$. □

[2] $c > 0$ leads to a contradiction: $\frac{d}{dt} V(x(t)) < 0$ only when $x(t) = 0$, that is, only when $V(x(t)) = 0$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The function V used in the trajectory-based proof is called a Lyapunov function. It converges to zero along the trajectories of the system because its time derivative is a negative definitie function of the state. Since $V(x)$ is positive for all $x \neq 0$, convergence of $V(x)$ to zero implies convergence of x to zero.

The idea of finding a positive definite Lyapunov function whose time derivative is negative definite extends naturally to nonlinear systems. Although nonlinear systems are beyond the scope of this course, we use the following example as a simple illustration.

Example 98. We use $V(x) = x^2$ to show asymptotic stability of $\dot{x} = -x^3$. Note that

$$\frac{d}{dt}V(x(t)) = 2x(t)\dot{x}(t) = -2x(t)^4$$

which is negative except when the state is zero. The arguments above then guarantee $x(t) \to 0$.

4 Input-Output Stability

Input-output notions of stability focus on the effect of the inputs on the outputs, rather than the unforced system behaviour.

Definition 99. A signal $u(t)$ is called bounded if there exists a constant K_u such that

$$\|u(t)\| \leq K_u < \infty \quad \text{for all } t$$

For example the signal $u(t) = \sin(t)$ is bounded, but the signal $u(t) = t$ is not.

A common input-output stability notion is bounded-input bounded-output (BIBO) stability.

Definition 100. A system is called BIBO stable if every bounded input produces a bounded output.

For LTI systems, our earlier notion of internal stability implies BIBO stability.

Theorem 101. *An asymptotically stable LTI system is BIBO stable.*

Proof: Recall that the forced response is

$$y(t) = \int_0^t Ce^{A(t-\tau)}Bu(\tau)d\tau + Du(t)$$

When u is bounded, the second term is bounded, so we need to show boundedness of the first term. We rewrite is as

$$\int_0^t h(t-\tau)u(\tau)d\tau$$

where $h(t) = Ce^{At}B$. With the shifted time variable $s = t - \tau$, the integral above becomes

$$\int_0^t h(s)u(t-s)ds$$

Note that

$$\left\| \int_0^t h(s)u(t-s)ds \right\| \leq \int_0^t \|h(s)u(t-s)\|ds \leq \int_0^t \|h(s)\|\|u(t-s)\|ds \leq K_u \int_0^t \|h(s)\|ds$$

Moreover, the rightmost term is bounded by $K_u \int_0^\infty \|h(s)\|ds$, which is well defined because $h(t) = Ce^{At}B$ is a combination of terms of the form $e^{\lambda t}, te^{\lambda t}, t^2e^{\lambda t}, \ldots, \lambda \in Spec(A)$ and all such terms are absolutely integrable when $Real(\lambda) < 0$ (guaranteed by the asymptotic stability assumption). Thus, the output is bounded and the system is BIBO stable.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

C. LTV Systems

1 Solution of LTV systems

Unlike the LTI system

$$\dot{x}(t) = Ax(t),\ x(t_0) = \xi$$

whose solution is given by $x(t) = e^{A(t-t_0)}\xi$, for the LTV system

$$\dot{x}(t) = A(t)x(t),\ x(t_0) = \xi$$

there is no explicit formula for the solution. Nevertheless the following characterization of the solution is useful:

Let $\phi_i(t, t_0)$ denote the solution of the unforced LTV system above when the initial condition is e_i, the ith unit vector. This means:

$$\frac{\partial}{\partial t}\phi_i(t, t_0) = A(t)\phi_i(t, t_0),\ \phi_i(t_0, t_0) = e_i$$

For a more compact representation, we define the $n \times n$ matrix

$$\Phi(t, t_0) := [\phi_1(t, t_0), \cdots, \phi_n(t, t_0)]$$

and note that it satisfies the matrix differential equation

$$\frac{\partial}{\partial t}\Phi(t, t_0) = A(t)\Phi(t, t_0),\ \Phi(t_0, t_0) = I$$

Indeed, the ith column of this matrix differential equation corresponds to the differential equation for ϕ_i above.

The solution of the unforced LTV system from an arbitrary initial condition ξ is then given by:

$$x(t) = \Phi(t, t_0)\xi$$

The matrix $\Phi(t, t_0)$ is called the state transition matrix because it determines where the state vector $x(t)$ ends up at time t when it starts at ξ at time t_0. In the special case of an LTI system, where A is constant, $\Phi(t, t_0) = e^{A(t-t_0)}$.

Note that the formula above is not an explicit solution because, to find $\Phi(t, t_0)$, we would have to solve the system equations (from the initial conditions $e_1, \cdots, e_n$). Instead, it is a characterization of solutions from an arbitrary initial condition ξ in terms of the solutions from n initial conditions. Once we compute these n solutions (typically using numerical integration), we can form the state transition matrix and use it to obtain the solution from any other initial condition.

The state transition matrix also enables to characterize the solutions of a forced LTV system

$$\dot{x}(t) = A(t)x(t) + B(t)u(t),\ x(t_0) = \xi$$

The solution is now

$$x(t) = \Phi(t, t_0)\xi + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau$$

which you can verify by substituting back into the differential equation and by using the matrix differential equation for Φ derived above.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

2 Stability of LTV systems can't be judged from eigenvalues

Checking whether the eigenvalues of A have negative real parts was a convenient stability test in the LTI case. For a time-varying $A(t)$ we can't conclude stability even if the eigenvalues have negative real parts at each time t. Here is a counterexample:

$$A(t) = \begin{bmatrix} -1 + 1.5\cos^2 t & 1 - 1.5\sin t \cos t \\ -1 - 1.5\sin t \cos t & -1 + 1.5\sin^2 t \end{bmatrix}$$

The eigenvalues are $-0.25 \mp i0.25\sqrt{7}$ for all t. They have negative real parts and don't even vary in time. Yet, the state transition matrix with $t_0 = 0$ has a term that grows unbounded:

$$\Phi(t, 0) = \begin{bmatrix} e^{0.5t}\cos t & e^{-t}\sin t \\ e^{-0.5t}\sin t & e^{-t}\cos t \end{bmatrix}$$

If $x(0) = e_1$, then the solution $x(t)$ is the first column of this matrix, which is unbounded. Therefore, the system is unstable.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

D. Controllability

1 T-Controllability

Definition 102. Consider the LTI system

$$\dot{x}(t) = Ax(t) + Bu(t),\ x(0) = \xi$$

where $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}^m$, and fix a time $T > 0$. The initial state $\xi \in \mathbb{R}^n$ is said to be T-controllable if there exists an input $u : [0, T] \to \mathbb{R}^m$ that drives the state from ξ to $x(T) = 0$.

Define $\mathcal{C}_T \subseteq \mathbb{R}^n$ to be the set of T-controllable states. The system is called T-controllable if every state $\xi \in \mathbb{R}^n$ is T-controllable, i.e., if $\mathcal{C}_T = \mathbb{R}^n$.

2 The Controllable Subspace $\mathcal{C}_T$

Recall that the state at $t = T$ is given by

$$x(T) = e^{AT}\xi + e^{AT}\int_0^T e^{-At}Bu(t)dt$$

Setting $x(T) = 0$ and using invertibility of e^{AT}, we reach the following characterization of $\mathcal{C}_T$:

Lemma 103. *$\xi \in \mathcal{C}_T$ if and only if there exists an input $u : [0, T] \to \mathbb{R}^m$ such that*

$$\xi = -\int_0^T e^{-At}Bu(t)dt$$

We can view the right-hand side of this as a linear operator mapping the input function u to $\mathbb{R}^n$:

$$\mathcal{L}u = \int_0^T G(t)u(t)dt, \quad G(t) := -e^{-At}B$$

Then, the lemma above implies that $\mathcal{C}_T$ is the range space of this operator: $\mathcal{C}_T = \mathcal{R}(\mathcal{L})$. An immediate consequence is the following:

Lemma 104. *The set $\mathcal{C}_T$ of T-controllable states is a subspace of $\mathbb{R}^n$.*

We now give a sharper characterization of $\mathcal{C}_T$ using the fact[3]

$$\mathcal{R}(\mathcal{L}) = \mathcal{R}(\mathcal{L}\mathcal{L}^*)$$

where $\mathcal{L}^*$ is the adjoint. The benefit of using $\mathcal{L}\mathcal{L}^*$ is that it maps $\mathbb{R}^n$ to $\mathbb{R}^n$ and can be represented as a $n \times n$ matrix, whose range space is easy to find. In contrast, the domain of $\mathcal{L}$ is an infinite dimensional vector space of input functions, which makes $\mathcal{L}$ difficult to analyze.

The operator $\mathcal{L}u = \int_0^T G(t)u(t)dt$ was studied as an example for adjoint operators in the previous chapter. It was shown that for $z \in \mathbb{R}^n$, $\mathcal{L}^*z$ is a function whose value at time t is $G(t)^Tz$. Thus,

$$\mathcal{L}\mathcal{L}^*z = \mathcal{L}(G(t)^Tz) = \int_0^T G(t)G(t)^Tzdt = \left(\int_0^T e^{-At}BB^Te^{-A^Tt}dt\right)z$$

Once the integral is computed, the bracketed term on the right is a constant $n \times n$ matrix. It is called the controllability Grammian and its range space is precisely the controllable subspace.

[3] We showed this for matrices before, but the same result is true for operators whose codomain is finite dimensional.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3 Controllability Grammians

Definition 105. The controllability Grammian for the system Σ on the interval $[0, T]$ is the matrix $W(0,T) \in \mathbb{R}^{n\times n}$ defined by

$$W(0,T) = \int_0^T e^{-At} B B^T e^{-A^T t} dt$$

Theorem 106. *Consider the system Σ and let $T > 0$. Then,*

(a) $\mathcal{C}_T = \mathcal{R}\left[W(0,T)\right]$

(b) Let $\xi \in \mathcal{C}_T$. From (a) above, there exists a vector $z \in \mathbb{R}^n$ such that $\xi = W(0,T)z$. Then, the input

$$u(t) = -B^T e^{-A^T t} z, \quad 0 \le t \le T$$

drives the state from initial state $x(0) = \xi$ to terminal state $x(T) = 0$.

Proof: We already argued part (a) in the previous section. To show part (b), recall that an input driving the state to zero at time T must satisfy

$$\xi = \mathcal{L}u$$

To find such u we can first find $z \in \mathbb{R}^n$ such that

$$\xi = \mathcal{L}\mathcal{L}^* z = W(0,T)z$$

and set $u = \mathcal{L}^* z$. The resulting function u is indeed the one proposed in part (b). □

4 Test for Controllability

Theorem 107. *Define the matrix $M_c \in \mathbb{R}^{n\times mn}$ as*

$$M_c = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}$$

Then, for any $T > 0$,

$$\mathcal{C}_T = \mathcal{R}(M_c)$$

and, thus, the system is controllable iff rank $(M_c) = n$.

M_c is called the controllability matrix. Note that it does not depend on the time T allotted for controllability. We therefore conclude that if a state $\xi \in \mathbb{R}^n$ is controllable on the interval $[0, T]$, it is controllable on any (nonzero) interval. We thus drop the superfluous argument T from the notion of T-controllability and the controllable subspace, which we now write as $\mathcal{C}$.

Proof of the Theorem: We already know $\mathcal{C}_T = \mathcal{R}(W(0,T))$, so we will show $\mathcal{R}(W(0,T)) = \mathcal{R}(M_c)$ which is equivalent to

$$\mathcal{R}(W(0,T))^\perp = \mathcal{R}(M_c)^\perp$$

But $\mathcal{R}(M_c)^\perp = \mathcal{N}(M_c^T)$ and $\mathcal{R}(W(0,T))^\perp = \mathcal{N}(W(0,T)^T) = \mathcal{N}(W(0,T))$, where the last equality is due to the symmetry of $W(0,T)$, so we need to prove

$$\mathcal{N}(W(0,T)) = \mathcal{N}(M_c^T)$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Since $W(0,T)$ is positive semidefinite (it is the integral of a Gram matrix which is positive semidefinite at each time), $x \in \mathcal{N}(W(0,T))$ is equivalent to

$$x^T W(0,T)x = 0.$$

Substituting $W(0,T)$ from its definition,

$$x^T W(0,T)x = \int_0^T x^T e^{-At} B B^T e^{-A^T t} x dt = \int_0^T w(t)^T w(t) dt = \int_0^T \|w(t)\|^2 dt$$

where $w(t) := B^T e^{-A^T t} x$. Thus,

$$x \in \mathcal{N}(W(0,T)) \;\Leftrightarrow\; x^T W(0,T)x = 0 \;\Leftrightarrow\; w(t) = 0 \;\forall t \in [0,T]$$

Note that

$$w(t)^T = x^T e^{-At} B = x^T \left(I - At + \frac{1}{2}A^2 t^2 - \frac{1}{3!}A^3 t^3 + \cdots \right) B = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \left(x^T A^k B \right) t^k$$

and, thus, $w(t) = 0 \;\forall t \in [0,T]$ means

$$x^T A^k B = 0, \;\; k = 0, 1, 2, \cdots$$

By Cayley-Hamilton Theorem, this is equivalent to the n equalities

$$x^T A^k B = 0, \;\; k = 0, 1, 2, \cdots, n-1$$

because, for $k \geq n$, A^k is a linear combination of $I, A, \ldots, A^{n-1}$.

To summarize, we have shown that $x \in \mathcal{N}(W(0,T))$ is equivalent to the n equalities above, which we rewrite compactly as

$$x^T \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix} = 0$$

But this means $M_c^T x = 0$, i.e., $x \in \mathcal{N}(M_c^T)$, proving $\mathcal{N}(W(0,T)) = \mathcal{N}(M_c^T)$.

5 Example

Consider the system $\Sigma(A, B, *, *)$ where

$$A = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-2} & -\alpha_{n-1} \end{bmatrix} \qquad B = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

It can be readily verified that the controllability matrix for this realization is

$$M_c = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \\ 0 & 0 & \cdots & 1 & * \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 1 & \cdots & * & * \\ 1 & * & \cdots & * & * \end{bmatrix}$$

Observe that rank $(M_c) = n$. As a result, Σ is controllable. This is the *canonical* example of a single-input controllable system.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

E. Observability

1 The Unobservable Subspace $\mathcal{UO}$

Definition 108. Consider the system Σ. A state $\xi \in \mathbb{R}^n$ is called unobservable if, with initial condition $x(0) = \xi$ and with input $u(t) = 0,\ t \geq 0$, the output trajectory is

$$y(t) = Ce^{At}\xi = 0 \quad \text{for all} \quad t \geq 0$$

Let $\mathcal{UO} \subseteq \mathbb{R}^n$ be the set of unobservable states of Σ.

The realization Σ is called observable if the only unobservable state is the zero state, i.e. $\mathcal{UO} = 0$.

2 Test for Observability

Theorem 109. *Consider the system Σ and define the matrix $M_o \in \mathbb{R}^{pn \times n}$ as*

$$M_o = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

Then

$$\mathcal{UO} = \mathcal{N}(M_o)$$

and, thus, the system is observable iff rank $(M_o) = n$.

M_o is called the observability matrix. Note that an immediate consequence of this theorem is that the set $\mathcal{UO}$ of unobservable states is a subspace of $\mathbb{R}^n$.

Proof of the Theorem: Note that

$$Ce^{At}\xi = C(I + At + \frac{1}{2}A^2t^2 + \cdots)\xi$$

Thus $Ce^{At}\xi = 0$ for all $t \geq 0$ means

$$CA^k\xi = 0,\ k = 1, 2, \cdots$$

By Cayley-Hamilton Theorem, this is equivalent to the n equalities

$$CA^k\xi = 0,\ k = 0, 1, 2, \cdots, n-1$$

because, for $k \geq n$, A^k is a linear combination of $I, A, \ldots, A^{n-1}$. Rewriting these n equalities compactly as

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \xi = M_0\xi = 0$$

we see that $Ce^{At}\xi = 0$ is equivalent to $\xi \in \mathcal{N}(M_o)$, which proves the theorem. □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

3 Observability Grammian

Definition 110. The observability Grammian for the system Σ on the interval $[0,T]$ is the matrix $U(0,T) \in \mathbb{R}^{n\times n}$ defined by

$$U(0,T) = \int_0^T e^{A^T t} C^T C e^{At} dt$$

Theorem 111. *Consider the system* Σ *and let* $T > 0$. *Then,*

$$\mathcal{UO} = \mathcal{N}(U(0,T))$$

Proof: Note that $U(0,T)$ is positive semidefinite by construction and, thus, $\xi \in \mathcal{N}(U(0,T))$ iff

$$\xi^T U(0,T)\xi = \int_0^T \xi^T e^{A^T t} C^T C e^{At} \xi dt = \int_0^T \|Ce^{At}\xi\|^2 dt = 0$$

This means $Ce^{At}\xi = 0$ for all $t \in [0,T]$ and, thus, $\xi \in \mathcal{N}(M_o)$ by the arguments in the proof of the previous theorem. Therefore, $\mathcal{N}(U(0,T)) = \mathcal{N}(M_o) = \mathcal{UO}$. □

Note that the value of $T > 0$ in the theorem is immaterial: the null space of $U(0,T)$ is the unobservable subspace for any $T > 0$. When A is Hurwitz, $\lim_{T\to\infty} U(0,T)$ exists because the integrand in the definition above converges exponentially and

$$U(0,\infty) = \int_0^\infty e^{A^T t} C^T C e^{At} dt$$

exists. The advantage of considering this limit is that this integral is the unique solution of the Lyapunov Equation

$$A^T P + PA + C^T C = 0$$

as we saw before. Thus we can evaluate the observability Grammian $U(0,\infty)$ by solving this Lyapunov equation algebraically rather than by evaluating an interval. The following theorem is a specialization of the previous one to the limit $T \to \infty$, which allows us to use the solution of the Lyapunov Equation for $U(0,\infty)$.

Theorem 112. *Consider the realization* Σ *and suppose* A *is Hurwitz. Then,*

$$\mathcal{UO} = \mathcal{N}(P)$$

where P *is the unique solution of the Lyapunov equation above.*

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

F. Modal Observability and Controllability Tests

1 Duality

The dual of the system Σ is defined as

$$\Sigma^{\text{dual}} = \left[\begin{array}{c|c} A^{\text{dual}} & B^{\text{dual}} \\ \hline C^{\text{dual}} & D^{\text{dual}} \end{array}\right] = \left[\begin{array}{c|c} A^T & C^T \\ \hline B^T & D^T \end{array}\right]$$

We now show that the dual system is observable iff the original system is controllable, and vice versa. To see this, note

$$M_o^{\text{dual}} = \begin{bmatrix} C^{\text{dual}} \\ C^{\text{dual}}A^{\text{dual}} \\ \vdots \\ C^{\text{dual}}{A^{\text{dual}}}^{n-1} \end{bmatrix} = \begin{bmatrix} B^T \\ B^TA^T \\ \vdots \\ B^T{A^T}^{n-1} \end{bmatrix} = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}^T = M_c^T$$

Thus $\text{rank}(M_o^{\text{dual}}) = n$ iff $\text{rank}(M_c) = n$; that is, the dual system is observable iff the original system is controllable.

2 Modal Observability and Controllability Tests

The following test uses the eigenvectors of A and the matrix C to judge observability:

Theorem 113. *The system Σ is unobservable iff A has an eigenvector v such that $Cv = 0$.*

If such an eigenvector exists, the corresponding eigenvalue is called an unobservable mode. If no such eigenvector exists, then the system is observable by the theorem.

Proof of the "if" statement: Let $v \neq 0$ be such that $Av = \lambda v$ and $Cv = 0$. Then,

$$\begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} v = \begin{bmatrix} Cv \\ CAv \\ \vdots \\ CA^{n-1}v \end{bmatrix} = \begin{bmatrix} Cv \\ C\lambda v \\ \vdots \\ C\lambda^{n-1}v \end{bmatrix} = 0$$

which means $v \in \mathcal{N}(M_o) = \mathcal{UO}$. Since $v \neq 0$, $\mathcal{UO}$ is nontrivial and the system is unobservable.

We will prove the "only if" statement after we discuss the Kalman Decomposition below. Note that, if the unobservability condition above holds, then the eigenvector v lies in the nullspace of

$$\begin{bmatrix} A - \lambda I \\ C \end{bmatrix}.$$

Since $v \neq 0$, this means the null space is nontrivial and, thus, the columns are linearly dependent. Conversely, if this matrix has full column rank, than no eigenvector satisfies the condition of the theorem above. This observation is known as the Popov-Belevich-Hautus (PBH) criterion:

Theorem 114. *The system Σ is observable iff, for all $\lambda \in Spec(A)$,*

$$\text{rank}\left(\begin{bmatrix} A - \lambda I \\ C \end{bmatrix}\right) = n.$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The results above can be translated to controllability using duality:

Theorem 115. *The system Σ is uncontrollable iff A^T has an eigenvector v such that $B^T v = 0$. Equivalently, the system is controllable iff, for all $\lambda \in Spec(A)$,*

$$\text{rank}\left(\begin{bmatrix} A - \lambda I, B \end{bmatrix}\right) = n.$$

3 A-Invariance of the Controllable and Unobservable Subspaces

Definition 116. Given a matrix $A \in \mathbb{R}^{n\times n}$, a subspace $\mathcal{S} \subset \mathbb{R}^n$ is called A-invariant if it closed under multiplication with A:

$$\xi \in \mathcal{S} \;\Rightarrow\; A\xi \in \mathcal{S}.$$

Note that the span of an eigenvector (or a set of eigenvectors) of A is A-invariant by this definition.

Lemma 117. *The unobservable and controllable subspaces of Σ are A-invariant.*

Proof of the Lemma: We start with the unobservable subspace $\mathcal{UO}$. Since $\mathcal{UO} = \mathcal{N}(M_o)$, $\xi \in \mathcal{UO}$ means

$$M_0\xi = \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} \xi = 0$$

By the Cayley-Hamilton Theorem, we also have $CA^n\xi = 0$, so

$$\begin{bmatrix} CA \\ CA^2 \\ \vdots \\ CA^n \end{bmatrix} \xi = 0 \quad \Rightarrow \quad \begin{bmatrix} C \\ CA \\ \vdots \\ CA^{n-1} \end{bmatrix} A\xi = 0$$

Thus, $A\xi \in \mathcal{N}(M_o) = \mathcal{UO}$, which means that the unobservable subspace is A-invariant.

Next, we show A-invariance of the controllable subspace $\mathcal{C}$, which is the range space of the controllability matrix

$$M_c = \begin{bmatrix} B & AB & \cdots & A^{n-1}B \end{bmatrix}.$$

Thus, if $\xi \in \mathcal{C}$, then $A\xi$ belongs to the range space of

$$AM_c = \begin{bmatrix} AB & A^2B & \cdots & A^nB \end{bmatrix}.$$

By the Cayley-Hamilton Theorem, A^nB can be written as a linear combination of $B, AB, \cdots, A^{n-1}B$; therefore $A\xi$ also lies in the range space of M_c, which is the controllable subspace $\mathcal{C}$. Thus, the controllable subspace is also A-invariant. □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

G. Kalman Decomposition

1 Similar Realizations

Recall that a similarity transformation of a state space representation consists of a change of variables. We choose a nonsingular matrix $T \in \mathbb{R}^{n\times n}$ and define *new* states via

$$Tx_{new} = x$$

If we rewrite the differential equations defining Σ in terms of these new states, we arrive at

$$\Sigma_{new} \quad \begin{cases} \dot{x}_{new}(t) &= T^{-1}AT x_{new}(t) + T^{-1}Bu(t) \\ y(t) &= CT x_{new}(t) + Du(t) \end{cases}$$

This new realization

$$\Sigma_{new} = \left[\begin{array}{c|c} T^{-1}AT & T^{-1}B \\ \hline CT & D \end{array}\right] = \left[\begin{array}{c|c} A_{new} & B_{new} \\ \hline C_{new} & D_{new} \end{array}\right] \tag{12}$$

is said to be similar to Σ.

Similar representations are fundamentally the same. They share identical properties such as stability, controllability, etc., and they yield the same transfer function:

$$H_{new}(s) = C_{new}(sI - A_{new})^{-1}B_{new} + D_{new} = C(sI - A)^{-1}B + D = H(s).$$

Indeed, we arrived at Σ_{new} from Σ with only a change of variables, and Σ_{new} is a different realization of the same transfer function.

2 Kalman Decomposition into Controllable and Uncontrollable Subsystems

We start with a state-space model Σ. By an appropriate choice of basis for the state-space, we can exhibit clearly the "controllable subsystem" and the "uncontrollable subsystem" as shown in Figure 15. Notice that the controllable subsystem is directly affected by the input. The uncontrollable subsystem is unaffected by the input and evolves autonomously. We can discard the uncontrollable subsystem without changing the input-output behavior of the realization Σ.

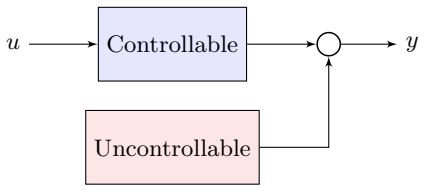

Figure 15: Controllable and Uncontrollable Subsystems.

Theorem 118. *Consider the system $\Sigma(A, B, C, D)$ with transfer function $H(s)$ and let $dim(\mathcal{C}) = r$. Let $\{t_1, \cdots, t_r\}$ be a basis for $\mathcal{C}$ and extend this by $\{t_{r+1}, \cdots, t_n\}$ to form a basis for $\mathbb{R}^n$. Then*

$$T = \left[\begin{array}{cccccc} t_1 & \cdots & t_r & t_{r+1} & \cdots t_n \end{array}\right]$$

is invertible, so we can do a similarity transformation using the change of variables $Tx_{new} = x$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

(a) Σ_{new} *has the structure*

$$\Sigma_{new} = \left[\begin{array}{c|c} T^{-1}AT & T^{-1}B \\ \hline CT & D \end{array}\right] = \left[\begin{array}{cc|c} A_{11} & A_{12} & B_1 \\ 0 & A_{22} & 0 \\ \hline C_1 & C_2 & D \end{array}\right]$$

where $A_{11} \in \mathbb{R}^{r\times r}$*, and* B *and* C *are partitioned conformably with the partition of* A*.*

(b) The reduced model

$$\Sigma_{reduced}(A_{11}, B_1, C_1, D)$$

is controllable and also realizes $H(s)$*.*

It is clear that the Kalman decomposition is not unique: different basis choices yield different decompositions, though they all share the structure described in the above theorem.

Proof: The key features of A_{new} and B_{new} are the zero blocks. To see how the zero block arises at the bottom left of A_{new}, recall that the controllable subspace $\mathcal{C}$ is A-invariant. This means that for each basis vector t_i, $i = 1, \cdots, r$, we have $At_i \in \mathcal{C}$, meaning that At_i can be written as a linear combination of $t_1, \cdots, t_r$:

$$At_i = \sum_{j=1}^{r} \alpha_{ji} t_j.$$

If we define a $r \times r$ matrix whose (i, j) entry is the coefficient α_{ij}, we get:

$$A[t_1 \cdots t_r] = [t_1 \cdots t_r] A_{11}.$$

Thus,

$$A \underbrace{\left[\begin{array}{cccccc} t_1 & \cdots & t_r & t_{r+1} & \cdots & t_n \end{array}\right]}_{T} = \underbrace{\left[\begin{array}{cccccc} t_1 & \cdots & t_r & t_{r+1} & \cdots & t_n \end{array}\right]}_{T} \left[\begin{array}{cc} A_{11} & A_{12} \\ 0 & A_{22} \end{array}\right]$$

with appropriate matrices A_{12}, A_{22}, and

$$A_{new} = T^{-1}AT = \left[\begin{array}{cc} A_{11} & A_{12} \\ 0 & A_{22} \end{array}\right].$$

To see the reason for the zero block at the bottom of B_{new}, recall that $\mathcal{C}$ coincides with the range space of the controllability matrix

$$M_c = \left[\begin{array}{cccc} B & AB & \cdots & A^{n-1}B \end{array}\right].$$

Thus, the columns of B lie in $\mathcal{C}$, and can be written as linear combinations of the basis vectors $t_1, \cdots, t_r$. This means that

$$B = [t_1 \cdots t_r] B_1 = \left[\begin{array}{cccccc} t_1 & \cdots & t_r & t_{r+1} & \cdots & t_n \end{array}\right] \left[\begin{array}{c} B_1 \\ 0 \end{array}\right] = T \left[\begin{array}{c} B_1 \\ 0 \end{array}\right]$$

for an appropriately dimensioned matrix B_1. It then follows that

$$B_{new} = T^{-1}B = \left[\begin{array}{c} B_1 \\ 0 \end{array}\right].$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Next, we show that

$$\Sigma_{reduced}(A_{11}, B_1, C_1, D)$$

is controllable. To see this, we note that the rank of the controllability matrix M_c above is r, since its columns span the r-dimensional controllable subspace. Thus,

$$T^{-1}M_c = \begin{bmatrix} T^{-1}B & T^1ATT^{-1}B & \cdots & T^1A^{n-1}TT^1B \end{bmatrix} = \begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{n-1}B_1 \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

also has rank r, which means

$$\begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{n-1}B_1 \end{bmatrix}$$

has rank r because the additional zero rows above do not change the rank. Since $r \leq n$, it follows from the Cayley-Hamilton theorem that the range space of this matrix is the same as that of

$$M_{c,reduced} = \begin{bmatrix} B_1 & A_{11}B_1 & \cdots & A_{11}^{r-1}B_1 \end{bmatrix}.$$

Thus, $M_{c,reduced}$ has rank r, proving the controllability of $\Sigma_{reduced}(A_{11}, B_1, C_1, D)$ which has r state variables.

Finally, to see that $\Sigma_{reduced}(A_{11}, B_1, C_1, D)$ also realizes $H(s)$, recall that

$$H(s) = C_{new}(sI - A_{new})^{-1}B_{new} + D_{new} = \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} sI - A_{11} & -A_{12} \\ 0 & sI - A_{22} \end{bmatrix}^{-1} \begin{bmatrix} B_1 \\ 0 \end{bmatrix} + D.$$

The upper left block of the inverse matrix above is $(sI - A_{11})^{-1}$, from which it floows that

$$H(s) = C_1(sI - A_{11})^{-1}B_1 + D,$$

which is the transfer function associated with $\Sigma_{reduced}(A_{11}, B_1, C_1, D)$. □

3 Kalman Decomposition into Observable and Unobervable Subsystems

Similarly, by an appropriate choice of basis, we can exhibit the "observable subsystem" and the "unobservable subsystem" as shown in Figure 16. Note that only the observable subsystem affects the output. We can discard the unobservable subsystem without changing the input-output behavior.

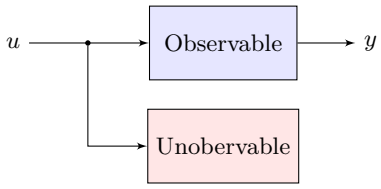

Figure 16: Observable and Unobervable Subsystems.

The following theorem is analogous to Theorem 118:

Theorem 119. *Consider the system $\Sigma(A, B, C, D)$ with transfer function $H(s)$ and let $dim(\mathcal{UO}) = r$. Let $\{t_1, \cdots, t_r\}$ be a basis for $\mathcal{UO}$ and extend this by $\{t_{r+1}, \cdots, t_n\}$ to form a basis for $\mathbb{R}^n$. Then*

$$T = \begin{bmatrix} t_1 & \cdots & t_r & t_{r+1} & \cdots t_n \end{bmatrix}$$

is invertible, so we can do a similarity transformation using the change of variables $Tx_{new} = x$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

(a) Σ_{new} *has the structure*

$$\Sigma_{new} = \left[\begin{array}{c|c} T^{-1}AT & T^{-1}B \\ \hline CT & D \end{array}\right] = \left[\begin{array}{cc|c} A_{11} & A_{12} & B_1 \\ 0 & A_{22} & B_2 \\ \hline 0 & C_2 & D \end{array}\right]$$

where $A_{11} \in \mathbb{R}^{r\times r}$*, and* B *and* C *are partitioned conformably with the partition of* A*.*

(b) The reduced model

$$\Sigma_{reduced}(A_{22}, B_2, C_2, D)$$

is observable and also realizes $H(s)$*.*

As an application of this decomposition, we return to Theorem 113 in the previous section, and complete the proof of its "only if" statement: *if the system* Σ *is unobservable then* A *must have an eigenvector* v *such that* $Cv = 0$*.* We will prove this statement for the decomposed system in Theorem 119; that is will show there exists an eigenvector v_{new} of A_{new} such that $C_{new}v_{new} = 0$. To construct such an eigenvector, pick an arbitrary eigenvalue/eigenvector pair (λ, v_1) for A_{11}:

$$A_{11}v_1 = \lambda v_1.$$

Then,

$$A_{new}\left[\begin{array}{c} v_1 \\ 0 \end{array}\right] = \left[\begin{array}{cc} A_{11} & A_{12} \\ 0 & A_{22} \end{array}\right]\left[\begin{array}{c} v_1 \\ 0 \end{array}\right] = \left[\begin{array}{c} A_{11}v_1 \\ 0 \end{array}\right] = \lambda\left[\begin{array}{c} v_1 \\ 0 \end{array}\right].$$

Moreover,

$$C_{new}\left[\begin{array}{c} v_1 \\ 0 \end{array}\right] = \left[\begin{array}{cc} 0 & C_2 \end{array}\right]\left[\begin{array}{c} v_1 \\ 0 \end{array}\right] = 0.$$

Thus,

$$v_{new} = \left[\begin{array}{c} v_1 \\ 0 \end{array}\right]$$

is an eigenvector of A_{new} that satisfies $C_{new}v_{new} = 0$. It then follows that $v = Tv_{new}$ satisfies

$$Av = (TA_{new}T^{-1})(Tv_{new}) = TA_{new}v_{new} = T\lambda v_{new} = \lambda v$$

and

$$Cv = (C_{new}T^{-1})(Tv_{new}) = C_{new}v_{new} = 0.$$

We thus conclude that A has an eigenvector v such that $Cv = 0$.

4 Kalman Decomposition: Most General Version

We now present the most general version of the Kalman decomposition that is based on both controllability and observability.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Theorem 120. *Consider the realization* $\Sigma(A,B,C,D)$ *of some transfer function* $H(s)$. *Let* $\{t_1,\cdots,t_{n_1}\}$ *be a basis for* $\mathcal{C}\cap\mathcal{UO}$. *Extend this basis by* $\{t_{n_1+1},\cdots,t_{n_2}\}$ *to form a basis for* $\mathcal{C}$, *and by* $\{t_{n_2+1},\cdots,t_{n_3}\}$ *to form a basis for* $\mathcal{UO}$. *Finally, extend the collection of vectors* $\{t_1,\cdots,t_{n_3}\}$ *by* $\{t_{n_3+1},\cdots,t_n\}$ *to complete a basis for* $\mathbb{R}^n$. *Define the invertible matrix*

$$T=\begin{bmatrix} t_1 & \cdots & t_n \end{bmatrix}$$

and let Σ_{new} *be the realization similar to* Σ *obtained by the state-space change of basis* $Tx_{new}=x$.

(a) Σ_{new} *has the structure*

$$\Sigma_{new}=\left[\begin{array}{c|c} T^{-1}AT & T^{-1}B \\ \hline CT & D \end{array}\right] = \left[\begin{array}{cccc|c} A_{11} & A_{12} & A_{13} & A_{14} & B_1 \\ 0 & A_{22} & 0 & A_{24} & B_2 \\ 0 & 0 & A_{33} & A_{34} & 0 \\ 0 & 0 & 0 & A_{44} & 0 \\ \hline 0 & C_2 & 0 & C_4 & D \end{array}\right]$$

where $A_{ii}\in\mathbb{R}^{n_i\times n_i}$ *for* $i=1,\cdots,4$, *and* B *and* C *are partitioned conformably with the partition of* A.

(b) The reduced system

$$\Sigma_{reduced}(A_{22},B_2,C_2,D)$$

is controllable and observable, and it also realizes $H(s)$.

5 Minimal Realizations

Definition 121. A realization Σ is called minimal if it is both controllable and observable.

The Kalman decompositions above showed us how to obtain a minimal realization starting from a nonminimal one. It can be shown further that the state dimension of two minimal realizations of a transfer function have the same state dimension and are related by a similarity transformation:

Theorem 122. *Let* Σ_1 *and* Σ_2 *be two minimal realizations of some transfer function* $H(s)$. *Then*

(a) Σ_1 *and* Σ_2 *have the same state-space dimension.*

(b) Σ_1 *and* Σ_2 *are similar.*

From the above result, the state dimension is independent of the particular minimal realization chosen and depends only on the transfer function $H(s)$. This dimension is called the *McMillan degree* of $H(s)$. In the case of a SISO system, McMillan degree is the order of the denominator polynomial after the transfer function has been simplified to remove any pole-zero cancellations. It corresponds to the number of state variables needed in the minimal realization of the system.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. State feedback

B. Observers

C. Output feedback

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. State Feedback

1 Introduction

In these notes, we will learn how to design controllers for LTI plants in state-space form. The methods we will learn form the basis of modern control system design and are numerically more attractive that transfer-function methods.

2 State Feedback

Consider the system

$$\Sigma \quad \begin{cases} \dot{x}(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t) \end{cases} \qquad x(t_0) = x_0$$

As we've learned, the behavior of the trajectories depends on the *eigenvalues* of A. For instance, the realization is internally stable if and only if all these eigenvalues are in the open left-half complex plane. If the eigenvalues of A have large negative real parts, we expect the impulse response of Σ to decay rapidly to zero. If these eigenvalues are complex, stable, and are close to the imaginary axis, we expect the step response of Σ to resemble a lightly damped sinusoid.

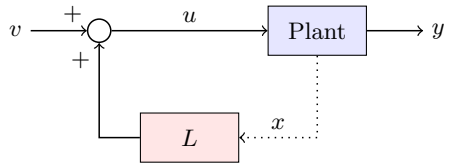

Figure 17: State feedback.

We now investigate the use of feedback to alter the behavior of the system by changing the eigenvalues of A. Suppose we measure the *entire* state of Σ and implement the *state-feedback law*

$$u = Lx + v$$

as illustrated in Figure 17. Here v is some new input which we could use for feedforward (not considered for now) and L is a $m \times n$ matrix where n is the number of states, and m is the number of inputs. In this case, the closed-loop system becomes

$$\Sigma_{cl} \quad \begin{cases} \dot{x}(t) &= [A + BL]\, x(t) + Bv(t) \\ y(t) &= Cx(t) + Du(t) \end{cases} \qquad x(t_0) = x_0$$

The behavior of the system is now governed by the eigenvalues of $A + BL$. The natural question to ask then is whether we can place the eigenvalues of $A_{cl} = A + BL$ as we wish by choice of L.

3 Controllable canonical form

Theorem 123. *Consider the controllable single-input system $\Sigma(A, b, *, *)$. Let*

$$det(sI - A) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

and define $\hat{\Sigma}(\hat{A}, \hat{b}, *, *)$ *by*

$$\hat{A} = \begin{bmatrix} 0 & 1 & \cdots & 0 & 0 \\ 0 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-2} & -\alpha_{n-1} \end{bmatrix} \quad \hat{b} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

Let $M, \hat{M} \in \mathbb{R}^{n \times n}$ *be the controllability matrices of* Σ *and* $\hat{\Sigma}$ *respectively. Then,* Σ *and* $\hat{\Sigma}$ *are similar:*

$$\hat{A} = T^{-1}AT, \quad \hat{b} = T^{-1}b$$

with $T = M\hat{M}^{-1}$.

Proof. First observe that by the Cayley-Hamilton theorem,

$$A^n b = -\sum_{0}^{n-1} \alpha_k A^k b$$

As a consequence,

$$\begin{aligned} AM &= A \begin{bmatrix} b & Ab & \cdots & A^{n-1}b \end{bmatrix} \\ &= \begin{bmatrix} Ab & A^2b & \cdots & A^n b \end{bmatrix} \\ &= M \begin{bmatrix} 0 & 0 & \cdots & 0 & -\alpha_0 \\ 1 & 0 & \cdots & 0 & -\alpha_1 \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & -\alpha_{n-2} \\ 0 & 0 & \cdots & 1 & -\alpha_{n-1} \end{bmatrix} \\ &=: MQ \end{aligned}$$

Since A and $\hat{A}$ have the same characteristic polynomial, we can analogously conclude that

$$\hat{A}\hat{M} = \hat{M}Q$$

Combining these we arrive at

$$\begin{aligned} T^{-1}AT &= \hat{M}M^{-1}AM\hat{M}^{-1} \\ &= \hat{M}Q\hat{M}^{-1} \\ &= \hat{A} \end{aligned}$$

Next, note that $Me_1 = b$ where e_1 is the first unit vector in $\mathbb{R}^n$. We therefore have

$$\begin{aligned} T^{-1}b &= \hat{M}M^{-1}b = \hat{M}e_1 \\ &= \hat{b} \end{aligned}$$

proving the claim. □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

4 Single-input pole placement

We now show that, when (A, B) is controllable, we can arbitrarily place all the eigenvalues of $A+BL$ by choice of state-feedback matrix L. We first focus on single-input single-output systems.

Theorem 124. *Consider the controllable single-input system* $\Sigma(A, b, *, *)$ *and let*

$$det\ (sI - A) = s^n + \alpha_{n-1}s^{n-1} + \cdots + \alpha_1 s + \alpha_0$$

Suppose we are given desired closed-loop pole locations $p_1, \cdots, p_n$*, leading to the desired characteristic polynomial*

$$(s - p_1)(s - p_2)\cdots(s - p_n) = s^n + \gamma_{n-1}s^{n-1} + \cdots + \gamma_1 s + \gamma_0$$

Define the matrices $\hat{A}, \hat{b}$ *in controllable canonical form as in the previous section, and let*

$$\hat{L} = \begin{bmatrix} (\alpha_0 - \gamma_0) & (\alpha_1 - \gamma_1) & \cdots & (\alpha_{n-2} - \gamma_{n-2}) & (\alpha_{n-1} - \gamma_{n-1}) \end{bmatrix}$$

Compute the $n \times n$ *controllability matrices* M *and* $\hat{M}$ *as in the previous section, and define the matrix*

$$L = \hat{L}T^{-1} \quad where \quad T = M\hat{M}^{-1}$$

Then, the closed-loop system under the state-feedback law $u = Lx+v$ *has the desired characteristic polynomial, i.e.,*

$$\det\left[sI - (A + bL)\right] = s^n + \gamma_{n-1}s^{n-1} + \cdots + \gamma_1 s + \gamma_0$$

Proof: The matrix T above transforms Σ to controllable canonical form:

$$\hat{A} = T^{-1}AT, \quad \hat{b} = T^{-1}b$$

Then, we have

$$\hat{A} + \hat{b}\hat{L} = T^{-1}AT + T^{-1}bLT = T^{-1}\left(A + bL\right)T$$

As a result, the eigenvalues of $(A + bL)$ are the same as those of the (similar) matrix $(\hat{A} + \hat{b}\hat{L})$. Equivalently, $(A + bL)$ and $(\hat{A} + \hat{b}\hat{L})$ have the same characteristic polynomial. To find this polynomial, note:

$$\begin{aligned} \hat{A} + \hat{b}\hat{L} &= \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\alpha_0 & -\alpha_1 & \cdots & -\alpha_{n-1} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} (\alpha_0 - \gamma_0) & (\alpha_1 - \gamma_1) & \cdots & (\alpha_{n-1} - \gamma_{n-1}) \end{bmatrix} \\ &= \begin{bmatrix} 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 1 \\ -\gamma_0 & -\gamma_1 & \cdots & -\gamma_{n-1} \end{bmatrix} \end{aligned}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

It is easy to read-off the characteristic polynomial of $\hat{A} + \hat{b}\hat{L}$ because it is still in the special controllable canonical form:

$$det\left[sI - (\hat{A} + \hat{b}\hat{L})\right] = s^n + \gamma_{n-1}s^{n-1} + \cdots + \gamma_1 s + \gamma_0$$

This matches the desired characteristic polynomial, concluding the proof. □

5 Multi-input pole placement

We now turn our attention to multi-input controllable systems. Let Σ be a multi-input controllable realization. We first show that by a preliminary choice of feedback it is possible to make Σ controllable from a *single* input.

Heymann's Lemma. *Let $\Sigma(A, B, *, *)$ be a multi-input controllable system. Let $v \in \mathbb{R}^m$ be such that $Bv \neq 0$ and denote $b = Bv$. Then there exists a matrix $L_0 \in \mathbb{R}^{m\times n}$ such that the single-input system $\Sigma_v(A + BL_0, b, *, *)$ is controllable.*

Armed with this result, we can adress the pole-placement problem for multi-input systems.

Theorem 125. *Let $\Sigma(A, B, *, *)$ be a multi-input controllable realization. Then, by choice of state-feedback gain L we can arbitrarily assign the closed-loop eigenvalues of $A_{cl} = A + BL$.*

Proof: Pick v such that $b = Bv \neq 0$ and let L_0 be as in Heymann's Lemma above. Since $(A + BL_0, b)$ is a single-input controllable pair, we can find L_1 such that $(A + BL_0) + bL_1$ has the desired eigenvalues. Note that

$$(A + BL_0) + bL_1 = (A + BL_0) + (Bv)L_1 = A + B(L_0 + vL_1)$$

Thus, $L = L_0 + vL_1$ assigns the eigenvalue of $A + BL$ as desired.

6 Pole placement and Controllability

We have shown that for controllable systems we can place closed-loop eigenvalues arbitrarily by state feedback. The obvious question is what happens if Σ is *uncontrollable* ?

To address this question the natural tool to employ is the Kalman decomposition. We know that there exists a similarity transformation T such that

$$A = T\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} T^{-1}, \quad B = T\begin{bmatrix} B_1 \\ 0 \end{bmatrix}$$

with (A_{11}, B_1) being controllable. Next observe that

$$\begin{aligned} A_{cl} &= A + BL \\ &= T\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} T^{-1} + T\begin{bmatrix} B_1 \\ 0 \end{bmatrix} L \\ &= T\left(\begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix} + \begin{bmatrix} B_1 \\ 0 \end{bmatrix} LT\right) T^{-1} \end{aligned}$$

If we partition the matrix LT as

$$LT = \begin{bmatrix} L_1 & L_2 \end{bmatrix}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

we obtain

$$A_{cl} = T \begin{bmatrix} A_{11} + B_1 L_1 & A_{12} + B_1 L_2 \\ 0 & A_{22} \end{bmatrix} T^{-1}$$

As a consequence, the eigenvalues of A_{cl} are precisely those of $A_{11} + B_1 L_1$ combined with those of A_{22}. Further, since (A_{11}, B_1) is controllable, we can arbitrarily assign the eigenvalues of $A_{11} + B_1 L_1$ by choice of L_1 (i.e. choice of L). However, regardless of the state-feedback gain L selected, the eigenvalues of A_{22} remain as eigenvalues of A_{cl}. These *uncontrollable* modes cannot be moved by state-feedback. We are simply stuck with these eigenvalues.

We summarize our conclusions in the following theorem.

Theorem 126. *The system $\Sigma(A, B, *, *)$ is controllable if and only if we can arbitrarily assign the closed-loop eigenvalues by state-feedback. The uncontrollable eigenvalues of A are unaffected by state-feedback.*

7 Stabilizability

Although we can't move uncontrollable eigenvalues (i.e., the eigenvalues of A_{22} in the Kalman decomposition above) by feedback, these eigenvalues may be acceptable for the closed-loop system. For example, the following definition describes the situation where the uncontrollable eigenvalues already have negative real parts, so that we can make the closed loop system stable by feedback (i.e., by designing L_1 such that the eigenvalues of $A_{11} + B_1 L_1$ also have negative real parts).

Definition 127. A realization $\Sigma(A, B, *, *)$ is called *stabilizable* if there exists a state-feedback gain L such that $A_{cl} = A + BL$ is stable; equivalently, the uncontrollable eigenvalues have negative real parts.

Note that all controllable realizations are stabilizable, but not vice versa.

8 Matlab Commands

The `place` command computes L so that eigenvalues of $A - BL$ are at specified locations in the complex plane. Note that this command assigns the eigenvalues of A minus BL. Therefore, we must enter $-B$ instead of B to be consistent with our convention where the feedback is $u = Lx$ and the closed-loop matrix is $A + BL$. Alternatively, you can enter B but apply the feedback $u = -Lx$.
Another quirk of the `place` command is that poles must be distinct, e.g.,

```
>> poles = [ -1; -1.000001, -1.0000002];
```

```
>> % start with any state-space realization Sigma(A,B,C)
>> % desired pole locations as a vector
>> poles = [p1; p2; ... ;pn];
>> L = place(A,-B, poles);
>>% then eig(A+BL) = poles
```

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

B. Observers

1 Introduction

In the previous section we studied the *state* feedback $u = Lx + v$. But this requires having access to the entire state vector, i.e. we need sensors that measure all of the states. We will now build a system called an *observer* that provides an estimate $\hat{x}$ of the states using the plant model, knowledge of the input u and measurement of the output y. This is illustrated in Figure 18.

In the subsequent section we will replace the states with their estimates in the feedback law, that is, $u = L\hat{x} + v$.

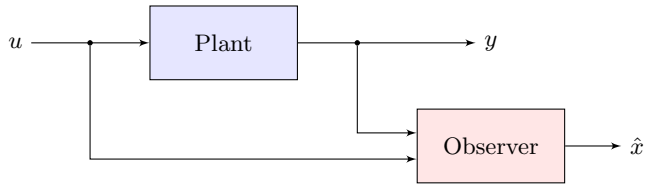

Figure 18: Block Diagram for an Observer.

2 Asymptotic observers

Imagine running a simulation copy of the system model in real time with the same input $u(t)$ applied to the actual system:

$$\begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + Bu(t) \\ \hat{y}(t) &= C\hat{x}(t) + Du(t) \end{aligned} \tag{13}$$

Then the error

$$e := \hat{x} - x$$

between the simulated state $\hat{x}$ and the actual state x satisfies

$$\dot{e}(t) = A\hat{x}(t) - Ax(t) = Ae(t)$$

Thus, if A is Hurwitz, $e(t)$ converges to zero asymptotically and we obtain an accurate estimate of the actual states from the simulation copy.

But what if A is *not* Hurwitz (so $e(t)$ doesn't go to zero) or if its eigenvalues are too close to the imaginary axis (so it takes a long time for $e(t)$ to die out)? In this case can use the discrepancy between the measured output y and the predicted output $\hat{y}$ to correct the simulation copy:

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + K(\hat{y}(t) - y(t))$$

This corrected simulation model is called an *observer* and the matrix K is called the *output injection* matrix. To design K so that $e(t)$ converges to zero, note that e now satisfies

$$\dot{e}(t) = A\hat{x}(t) + K(C\hat{x}(t) - Cx(t)) - Ax(t) = (A + KC)e(t)$$

Thus, if we choose K such that $A + KC$ is Hurwitz, then

$$\lim_{t\to\infty} e(t) = 0$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Furthermore, the eigenvalues of $A + KC$ determine the rate of convergence of $e(t)$.

Can we design K to assign the eigenvalues of $A+KC$ arbitrarily? The answer is greatly simplified if we use duality. The eigenvalues of $A+KC$ are the same as those of its transpose $A^T + C^T K^T$. Denote

$$A_{\text{dual}} = A^T, \quad B_{\text{dual}} = C^T, \quad L_{\text{dual}} = K^T$$

and note that if we can design L_{dual} to assign the eigenvalues of $A_{\text{dual}} + B_{\text{dual}} L_{\text{dual}}$ arbitrarily, that means we can assign the eigenvalues of $A + KC$ arbitrarily with $K = L_{\text{dual}}^T$.

We know we can assign the eigenvalues of $A_{\text{dual}} + B_{\text{dual}} L_{\text{dual}}$ arbitrarily iff $(A_{\text{dual}}, B_{\text{dual}})$ is controllable. We also know $(A_{\text{dual}}, B_{\text{dual}})$ is controllable iff (C, A) is observable. Thus, we reach the following conclusion:

Theorem 128. *We can choose K to assign the eigenvalues of $A + KC$ arbitrarily if and only if (C, A) is observable.*

On the other hand, if (C, A) is not observable but the unobservable eigenvalues already have negative real parts, we can choose K such that $A + KC$ is Hurwitz. Thus, the error e converges to zero even if we can't choose the rate at which it does so. This leads to the notion of *detectability*:

Definition 129. The system $\Sigma(A, *, C, *)$ is called *detectable* if there exists K such that $A + KC$ is Hurwitz.

Note that the system is detectable iff its dual is stabilizable.

3 Example

We will work out a detailed example of using an observer. Our plant is a boat traveling in one dimension.It has two states: position y, and velocity $\dot{y}$. The boat has an outboard motor which provides thrust u. There is a drag force from the water proportional to the velocity of the boat. The drag coefficient is b. The boat has mass $m = 300$Kg, and let $b = 2.5$Newtons-sec/m. The dynamics are

$$m\ddot{y} = -b\dot{y} + u + d \tag{14}$$

where d is a disturbance force due to waves and currents. The waves add a periodic force at two frequencies and the currents add a constant force. So we will write

$$d = 3 + 2\cos(t) - \sin(2t) \text{ Newtons}$$

We do not know d in advance. The throttle u is known, say by measuring the diesel engine torque. Let us say that the boat starts at $y = 0$, and we apply full throttle for 20 minutes, and then idle the engine. So we write

$$u = \begin{cases} 7 & t \in [0, 20) \\ 0 & t \in [20, 30] \end{cases}$$

The throttle u and the disturbance force d are shown in Figure 19. We have available a noisy measurement z of the boat position, say by range finders to known buoys. We write this as

$$z = y + n$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

where n is sensor noise.

Define the state of the boat as

$$x_1 = y, \quad x_2 = \dot{y}$$

We will design an observer that estimates the sate from the measurement of the position y. We ignore the sensor noise n and the disturbance d in designing our observer. We write the boat dynamics in state-space form as

$$\left[\begin{array}{c} \dot{x}_1 \\ \dot{x}_2 \\ \hline y \end{array}\right] = \left[\begin{array}{cc|c} 0 & 1 & 0 \\ 0 & -b/m & 1/m \\ \hline 1 & 0 & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \\ \hline u \end{array}\right]$$

All we have to do is choose stable observe pole locations. The Matlab command `place` will calculate the observer gain K if we use duality: enter A^T instead of A and C^T instead of B. The command will return feedback matrix L, but we use $K = L^T$ as the output injection matrix. Recall the `place` command has a quirk that you cannot give repeated pole locations.

```
>> obs_poles = [-5, -5.01];
>> b = 2.5; m =300;
>> A = [0 1; 0 -b/m];
>> C = [1 0];
>> foo = place(A', - C');
>> K = foo';
% plant model
% 1/(s^2 + 0.5s)  ==>  y'' +0.5y' = u
% viscous damping coefficient is 0.5
A = [0 1; 0 -0.5];
B = [0 ; 1];
C = [1 0];
D = 0;
nx = size(A,1);
plant = ss(A,B,eye(nx),zeros(nx,1));

% input and disturbance
tt = [0:0.1:40]';
%uu = cos(1*tt) - 2*sin(2*tt)+ 3*(tt >10) - 2*(tt>20);
dd = 0.3 + 0.2*cos(1*tt) - 0.5*sin(2*tt); %+ 3*(tt >10) - 2*(tt>20);
uu = ones(401,1)*10;
uu(201:end) = 0; uu(201) = 8; uu(202) = 6; uu(203) = 4; uu(204) = 2;

% simulate plant
xx = lsim(plant, uu+dd, tt);
zz = xx(:,1) + randn(size(tt));
zzdot = [0; diff(zz)];
```

Now we can simulate the observer. This time we put in sensor noise and the disturbance.

Our objective is to estimate the position y and the velocity $\dot{y}$ of the boat.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

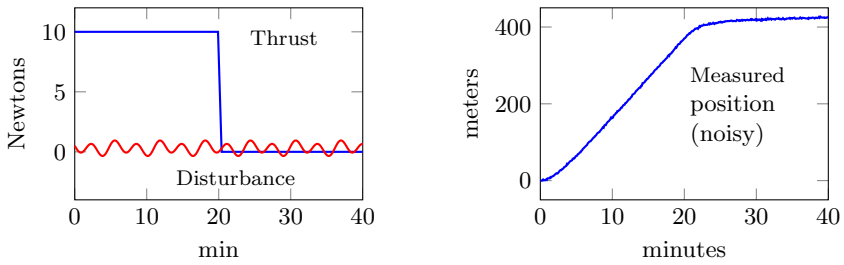

Figure 19: (a) Engine thrust u and disturbance d, (b) measured position of boat.

As a baseline, we can ignore the ship model, and use the position measurement z as an estimate of the ship position x. We can also numerically differentiate z to estimate the ship velocity $\dot{y}$. Of course, differentiating a noisy signal is not advisable, but let us see what we get. The results are shown in Figure 20. The position estimate is noisy but acceptable with an accuracy of $\pm$2m. But the velocity estimate is awful. It is off by 20m/min when the boat is moving, and has an accuracy of $\pm$3m/min when the boat is idling. This approach discards useful information: the model of the boat and the known throttle.

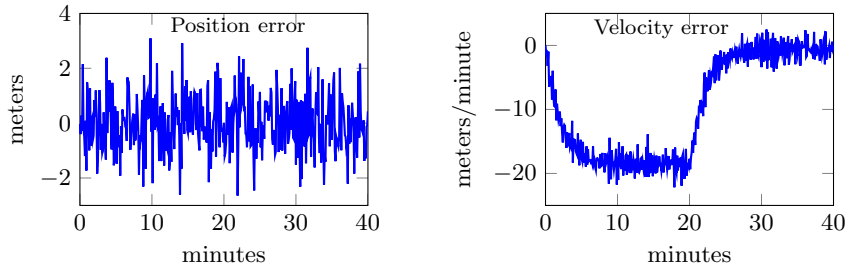

Figure 20: State estimation using position sensor alone: (a) position error, (b) velocity error.

Let us now try an observer to estimate the state. Here, we use both the measured position and the boat model. We try three cases: case A is the least aggressive with a low observer gain, case B has a medium gain, and case C is the most aggressive with observer poles deep into the left half lane and high observer gain. The position and velocity error plots are shown in Figure 22, and the results are summarized in Figure 21. We notice that there is a sweet spot: being either too aggressive or too conservative with observer poles results in poor state estimates. Observers combine measurements and the model to compute state estimates. We have to strike a balance between how good the measurements are (i.e. how big is the sensor nose) versus how good the model is (i.e. how big are the disturbances). Kalman Filtering is a technique that does this in an optimal way, but this is beyond the scope of this class.

Case	observer poles	observer gain K	position error (m)	velocity error (m/min)
A	-0.1, -0.1	$-[0.290.156]$	$\pm$20	$\pm$6
B	-5, -5	$-[5.66.5]$	$\pm$1	$\pm$1
C	-100, -100	$-[199.6991.0]$	$\pm$2.2	$\pm$25

Figure 21: Summary of results: A. low observer gain, B. medium observer gain, C. high observer gain

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

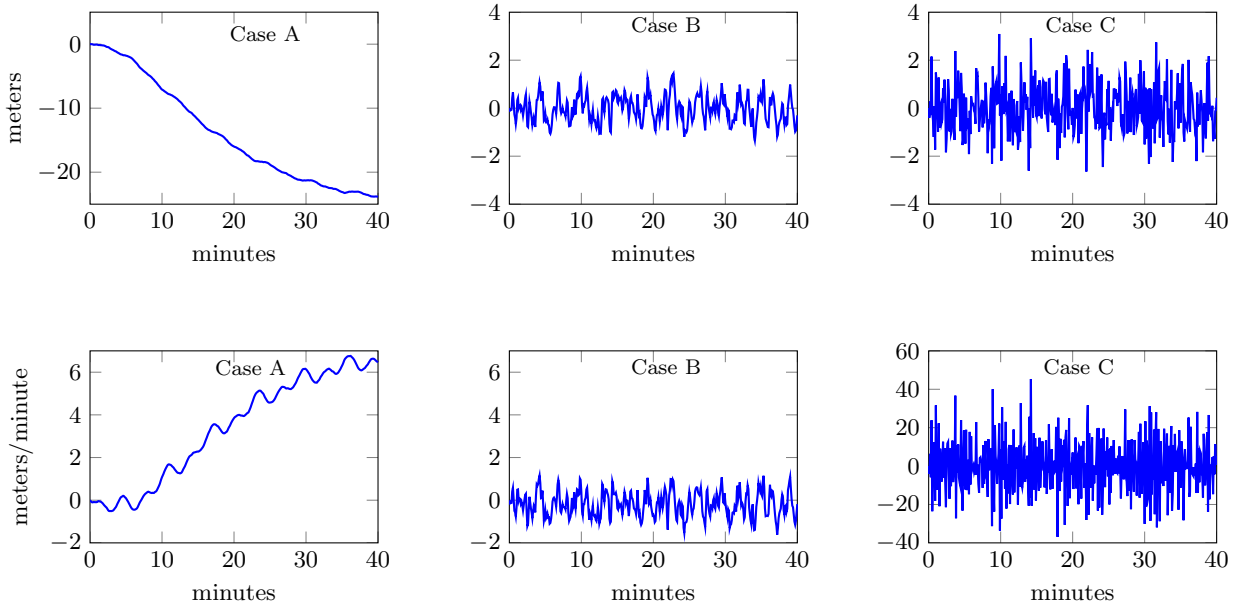

Figure 22: State estimation using observers. Case A: low gain, case B: medium gain, case C: high gain. Top row: position errors, bottom row: velocity errors.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

C. Output Feedback

1 Output Feedback

Previously, our scheme for state feedback stabilization assumed that we had access to the plant state x, as illustrated in Figure 23. In practice, we usually do not have sensors to measure every state variable. But now that we know how to construct an estimate $\hat{x}$ of the plant state x, we can use this estimate for state feedback. This idea is illustrated in Figure 24.

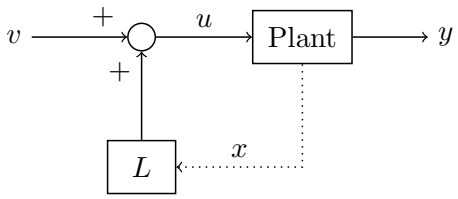

Figure 23: State-Feedback Stabilization

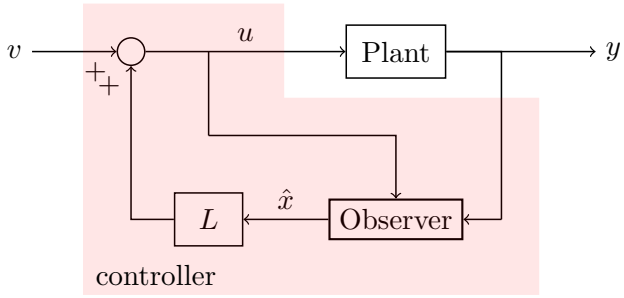

Figure 24: Output Feedback Stabilization

We will now derive state-space equations for the output feedback controller shown in Figure 24. The observer and state feedback equations are:

$$\dot{\hat{x}} = (A + KC)\hat{x} + Bu - K(y - Du)$$
$$u = L\hat{x} + v$$

Combining these we get the controller dynamics:

$$\dot{\hat{x}} = (A + KC + BL + KDL)\hat{x} + (B + KD)v - Ky$$
$$u = L\hat{x} + v$$

Note the controller itself has n states, where n is the number of states in the plant model. Thus, the closed-loop system of Figure 24 has $2n$ states - n plant model states, and n controller states.

2 The Separation Principle

When we use the observer estimate of the state for feedback control, it is not clear whether we have the same stability properties as state feedback. The following central result guarantees that implementing the state feedback law with observer states does indeed maintain stability.

Theorem 130. *(Separation Principle) Consider a LTI plant with state space realization*

$$\left[\begin{array}{c} \dot{x} \\ \hline y \end{array}\right] = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \left[\begin{array}{c} x \\ \hline u \end{array}\right]$$

Suppose we control this plant with the controller

$$\left[\begin{array}{c} \dot{\hat{x}} \\ \hline u \end{array}\right] = \left[\begin{array}{c|cc} A + KC + BL + KDL & B + KD & -K \\ \hline L & I & 0 \end{array}\right] \left[\begin{array}{c} \hat{x} \\ \hline v \\ y \end{array}\right]$$

as illustrated in Figure 24. Then,

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

(a) the closed loop system can be written as

$$\left[\begin{array}{c} \dot{x} \\ \dot{\hat{x}} \\ \hline y \end{array}\right] = \left[\begin{array}{cc|c} A & BL & B \\ -KC & A+KC+BL & B \\ \hline C & DL & D \end{array}\right] \left[\begin{array}{c} x \\ \hat{x} \\ \hline v \end{array}\right]$$

(b) the eigenvalues of

$$A_{\text{cl}} = \left[\begin{array}{cc} A & BL \\ -KC & A+KC+BL \end{array}\right]$$

are those of $A+BL$ and $A+KC$ combined:

$$Spec(A_{\text{cl}}) = Spec(A+BL) \ \cup \ Spec(A+KC)$$

Proof: We simply combine the plant and controller dynamics to write down the state-space equations for the closed-loop system. For this, first notice that

$$\begin{aligned} u &= L\hat{x} + v \\ y &= Cx + Du \end{aligned} \tag{15}$$

Using these equations, we can write

$$\dot{x} = Ax + Bu = Ax + BL\hat{x} + Bv \tag{16}$$

$$\begin{aligned} \dot{\hat{x}} &= (A+KC+BL+KDL)\hat{x} + (B+KD)v - Ky \\ &= (A+KC+BL+KDL)\hat{x} + (B+KD)v - KCx - KDu \\ &= (A+KC+BL+KDL)\hat{x} + (B+KD)v - KCx - KDL\hat{x} - KDv \\ &= -KCx + (A+KC+BL)\hat{x} + Bv \end{aligned} \tag{17}$$

Equations (15) through (17) can be written collectively as in part (a) of the Theorem.

We now examine the eigenvalues of A_{cl}. Recall that for any invertible matrix T, the matrix $T^{-1}A_{\text{cl}}T$ has the same eigenvalues as A_{cl}. If we choose

$$T = \left[\begin{array}{cc} I & 0 \\ I & I \end{array}\right]$$

then

$$T^{-1} = \left[\begin{array}{cc} I & 0 \\ -I & I \end{array}\right]$$

and

$$\begin{aligned} T^{-1}A_{\text{cl}}T &= \left[\begin{array}{cc} I & 0 \\ -I & I \end{array}\right] \left[\begin{array}{cc} A & BL \\ -KC & A+KC+BL \end{array}\right] \left[\begin{array}{cc} I & 0 \\ I & I \end{array}\right] \\ &= \left[\begin{array}{cc} A & BL \\ -(A+KC) & A+KC \end{array}\right] \left[\begin{array}{cc} I & 0 \\ I & I \end{array}\right] = \left[\begin{array}{cc} A+BL & BL \\ 0 & A+KC \end{array}\right] \end{aligned}$$

This final matrix is block upper triangular, so its spectrum is $\text{Spec}(A+BL) \ \cup \ \text{Spec}(A+KC)$, proving part (b). □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The Separation Principle assures us that we can *separately* design the state-feedback gain L and the observer gain K. So as long as we design these gain matrices so that $A + BL$ and $A + KC$ are stable, the output feedback controller will stabilize the plant model. This is true only for LTI plant models. It breaks down for nonlinear systems. It also breaks down for situations where the controller structure is constrained (such as decentralized control). In these cases, the design of the observer and state-feedback gain become intimately coupled.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Finite-horizon optimal control in discrete time

B. Finite-horizon optimal control in continuous time

C. Infinite-horizon LQR

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

A. Finite-Horizon Optimal Control in Discrete Time

1 Optimality

The following principle plays a key role in derivations leading to optimal control policies:

Lemma 131. *(Bellman's Principle of Optimality)*

$$\min_{V_1,V_2}\{F_1(V_1)+F_2(V_1,V_2)\}=\min_{V_1}\left\{F_1(V_1)+\min_{V_2}F_2(V_1,V_2)\right\}$$

Proof. We prove this by showing (LHS $\leq$ RHS) and (LHS $\geq$ RHS), which together imply that the left-hand side must be equal to the right-hand side.

($\leq$) Let $\bar{V}_1,\bar{V}_2$ be the minimizer of RHS, *i.e.,* RHS $= F_1(\bar{V}_1)+F_2(\bar{V}_1,\bar{V}_2)$. Since LHS minimizes $F_1(V_1)+F_2(V_1,V_2)$ over all V_1,V_2, we have LHS $\leq$ RHS.

($\geq$) Let V_1^*,V_2^* be the minimizer of LHS, *i.e.*, LHS $= F_1(V_1^*)+F_2(V_1^*,V_2^*)$. Then,

$$\begin{aligned}\text{LHS} &= F_1(V_1^*)+F_2(V_1^*,V_2^*)\\ &\geq F_1(V_1^*)+\min_{V_2}F_2(V_1^*,V_2)\\ &\geq \min_{V_1}\left\{F_1(V_1)+\min_{V_2}F_2(V_1,V_2)\right\}=\ \text{RHS}\end{aligned}$$

Note that the last inequality holds since the choice V_1^* can't lead to a smaller value than the minimum over all V_1. □

This principle states that, in an optimal sequence of decisions, the remaining subsequence after the first decision is also optimal. That is, if V_1^*,V_2^* is the optimal sequence, then after taking the first step V_1^*, there is no better action than V_2^* for the remaining problem $\min_{V_2}F_2(V_1^*,V_2)$.

2 Finite-horizon optimal control and Linear Quadratic Regulator in Discrete Time

Consider the following discrete time system with a given initial state x_0,

$$x_{k+1}=f(x_k,u_k),\quad x_k\in\mathbb{R}^n,u_k\in\mathbb{R}^m.$$

We want to design a sequence of inputs $u_0,\cdots,u_{N-1}$ to minimize the following cost,

$$\sum_{k=0}^{N-1}g_k(x_k,u_k)+g_N(x_N),$$

where g_i's are user-defined cost functions.

Note that the initial condition x_0 and the input sequence $u_0,u_1,\ldots,u_{N-1}$ uniquely determine the trajectory $x_1,\cdots,x_N$. Therefore, the only independent variables above are the input values $u_0,u_1,\ldots,u_{N-1}$. With this in mind, we rewrite the optimal control problem above as:

$$\begin{aligned}\min_{u_0,\cdots,u_{N-1}}\quad & J(x_0;u_0,u_1,\cdots,u_{N-1}):=\underbrace{\sum_{k=0}^{N-1}g_k(x_k,u_k)}_{\text{stage cost}}+\underbrace{g_N(x_N)}_{\text{terminal cost}}\\ s.t.\quad & x_{k+1}=f(x_k,u_k).\end{aligned}\tag{18}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

The first term in the objective function is called the stage cost, while the second term is called the terminal cost which only depends on the final state x_N. Moreover, we denote the minimum cost of the problem as $V_0(x_0)$, the so-called "value function", *i.e.*,

$$V_0(x_0) = \min_{u_0,\cdots,u_{N-1}} J(x_0; u_0, u_1, \cdots, u_{N-1}).$$

An important special case of (1) is the linear quadratic regulator.

Dicrete-time Linear Quadratic Regulator (LQR). Consider the special case of:

- Linear system dynamics: $x_{k+1} = Ax_k + Bu_k$
- Quadratic stage cost independent of k: $g_k(x,u) = g(x,u) = x^\top Qx + u^\top Ru, Q \succeq 0, R \succ 0$
- Quadratic terminal cost: $g_N(x) = x^\top Sx, S \succeq 0$.

That is, the LQR is a specialization of the general problem (18) to:

$$\begin{aligned} \min_{u_0,\cdots,u_{N-1}} \quad & J(x_0; u_0, \cdots, u_{N-1}) = \sum_{k=0}^{N-1} (x_k^\top Q x_k + u_k^\top R u_k) + x_N^\top S x_N \\ s.t \quad & x_{k+1} = Ax_k + Bu_k. \end{aligned} \tag{19}$$

Example 132. Suppose we want to bring X_N close to origin without expending too much control effort, and we are not worried about the trajectory $x_1, \cdots, x_{N-1}$ prior to the Nth time step. Then, we can select the cost function as

$$J = \sum_{k=0}^{N-1} \gamma ||u_k||^2 + ||x_N||^2 \text{ with } \gamma \gg 1,$$

which means $Q = 0, R = \gamma I, S = I$ in the general formulation (19). A larger value of γ places more priority on spending the least control effort; a small value of γ aims to bring x_N closer to the origin at the cost of more control effort.

3 Solutions of finite-horizon optimal control problem and LQR in discrete time

We first apply Lemma 131 to demonstrate the process for solving the general problem (18); then we specialize to the LQR problem (19) where the solution becomes computationally tractable.

Recall that

$$\begin{aligned} V_0(x_0) &= \min_{u_0,u_1,\cdots,u_{N-1}} \left\{ \sum_{k=0}^{N-1} g_k(x_k,u_k) + g_N(x_N) \right\} \\ &= \underbrace{u_0}_{:=V_1}, \underbrace{\min_{u_1,\cdots,u_{N-1}}}_{:=V_2} \left\{ \underbrace{g_0(x_0,u_0)}_{:=F_1(V_1)} + \underbrace{\sum_{k=1}^{N-1} g_k(x_k,u_k) + g_N(x_N)}_{:=F_2(V_1,V_2)} \right\} \\ &= \min_{u_0} \left\{ g_0(x_0,u_0) + \underbrace{\min_{u_1,\cdots,u_{N-1}} \sum_{k=1}^{N-1} g_k(x_k,u_k) + g_N(x_N)}_{:=V_1(x_1) \quad \text{"cost to go"}} \right\} \\ &= \min_{u_0} \; g_0(x_0,u_0) + V_1(x_1), \end{aligned}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

where the third equality is due to Lemma 131. Note that $V_1(x_1)$ is implicitly a function of u_0 since $x_1 = f(x_0, u_0)$. Applying 131 recursively, we get

$$V_{k-1}(x_{k-1}) = \min_{u_{k-1}} \{g_{k-1}(x_{k-1}, u_{k-1}) + V_k(x_k)\}, \ k = 1, \cdots, N \tag{20}$$

$$V_N(x_N) = g_N(x_N), \tag{21}$$

which are called the Bellman Equations. Here V_k is called the cost-to-go function from the kth time instant, and it depends on u_{k-1} as $V_k(x_k) = V_k(f(x_{k-1}, u_{k-1}))$. Thus, if we know the cost-to-go function V_k, then we can in principle find the minimizer u_{k-1} in terms of x_{k-1} for (20), and substitute it to find the function V_{k-1}. Since the cost-to-go V_N at the final time is equal to the terminal cost g_N, which is known, we can start the process at $k = N$ and solve the Bellman equations backwards from $k = N$ to 1. In the end, we obtain the value function $V_0(x_0)$ and, along the way, we generate the optimal control input u_{k_1} as a function of x_{k-1}, $k = N, N-1, \cdots, 1$.

In general, finding an analytical solution to the optimization problem (20) may be impossible. However, when we specialize to the LQR problem, the function being minimized in (20) is quadratic in x_{k-1} and u_{k-1}, and it is possible to derive an analytic solution.

Solution of LQR. The Bellman equations (20) and (21) for LQR become

$$V_N(x_N) = x_N^\top S x_N$$
$$V_{k-1}(x_{k-1}) = \min_{u_{k-1}} \left\{ x_{k-1}^\top Q x_{k-1} + u_{k-1}^\top R u_{k-1} + V_k(x_k) \right\}, \ k = 1, \cdots, N$$

For $k = N$,

$$V_{N-1}(x_{N-1}) = \min_{u_{N-1}} \ x_{N-1}^\top Q x_{N-1} + u_{N-1}^\top R u_{N-1} + x_N^\top S x_N \tag{22}$$

$$= \min_{u_{N-1}} \ x_{N-1}^\top Q x_{N-1} + u_{N-1}^\top R u_{N-1} + (Ax_{N-1} + Bu_{N-1})^\top S (Ax_{N-1} + Bu_{N-1})$$

$$= \min_{u_{N-1}} \begin{bmatrix} x_{N-1} \\ u_{N-1} \end{bmatrix}^\top \begin{bmatrix} Q + A^\top S A & A^\top S B \\ B^\top S A & R + B^\top S B \end{bmatrix} \begin{bmatrix} x_{N-1} \\ u_{N-1} \end{bmatrix} \tag{23}$$

To solve the minimization problem (23), we first note the following two facts and provide a lemma.

- $\begin{bmatrix} Q + A^\top S A & A^\top S B \\ B^\top S A & R + B^\top S B \end{bmatrix} \succeq 0$ since (22) ≥ 0.
- $R + B^\top S B \succ 0$ since $R \succ 0$.

Lemma 133. *Suppose* $\begin{bmatrix} K & L^\top \\ L & M \end{bmatrix} \succeq 0$ *and* $M \succ 0$. *Then,*

(a) $\arg\min_u \begin{bmatrix} x \\ u \end{bmatrix}^\top \begin{bmatrix} K & L^\top \\ L & M \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = -M^{-1} L x.$

(b) $\min_u \begin{bmatrix} x \\ u \end{bmatrix}^\top \begin{bmatrix} K & L^\top \\ L & M \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = x^\top (K - L^\top M^{-1} L) x.$

(c) $K - L^\top M^{-1} L \succeq 0.$

Proof. The proof is left as a homework problem. □

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Applying Lemma 133 to (23), we get

$$\begin{aligned} u_{N-1} &\overset{(a)}{=} -(R+B^\top SB)^{-1}B^\top SAx \\ V_{N-1}(x_{N-1}) &\overset{(b)}{=} x_{N-1}^\top \underbrace{\left(Q+A^\top SA - A^\top SB(R+B^\top SB)^{-1}B^\top SA\right)}_{:=P_{N-1}} x_{N-1}. \end{aligned}$$

Note that $P_{N-1} \overset{(c)}{\succeq} 0$. Therefore, $V_{N-1}(x_{N-1}) = x_{N-1}^\top P_{N-1} x_{N-1}$ is still in the quadratic form which allows us to similarly consider for $k = N-1$,

$$\begin{aligned} V_{N-2}(x_{N-2}) &= \min_{u_{N-2}} \; x_{N-2}^\top Q x_{N-2} + u_{N-2}^\top R u_{N-2} + V_{N-1}(x_{N-1}) \\ &= \min_{u_{N-2}} \; x_{N-2}^\top Q x_{N-2} + u_{N-2}^\top R u_{N-2} + x_{N-1}^\top P_{N-1} x_{N-1} \\ &= \min_{u_{N-2}} \begin{bmatrix} x_{N-2} \\ u_{N-2} \end{bmatrix}^\top \begin{bmatrix} Q+A^\top P_{N-1}A & A^\top P_{N-1}B \\ B^\top P_{N-1}A & R+B^\top P_{N-1}B \end{bmatrix} \begin{bmatrix} x_{N-2} \\ u_{N-2} \end{bmatrix}. \end{aligned}$$

We can once again apply Lemma 133 and continue recursively. Thus, we have for $k = N, \cdots, 1$,

$$V_{k-1}(x_{k-1}) = \min_{u_{k-1}} \; x_{k-1}^\top Q x_{k-1} + u_{k-1}^\top R u_{k-1} + (Ax_{k-1}+Bu_{k-1})^\top P_k (Ax_{k-1}+Bu_{k-1}),$$

and Lemma 133 yields

$$u_{k-1} = -(R+B^\top P_k B)^{-1}B^\top P_k A x_{k-1} \tag{24}$$

$$V_{k-1}(x_{k-1}) = x_{k-1}^\top \underbrace{\left(Q+A^\top P_k A - A^\top P_k B(R+B^T P_k B)^{-1}B^\top P_k A\right)}_{:=P_{k-1}} x_{k-1}. \tag{25}$$

We can rewrite the last equation as an iteration on matrices P_k:

$$\begin{aligned} P_N &= S \\ P_{k-1} &= Q + A^\top P_k A - A^\top P_k B(R+B^\top P_k B)^{-1}B^\top P_k A, \quad k = N, \cdots, 1. \end{aligned} \tag{26}$$

Summary of the steps for solving LQR:

- Starting with $P_N = S$, solve (26) backward for $N, \cdots, 1$ to find $P_{N-1}, P_{N-2}, \cdots, P_0$.
- Substitute $P_1, \cdots, P_N$ in (24) to find $u_0, \cdots, u_{k-1}$
- The value function is $V_0(x_0) = x_0^\top P_0 x_0$.

Note that since P_k depends on time instant k, (24) is a time-varying state feedback law.

Example 134. Suppose we select $Q = 0$, $R = \gamma$, $S = I$ as in Example 2 and time horizon $N = 3$:

$$J = \gamma(u_0^2 + u_1^2 + u_2^2) + ||x_3||^2.$$

Suppose further the linear system is given by:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

What is the optimal sequence u_0, u_1, u_2 to minimize the cost J above? To solve this LQR problem, we first determine the P_k's:

$$P_3 = S = I$$

$$P_2 = 0 + A^\top P_3 A - A^\top P_3 B(\gamma + B^\top P_3 B)^{-1} \underbrace{B^\top P_3 A}_{=0} = A^\top P_3 A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P_1 = 0 + A^\top P_2 A - A^\top P_2 B(R + B^\top P_2 B)^{-1} \underbrace{B^\top P_2 A}_{=0} = A^\top P_2 A = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$P_0 = 0 + A^\top P_1 A - A^\top P_1 B(R + B^\top P_1 B)^{-1} \underbrace{B^\top P_1 A}_{=0} = A^\top P_1 A = 0.$$

Thus, the value function is $V_0(x_0) = x_0^\top P_0 x_0 = 0$ and the optimal control inputs are:

$$u_0 = -(R + B^\top P_1 B)^{-1} \underbrace{B^\top P_1 A}_{=0} x_0 = 0$$

$$u_1 = -(R + B^\top P_2 B)^{-1} \underbrace{B^\top P_2 A}_{=0} x_1 = 0$$

$$u_2 = -(R + B^\top P_3 B)^{-1} \underbrace{B^\top P_3 A}_{=0} x_0 = 0$$

for any initial condition x_0. That is, the best control action in this example is to do nothing! This is not surprising if we observe that $A^3 = 0$, which implies $x_3 = 0$ from any x_0 without inputs. Thus, the cost function is zero with $u_0 = u_1 = u_2 = 0$, and applying nonzero inputs only increases the cost.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

B. Finite-Horizon Optimal Control in Continuous Time

1 Finite-horizon optimal control and Linear Quadratic Regulator in Continuous Time

We now consider the continuous time system

$$\dot{x} = f(x(t), u(t)), \quad x(t) \in \mathbb{R}^n, u(t) \in \mathbb{R}^m,$$

with initial state $x(0) = x_0$. We want to design $u : [0, T] \mapsto \mathbb{R}^m$ to minimize

$$\int_0^T g(x(t), u(t))\ dt + \sigma(x(T)),$$

where $g(\cdot, \cdot)$ and $\sigma(\cdot)$ are user-defined cost functions. Note that we could allow g to also depend on time, as we did in the discrete time case, but we drop the time dependency for simplicity.

Assuming the differential equation above satisfies conditions that guarantee uniqueness of solutions, the cost defined above is a function of the initial condition x_0 and input signal u only, as the state trajectory x is determined by those. Thus, we rewrite the problem as:

$$\begin{aligned} \min_u \ & J(x_0; u) := \underbrace{\int_0^T g(x(t), u(t))\ dt}_{\text{stage cost}} + \underbrace{\sigma(x(T))}_{\text{terminal cost}} \\ s.t. \ & \dot{x} = f(x(t), u(t)) \\ & x(0) = x_0. \end{aligned} \tag{27}$$

As in discrete time, the first term in the objective function is called the stage cost, while the second term is called the terminal cost.

Similar to the discrete-time case, we define the value function as $V(0, x_0 := \min_u J(x_0, u)$.

Likewise, for arbitrary $s \in [0, T]$, we define the "cost to go" as

$$V(s, x(s)) = \min_{u:[s,T] \mapsto \mathbb{R}^m} \int_s^T g(x(t), u(t))\ dt + \sigma(x(T)),$$

An important special case of (1) is the linear quadratic regulator.

Continuous-time Linear Quadratic Regulator (LQR). Consider the special case of:

- Linear system dynamics: $\dot{x} = Ax + Bu$
- Quadratic stage cost independent of k: $g(x, u) = x^\top Q x + u^\top R u, Q \succeq 0, R \succ 0$
- Quadratic terminal cost: $\sigma(x) = x^\top S x, S \succeq 0$.

Thus, LQR is a specialization of the general problem (27) to

$$\begin{aligned} \min_u \ & J(x_0; u) = \int_0^\top (x(t)^\top Q x(t) + u(t)^\top R u(t))\ dt + x(T)^\top S x(T) \\ s.t \ & \dot{x} = Ax + Bu \\ & x(0) = x_0. \end{aligned} \tag{28}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

2 Solutions of finite-horizon optimal control problem and LQR in continuous time

We now adapt the derivation of the Bellman equations from discrete to continuous time:

Pick an arbitrary $s \in [0,T]$ and $h > 0$ such that $s+h \leq T$. Let $u_{[s,s+h]}$, $u_{[s+h,T]}$, $u_{[s,T]}$ denote the snippets of the signal $u : [0,T] \mapsto \mathbb{R}^m$ on the time segments indicated by the subscript. Then,

$$\begin{aligned}
V(s,x(s)) &= \min_{u_{[s,T]}} \int_s^T g(x(t),u(t))\ dt + \sigma(x(T)) \\
&= \min_{u_{[s,T]}} \underbrace{\int_s^{s+h} g(x(t),u(t))\ dt}_{\text{depends on } u_{[s,s+h]}} + \underbrace{\int_{s+h}^T g(x(t),u(t))\ dt + \sigma(x(T))}_{\text{depends on } u_{[s,s+h]} \text{ and } u_{[s+h,T]}} \\
&= \min_{u_{[s,T]}} \left\{ \int_s^{s+h} g(x(t),u(t))\ dt + \min_{u \in [s+h,T]} \int_{s+h}^T g(x(t),u(t))\ dt + \sigma(x(T)) \right\} \\
&= \min_{u_{[s,s+h]}} \left\{ \int_s^{s+h} g(x(t),u(t))\ dt + V(s+h,x(s+h)) \right\}
\end{aligned}$$

where the third equality follows from the principle stated in Lemma 131.

For the two terms on the right hand side, we do a Taylor expansion around $h = 0$ as follows:

$$\begin{aligned}
V(s+h,x(s+h)) &= V(s,x(s)) + \frac{d}{ds}V(s,x(s))h + O(h^2) \\
\int_s^{s+h} g(x(t),u(t))\ dt &= g(x(s),u(s))h + O(h^2).
\end{aligned}$$

Then,

$$\begin{aligned}
V(s,x(s)) &= \min_{u_{[s,s+h]}} \left\{ \int_s^{s+h} g(x(t),u(t))\ dt + V(s+h,x(s+h)) \right\} \\
&= \min_{u_{[s,s+h]}} \left\{ V(s,x(s)) + \left(\frac{d}{ds}V(s,x(s)) + g(x(s),u(s)) \right) h + O(h^2) \right\}.
\end{aligned}$$

Subtracting $V(s,x(s))$ from both sides, we get

$$0 = \min_{u_{[s,s+h]}} \left\{ \left(\frac{d}{ds}V(s,x(s)) + g(x(s),u(s)) \right) h + O(h^2) \right\}.$$

Since this holds for arbitrarily small h, we conclude

$$\begin{aligned}
0 &= \min_{u(s)} \left\{ \frac{d}{ds}V(s,x(s)) + g(x(s),u(s)) \right\} \\
&= \min_{u(s)} \left\{ \nabla_s V(s,x(s)) + \nabla_x V(s,x(s))^\top f(x(s),u(s)) + g(x(s),u(s)) \right\},
\end{aligned}$$

where the second equality follows from the chain rule. Dropping the argument s to make the equation more concise, we get

$$0 = \min_u \left\{ \nabla_s V(s,x) + \nabla_x V(s,x)^\top f(x,u) + g(x,u) \right\}.$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Since $\nabla_s V(s,x)$ does not depend on u, we move it to the left-hand side and obtain the Bellman equations for the case of continuous time:

$$-\nabla_s V(s,x) = \min_u \Big\{ g(x,u) + \nabla_x V(s,x)^\top f(x,u) \Big\} \tag{29}$$

$$V(T,x) = \sigma(x), \tag{30}$$

where the last equality follows because the cost-to-go at time T is the terminal cost.

Compare this with the Bellman equations (20)-(21) in discrete time. Instead of the recursion in (20), we now have the partial differential equation (29), which is impossible to solve analytically in general. It does, however, admit an explicit solution in the special case of LQR:

Solution of LQR. When $f(x,u) = Ax + Bu$, $g(x,u) = x^\top Qx + u^\top Ru$, $\sigma(x) = x^\top Sx$, the Bellman equations (29)-(30) become

$$-\nabla_t V(t,x) = \min_u \Big\{ x^\top Qx + u^\top Ru + \nabla_x V(t,x)^\top (Ax + Bu) \Big\}$$
$$V(T,x) = x^\top Sx.$$

Since the boundary value is $V(T,x) = x^\top Sx$, we will look for a quadratic solution $V(t,x) = x^\top P(t)x$, where $P(t)$ is symmetric and $P(T) = S$, and show that such a solution indeed exists. Then, $-\nabla_t V(t,x) = -x^\top \dot{P}(t)x$, $\nabla_x V(t,x) = 2P(t)x$, and the first equation above becomes

$$\begin{aligned}
-x^\top \dot{P}(t)x &= \min_u \Big\{ x^\top Qx + u^\top Ru + 2x^\top P(t)(Ax + Bu) \Big\} \\
&= \min_u \Big\{ x^\top Qx + u^\top Ru + x^\top (P(t)A + A^\top P(t))x + x^\top P(t)Bu + u^\top B^\top P(t)x \Big\} \\
&= \min_u \begin{bmatrix} x \\ u \end{bmatrix}^\top \begin{bmatrix} PA + A^\top P + Q & PB \\ B^\top P & R \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} \\
&= x^\top (PA + A^\top P + Q - PBR^{-1}B^\top P)x,
\end{aligned}$$

where the second equality holds because $x^\top P(t)Ax$ and $x^\top P(t)Bu$ are scalars and, thus, equal to their transposes $x^\top A^\top P(t)x$ and $u^\top B^\top P(t)x$, respectively. This allows us to write the quadratic expression in the third equation with a symmetric matrix. The fourth equality is by Lemma 133(b) and, since it must hold for all x, we obtain[4] the matrix differential equation

$$-\dot{P}(t) = P(t)A + A^\top P(t) + Q - P(t)BR^{-1}B^\top P(t), \quad P(T) = S, \tag{31}$$

which must be solved backwards to obtain $P(t), t \in [0,T]$. This is called the Riccati Differential Equation. Finally, note from Lemma 133(a) that the minimization problem above is solved by

$$u = -R^{-1}B^\top P(t)x.$$

Summary of the solution of LQR:

- Solve the matrix Riccati Differential Equation (31) backwards to obtain $P(t), t \in [0,T]$.
- The optimal control is $u(t) = -R^{-1}B^\top P(t)x(t)$.

[4]Here we use the implication $x^\top Mx = 0 \, \forall x \Rightarrow M = 0$. To see why this holds, first let $x = e_i$, the ith unit vector, and note $e_i^\top M e_i = M_{ii} = 0$. Thus, all diagonal entries of M must be zero. Next, substitute $x = e_i - e_j$ and use symmetry of M to conclude $m_{ij} = m_{ji} = 0$ when $i \neq j$. Thus, the off-diagonal entries of M must also be zero.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

- The value function is $V(0, x_0) = x_0^\top P(0) x_0$.

Since $P(t)$ depends on time, the optimal control u is a time-varying state feedback law.

3 Solving the Riccati Differential Equation (RDE)

In general, the Riccati Differential Equation (31), abbreviated as RDE, is solved numerically. Below we first study two simple examples where the solution can obtained analytically. Next we show that, although the RDE is nonlinear in P, its solution can be obtained from the solution of an auxiliary linear matrix differential equation.

Example 135. Suppose $Q = S = 0$; that is, the cost function is $\int_0^T u^\top R u$. It is straightforward to see that the optimal control is $u \equiv 0$, since there is no cost on the state. We can confirm this by solving the RDE:

$$\begin{aligned} -\dot{P}(t) &= P(t)A + A^\top P(t) - P(t)BR^{-1}B^\top P(t) \\ P(T) &= 0, \end{aligned}$$

whose solution is indeed $P(t) = 0$, $t \in [0, T]$, and the optimal control is $u = -R^{-1}B^\top P(t)x = 0$.

Example 136. Consider the LQR problem where $A = 0, B = 1, Q = 0, R = \gamma, S = 1$; that is,

$$\begin{aligned} \dot{x} &= u \\ J(x_0, u) &= \int_0^1 \gamma u(t)^2 \, dt + x(1)^2. \end{aligned}$$

The Riccati Differential Equation is $-\dot{P} = -P^2/\gamma, P(1) = 1$, which can be solved as follows:

$$\begin{aligned} & \frac{dP}{d\tau} = \frac{P^2}{\gamma} \\ \Rightarrow\ & \frac{dP}{P^2} = \frac{d\tau}{\gamma} \\ \Rightarrow\ & \int_{P(t)}^{P(1)} \frac{dP}{P^2} = \int_t^1 \frac{d\tau}{\gamma} \\ \Rightarrow\ & \frac{-1}{P}\Big|_{P(t)}^{P(1)} = \frac{1}{P(t)} - \frac{1}{P(1)} = \frac{1-t}{\gamma} \\ \Rightarrow\ & \frac{1}{P(t)} = 1 + \frac{1-t}{\gamma} \quad \Rightarrow \quad P(t) = \frac{\gamma}{\gamma + 1 - t}. \end{aligned}$$

The optimal control u is then $u(t) = -\frac{1}{\gamma+1-t}x(t)$ and the closed-loop system is $\dot{x}(t) = -\frac{1}{\gamma+1-t}x(t)$, which gives the trajectory

$$x(t) = e^{-\int_0^t \frac{1}{\gamma+1-\tau} d\tau} x(0) = \frac{\gamma + 1 - t}{\gamma + 1} x(0).$$

Note that, as $\gamma \to 0$, $x(1) = \frac{\gamma}{1+\gamma}x(0) \to 0$. Since small γ means control is "cheap," we are able to take aggressive control actions to bring the terminal cost close to zero. If, on the other hand, γ is large, the priority is on spending the control effort parsimoniously at the cost of a larger terminal cost. Indeed, if we let $\gamma \to \infty$, then $u(t) = -\frac{1}{\gamma+1-t}x(t) \approx 0$ and $x(1) = \frac{\gamma}{1+\gamma}x(0) \approx x(0)$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Making the RDE linear. Define the auxiliary matrix differential equation:

$$\dot{X}(t) = \Big(A - BR^{-1}B^\top P(t)\Big)\, X(t), \quad X(T) = I. \tag{32}$$

Then, define $Y(t) = P(t)X(t)$ which satisfies $Y(T) = P(T)X(T) = S$ and

$$\begin{aligned}
\dot{Y} &= \dot{P}X + P\dot{X} \\
&= (-PA - A^\top P - Q + PBR^{-1}B^\top P)X + P(A - BR^{-1}B^\top P)X \\
&= -A^\top PX - QX \\
&= -A^\top Y - QX,
\end{aligned}$$

where we substituted $PX = Y$ in the last step. With the same substitution, we rewrite (32) as:

$$\dot{X} = AX - BR^{-1}B^\top Y$$

The two differential equations can be rewritten as the following linear matrix differential equation:

$$\begin{bmatrix} \dot{X}(t) \\ \dot{Y}(t) \end{bmatrix} = \begin{bmatrix} A & -BR^{-1}B^\top \\ -Q & -A^\top \end{bmatrix} \begin{bmatrix} X(t) \\ Y(t) \end{bmatrix}, \quad \begin{bmatrix} X(T) \\ Y(T) \end{bmatrix} = \begin{bmatrix} I \\ S \end{bmatrix}$$

Therefore, we can solve for $X(t)$ and $Y(t)$ from this linear equation, and obtain the solution of the RDE from

$$P(t) = Y(t)X(t)^{-1}.$$

Note that $X(t)$ is invertible since $X(t) = \Phi(t,T)X(T) = \Phi(t,T)$, where $\Phi(t,T)$ is the state transition matrix of (32), which is indeed invertible.

Since linear differential equations admit well defined solutions, the derivation in this section implies that the nonlinear RDE (31) also has a well-defined solution $P(t)$.

Solving the RDE forward in time. Note that the RDE (31) must be solved backwards in time. Alternatively we can define

$$\Pi(t) := P(T - t),$$

and observe that

$$\Pi(0) = P(T), \quad \dot{\Pi}(t) = \frac{d}{dt}P(T-t) = -\dot{P}(T-t).$$

It then follows from (31) that

$$\begin{aligned}
\dot{\Pi}(t) &= A^\top P(T-t) + P(T-t)A + Q - P(T-t)BR^{-1}B^\top P(T-t) \\
&= A^\top \Pi(t) + \Pi(t)A + Q - \Pi(t)BR^{-1}B^\top \Pi(t) \\
\Pi(0) &= P(T) = S.
\end{aligned} \tag{33}$$

We can solve this equation forward in time and obtain the value function from

$$V(0, x_0) = x_0^\top P(0) x_0 = x_0^\top \Pi(T) x_0.$$

Since the solution of the original RDE is $P(t) = \Pi(T-t)$, the optimal control can be written as

$$u(t) = -R^{-1}B^\top P(t)x(t) = -R^{-1}B^\top \Pi(T-t)x(t).$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

C. Infinite-Horizon LQR

1 Infinite-horizon LQR in Continuous Time

The infinite-horizon LQR problem is

$$\begin{aligned} \min_u \quad & J_\infty(x_0;u) = \int_0^\infty (x(t)^\top Q x(t) + u(t)^\top R u(t))\, dt \\ s.t \quad & \dot{x}(t) = Ax(t) + Bu(t) \\ & x(0) = x_0 \end{aligned} \tag{34}$$

Note that there is no terminal cost, since there is no terminal time. We will formulate a solution to this problem by examining the finite horizon problem with no terminal cost (*i.e.*, $S = 0$) in the limit as $T \to \infty$. Apply the forward RDE (33) with $S = 0$:

$$\dot{\Pi}(t) = A^\top \Pi(t) + \Pi(t)A + Q - \Pi(t)BR^{-1}B^\top \Pi(t), \quad \Pi(0) = 0, \tag{35}$$

and recall that the value function is

$$V_T(x_0) = x_0^\top \Pi(T) x_0.$$

We added the subscript T to emphasize the dependence of the value function on the time horizon. We first observe that increasing the horizon cannot decrease the value function:

Lemma 137. *Consider the finite horizon LQR problem (28) with $S = 0$, $Q \succeq 0$, $R \succ 0$. Then,*

(a) $V_T(x_0)$ is nondecreasing in T.

(b) $V_T(x_0) \leq J_\infty(x_0, u)$ for all T, u.

Proof.

(a) Let $\bar{u}$ be the minimizer of $J_T(x_0, u)$ and let $T' \leq T$. Then,

$$V_T(x_0) = \underbrace{\int_0^{T'} (x^\top Q x + \bar{u}^\top R \bar{u}) dt}_{\geq V_{T'}(x_0)} + \underbrace{\int_{T'}^{T} (x^\top Q x + \bar{u}^\top R \bar{u}) dt}_{\geq 0}$$

Thus, $V_T(x_0) \geq V_{T'}(x_0)$.

(b) Similarly, regardless of the choice of u,

$$J_\infty(x_0, u) = \underbrace{\int_0^{T} (x^\top Q x + u^\top R u) dt}_{\geq V_T(x_0)} + \underbrace{\int_{T}^{\infty} (x^\top Q x + u^\top R u) dt}_{\geq 0}.$$

□

Now, if we can show that $J_\infty(x_0, u) < \infty$ for some input function u, then we can conclude from part (b) of Lemma 137 above that $V_T(x_0)$ is bounded in T. The next lemma explicates when an input u exists such that $J_\infty(x_0, u) < \infty$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Lemma 138. *If (A, B) is stabilizable, then $\exists u$ s.t. $J_\infty(x_0, u) < \infty$.*

Proof. Recall that, when (A, B) is stabilizable, we can find a state feedback matrix L such that $A + BL$ is Hurwitz (i.e., it has eigenvalues with negative real parts). Then, the input $u(t) = Lx(t)$ ensures that $x(t)$ converges to zero exponentially, and so does $u(t) = Lx(t)$ itself. Thus, $x(t)^\top Qx(t)+u(t)^\top Ru(t) \to 0$ exponentially and its integral exists, *i.e.*, $J_\infty(x_0, u) < \infty$. □

Note that stabilizability is not a superfluous assumption. The following example demonstrates that no input can make $J_\infty(x_0, u)$ bounded if the system is not stabilizable.

Example 139. Consider the following system

$$\dot{x}_1 = ax_1$$
$$\dot{x}_2 = u,$$

which is uncontrollable. It is stabilizable if $a < 0$ and not stabilizable if $a \geq 0$.

Let $Q = I$ and $R = 1$. Then

$$J_\infty = \int_0^\infty (x_1^2 + x_2^2 + u^2)dt \geq \int_0^\infty x_1^2 \, dt = \int_0^\infty \left(e^{at}x_1(0)\right)^2,$$

which is unbounded whenever $x_1(0) \neq 0$ if the system is not stabilizable ($a \geq 0$).

Thus, with the stabilizability condition, $V_T(x_0)$ has a bound that does not depend on T by part (b) of Lemma 137. Since it is also nondecreasing in T by part (a), it must have a limit as $T \to \infty$. We state this as a corollary to Lemmas 137 and 138:

Corollary 140. *Consider the finite horizon LQR problem (28) with $S = 0$, $Q \succeq 0$, $R \succ 0$, and suppose (A, B) is stabilizable. Then, $V_T(x_0)$ is bounded (by Lemma 137(b) and Lemma 138) and nondecreasing in T (by Lemma 137(a)). Thus, $lim_{T\to\infty} V_T(x_0)$ exists.*

Since $\lim_{T\to\infty} V_T(x_0) = \lim_{T\to\infty} x_0^\top \Pi(T)x_0$ exists by Collorary 140,

$$\lim_{T\to\infty} \Pi(T) =: \overline{\Pi} \tag{36}$$

exists. This means that the solution $\Pi(t)$ of the RDE (35) converges to $\overline{\Pi}$ and, thus, $\dot{\Pi}(t)$ converges to zero. Therefore, $\overline{\Pi}$ must make the right-hand side of (35) zero; that is, it must solve the equation

$$0 = A^\top \Pi + \Pi A + Q - \Pi B R^{-1} B^\top \Pi, \tag{37}$$

which is called the Algebraic Riccati Equation (ARE).

To summarize, if (A, B) is stabilizable, then $\lim_{T\to\infty} V_T(x_0) = x_0^\top \overline{\Pi} x_0$, where $\overline{\Pi}$ is the limit of the solution of the RDE (35) and satisfies the ARE (37). Moreover, from Lemma 137, we have

$$x_0^\top \overline{\Pi} x_0 \leq J_\infty(x_0, u), \ \forall u. \tag{38}$$

The ARE (37) may admit multiple solutions, raising the question: which one is the limit $\overline{\Pi}$ of the solution of the RDE (35)? First observe that, the cost function $V_T(x_0) = x_0^\top \Pi(T)x_0 \geq 0$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

for any x_0; thus $\Pi(T)$ as well as its limit $\overline{\Pi}$ must be at least positive semidefinite. In addition, if Q is strictly positive definite, then so is $\overline{\Pi}$. This follows from (35) because $\Pi(0) = 0$ and $\dot{\Pi}(0) = Q \succ 0$, which means that at $T = 0$, $\frac{dV_T(x_0)}{dT} = x_0^\top Q x_0 > 0$ if $x_0 \neq 0$. Thus, $V_T(x_0) > 0$ for arbitrarily small $T > 0$ when $x_0 \neq 0$. Since $V_T(x_0)$ is nondecreasing in T, we conclude $V_T(x_0) = x_0^\top \Pi(T) x_0 > 0$ for all $T > 0$ when $x_0 \neq 0$; that is $\Pi(T) \succ 0$ for all T.

The following lemma summarizes the observations above:

Lemma 141. *If $Q \succeq 0$ and $R \succ 0$, then $\overline{\Pi} \succeq 0$. Moreover, if $Q \succ 0$, then $\overline{\Pi} \succ 0$.*

Example 142. Suppose $A = 1, B = 1, Q = 1, R = 1$; that is,

$$\begin{aligned} \dot{x} &= x + u \\ J_\infty(x_0, u) &= \int_0^\infty (x(t)^2 + u(t)^2)\, dt. \end{aligned}$$

Then, the ARE (37) becomes $2\pi + 1 - \pi^2 = 0$, where we used lower case since π is a scalar in this example. This quadratic equation has solutions $\pi_1 = 1 + \sqrt{2}$ and $\pi_2 = 1 - \sqrt{2}$. By Lemma 141, the relevant solution is the positive one, $\pi_1 = 1 + \sqrt{2}$.

Now that we understand the limit of the finite horizon LQR problem as $T \to \infty$, we are ready to state the **main result for the infinite horizon LQR:**

Theorem 143. *Let $Q \succeq 0, R \succ 0$, and suppose (A, B) is stabilizable. Denote by $\overline{\Pi}$ the limit of the solution of RDE (35). Then:*

(a) The input u^* generated by the feedback law

$$u^*(t) = -R^{-1} B^\top \overline{\Pi} x(t) \tag{39}$$

is the optimal solution to (34).

(b) The feedback law (39) also guarantees $\lim_{t\to\infty} x(t)^\top \overline{\Pi} x(t) = 0$.

(c) If $Q \succ 0$, then $x(t) \to 0$.

Proof. To prove (a) and (b), note that

$$\begin{aligned} J_\infty(x_0, u^*) &= \int_0^\infty \left(x(t)^\top Q x(t) + \left(-R^{-1} B^\top \overline{\Pi} x(t)\right)^\top R \left(-R^{-1} B^\top \overline{\Pi} x(t)\right) \right) dt \\ &= \int_0^\infty x(t)^\top (Q + \overline{\Pi}^\top B R^{-1} B^\top \overline{\Pi}) x(t)\, dt \end{aligned}$$

Note that the closed-loop system is $\dot{x} = Ax + B(-R^{-1}B^\top \overline{\Pi} x) = (A - BR^{-1}B^\top \overline{\Pi})x =: A_{cl} x$. Then,

$$A_{cl}^\top \overline{\Pi} + \overline{\Pi} A_{cl} = A^\top \overline{\Pi} + \overline{\Pi} A - 2\overline{\Pi} B R^{-1} B^\top \overline{\Pi} = -(Q + \overline{\Pi}^\top B R^{-1} B^\top \overline{\Pi}),$$

where the last equality follows because $\overline{\Pi}$ satisfies the ARE (37). Substituting this in our integral for $J_\infty(x_0, u^*)$ above, we get

$$J_\infty(x_0, u^*) = -\int_0^\infty x(t)^\top (A_{cl}^\top \overline{\Pi} + \overline{\Pi} A_{cl}) x(t)\, dt$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

$$
\begin{aligned}
&= -\int_0^\infty \left(\dot{x}(t)^\top \overline{\Pi} x(t) + x(t)^\top \overline{\Pi} \dot{x}(t)\right) dt \\
&= -\int_0^\infty \frac{d}{dt} x(t)^\top \Pi x(t)\ dt \\
&= -\ x(t)^\top \overline{\Pi} x(t)\Big|_0^\infty \\
&= x_0^\top \overline{\Pi} x_0 - \underbrace{\lim_{t\to\infty} x(t)^\top \overline{\Pi} x(t)}_{\geq 0} \qquad (40) \\
&\leq x_0^\top \overline{\Pi} x_0
\end{aligned}
$$

Combining with (38), we get

$$J_\infty(x_0, u^*) \leq x_0^\top \overline{\Pi} x_0 \leq J_\infty(x_0, u),\ \forall u.$$

Since the second inequality holds for all u, it holds when $u = u^*$; thus,

$$J_\infty(x_0, u^*) \leq x_0^\top \overline{\Pi} x_0 \leq J_\infty(x_0, u^*) \quad \Rightarrow \quad J_\infty(x_0, u^*) = x_0^\top \overline{\Pi} x_0.$$

In addition, (38) states that no choice of u can lower $J_\infty(x_0, u)$ below $x_0^\top \overline{\Pi} x_0$. Thus u^* achieves the minimum possible cost, which proves part (a).

Part (b) follows from (40) because we now know $J_\infty(x_0, u^*) = x_0^\top \overline{\Pi} x_0$ and, thus, the underbraced limit in (40) must be zero.

For part (c), recall from Lemma 141 that $Q \succ 0$ implies $\overline{\Pi} \succ 0$. Since $\lim_{t\to\infty} x(t)^\top \overline{\Pi} x(t) = 0$ by part (b) and since $\overline{\Pi}$ is positive definite, we conclude $x(t) \to 0$. □

Note that the optimal control (39) is a state feedback law and, unlike the finite horizon case, it is time invariant. The closed-loop system with this feedback is $\dot{x} = (A - BR^{-1}B^\top \Pi)x =: A_{cl}x$. If $Q \succ 0$, then $x(t) \to 0$, which implies that A_{cl} is necessarily Hurwitz.

The **Matlab command** $[K, P] = \text{lqr}(A, B, Q, R)$ can be used to compute the solution of LQR. It returns the optimal feedback gain K to be substituted in $u = -Kx$. That is, $K = R^{-1}B^\top P$, where P corresponds to $\overline{\Pi}$.

Example 144. Let $A = 3, B = 4, Q = 1, R = 1$. Thus,

$$
\begin{aligned}
\dot{x} &= 3x + 4u \\
J_\infty(x_0, u) &= \int_0^\infty (x^2 + u^2)\ dt.
\end{aligned}
$$

Then, the ARE is $6\pi + 1 - 16\pi^2 = 0$, whose solutions are $\Pi_1 = \frac{1}{2}$ and $\Pi_2 = -\frac{1}{8}$. The relevant one is the positive solution $\pi_1 = \frac{1}{2}$ and the optimal feedback is $u^* = -4\pi_1 x = -2x$. The corresponding closed-loop system is $\dot{x} = 3x + 4(-2x) = -5x$.

Example 145. Now suppose $A = 0, B = 1, Q = 1, R = 1$. That is,

$$
\begin{aligned}
\dot{x} &= u \\
J_\infty(x_0, u) &= \int_0^\infty (x^2 + u^2)\ dt.
\end{aligned}
$$

The ARE is $1 - \pi^2 = 0$, whose solutions are $\pi = \pm 1$. Selecting the positive one, we get the optimal control $u^* = -x$.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Example 146. Consider the following system,

$$\begin{aligned}\dot{x}_1 &= x_2\\ \dot{x}_2 &= u\end{aligned} \quad \text{with the cost function: } J_\infty = \int_0^\infty (x_1^2 + x_2^2 + u^2)\, dt.$$

That is, $Q = I, R = 1, A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$.

The corresponding ARE is $\Pi A + A^T\Pi + I - \Pi B B^T \Pi = 0$. Since Π is symmetric, we write it as $\Pi = \begin{bmatrix} \pi_{11} & \pi_{12} \\ \pi_{12} & \pi_{22} \end{bmatrix}$ with only three independent entries. Then, we have

$$\begin{aligned}&\begin{bmatrix} 0 & \pi_{11} \\ 0 & \pi_{12} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \pi_{11} & \pi_{12} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} \pi_{12} \\ \pi_{22} \end{bmatrix} \begin{bmatrix} \pi_{12} & \pi_{22} \end{bmatrix} = 0\\ \implies\ & 1 - \pi_{12}^2 = 0\\ & \pi_{11} - \pi_{12}\pi_{22} = 0\\ & 2\pi_{12} + 1 - \pi_{22}^2 = 0\\ \implies\ & \pi_{12} = \pm 1, \pi_{11} = \pi_{12}\pi_{22}, \pi_{22} = \pm\sqrt{1 + 2\pi_{12}}\end{aligned}$$

Since $\overline{\Pi} \succ 0$ (from $Q \succ 0$), we discard $\pi_{12} = -1$ because that would have implied $\pi_{11} = -\pi_{22}$, which means zero or negative diagonal entries, contradicting positive definiteness. It follows that $\pi_{12} = 1, \pi_{11} = \pi_{22} = \sqrt{3}$:

$$\overline{\Pi} = \begin{bmatrix} \sqrt{3} & 1 \\ 1 & \sqrt{3} \end{bmatrix}.$$

The optimal controller is $u^* = -R^{-1}B^\top \overline{\Pi} x = -\begin{bmatrix} 1 & \sqrt{3} \end{bmatrix} x$.

The closed-loop system is $\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & -\sqrt{3} \end{bmatrix} x$ with characteristic polynomial $s^2 + \sqrt{3}s + 1 = 0$ and eigenvalues at $\frac{-\sqrt{3} \pm j}{2}$, which confirms asymptotic stability.

2 When is the optimal control asymptotically stabilizing?

As mentioned in Theorem 143, Part (c), if $Q \succ 0$, then the optimal solution will asymptotically stabilize the system. This, however, is only a sufficient condition. To see this, suppose $Q = 0$, which means $J_\infty(x_0, u) = \int_0^\infty u^\top R u\, dt$ and the optimal control is $u^* = 0$ since there is no penalty on the state. If the matrix A is Hurwitz, we have asymptotic stability even with $u^* = 0$.

This suggests that a sharper condition for the optimal control to be asymptotically stabilizing can allow $Q \succeq 0$ as long A has suitable properties. The following example will help us identify the precise property needed.

Example 147. We return to Example 146, but this time choose

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix},$$

which is only semidefinite. Thus, the new cost is:

$$J_\infty = \int_0^\infty (x_1^2 + u^2)\, dt, \tag{41}$$

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

but the system is unchanged:

$$\dot{x}_1 = x_2$$
$$\dot{x}_2 = u.$$

Below is the modified ARE resulting from the new Q, with changes marked in red:

$$\begin{aligned}
&\begin{bmatrix} 0 & \pi_{11} \\ 0 & \pi_{12} \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \pi_{11} & \pi_{12} \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \pi_{12} \\ \pi_{22} \end{bmatrix} \begin{bmatrix} \pi_{12} & \pi_{22} \end{bmatrix} = 0 \\
\implies\ & 1 - \pi_{12}^2 = 0 \\
& \pi_{11} - \pi_{12}\pi_{22} = 0 \\
& 2\pi_{12} + 0 - \pi_{22}^2 = 0 \\
\implies\ & \pi_{12} = \pm 1, \pi_{11} = \pi_{12}\pi_{22}, \pi_{22} = \pm\sqrt{0 + 2\pi_{12}}
\end{aligned}$$

We again select[5] $\pi_{12} = 1$, which leads to $\pi_{11} = \pi_{22} = \sqrt{2}$:

$$\overline{\Pi} = \begin{bmatrix} \sqrt{2} & 1 \\ 1 & \sqrt{2} \end{bmatrix},$$

and the optimal controller is $u^* = -R^{-1}B^\top \overline{\Pi} x = -\begin{bmatrix} 1 & \sqrt{2} \end{bmatrix} x$. The closed-loop system is $\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & -\sqrt{2} \end{bmatrix} x$, which again has eigenvalues with negative real parts. Thus, the optimal control is asymptotically stabilizing even though Q is only semidefinite.

On the other hand, if we select

$$Q = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix};$$

that is,

$$J_\infty = \int_0^\infty (x_2^2 + u^2)\, dt, \tag{42}$$

the optimal control is $u^* = -x_2$, which is not asymptotically stabilizing. You can show this by modifying and solving the ARE above, and verify the closed-loop system has an eigenvalue at zero, ruling out asymptotic stability. Alternatively, note that the problem (42) with the equation $\dot{x}_2 = u$ is oblivious to the x_1 subsystem, and it is identical to Example 145 with the single state replaced with x_2. The resulting optimal control $u^* = -x_2$ drives $x_2(t)$ to zero, but the remaining subsystem $\dot{x}_1 = x_2$ integrates $x_2(t)$; thus, $x_1(t)$ does not converge to zero.

If we interpret x_2 appearing in the cost (42) as the system output (that is, $C = [0\ 1]$), then the x_1 subsystem is unobservable; in fact, *undetectable* since the unobservable x_1 subsystem is not asymptotically stable by itself. By contrast, the cost (41) depends on x_1 ($C = [1\ 0]$), from which the system is observable, thus also detectable.

[5] The argument for discarding $\pi_{12} = -1$ is slightly modified, as we no longer have $Q \succ 0$ and can only restrict our search to $\overline{\Pi} \succeq 0$, rather than $\overline{\Pi} \succ 0$. The choice $\pi_{12} = -1$ contradicts $\overline{\Pi} \succeq 0$ as well because it implies $\pi_{11} = -\pi_{22}$, meaning either a negative diagonal entry (contradicting positive semidefiniteness) or two zero diagonal entries (which also contradicts positive semidefiniteness since the off-diagonal entry π_{12} is nonzero).

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

Generalizing the example above, it can be shown that the optimal control (39) for the stabilizable system $\dot{x} = Ax + Bu$ with cost

$$J_\infty = \int_0^\infty (y^\top y + u^\top R u)\ dt, \tag{43}$$

where $y = Cx$ and the pair (C, A) is detectable, is asymptotically stabilizing. Furthermore, stabilizability and detectability guarantee that the ARE (37) admits a unique positive semidefinite solution, which is necessarily $\overline{\Pi}$, the limit of the RDE (35) used in the optimal control (39).

Note that (44) corresponds to $Q = C^\top C$ in the LQR formulation, and any $Q \succeq 0$ can be written in this form with a suitable C (using a Schur or Cholesky decomposition). The statement above (whose proof is omitted) is less restrictive than the condition $Q \succ 0$ used in Theorem 143, Part (c) to prove asymptotic stability. On the other hand, Q is a design choice, so the practical distinction between $Q \succ 0$ and $Q \succeq 0$ is insignificant. For example, instead of the cost (42), we can choose

$$J_\infty = \int_0^\infty (\varepsilon x_1^2 + x_2^2 + u^2)\ dt \tag{44}$$

with a small $\varepsilon > 0$, so that $Q = \begin{bmatrix} \varepsilon & 0 \\ 0 & 1 \end{bmatrix} \succ 0$ while the emphasis remains on regulating x_2.

3 Infinite-horizon LQR in Discrete Time

The infinite-horizon LQR problem in discrete time,

$$\begin{aligned} &\min_u\ J_\infty(x_0; u) = \sum_{k=0}^{\infty} \left(x_k^\top Q x_k + u_k^\top R u_k \right) \\ &\text{s.t. } x_{k+1} = Ax_k + Bu_k, \end{aligned}$$

is again solved using the limit as $N \to \infty$ of the finite horizon problem (19) with zero terminal cost, i.e., $S = 0$. We define $\Pi_k = P_{N-k}$ and rewrite the Riccati Difference Equation (26) as

$$\begin{aligned} \Pi_0 &= S = 0 \\ \Pi_{k+1} &= Q + A^\top \Pi_k A - A^\top \Pi_k B (R + B^\top \Pi_k B)^{-1} B^\top \Pi_k A, \quad k = 1, \cdots, N, \end{aligned} \tag{45}$$

so that it is solved forward in time instead of backwards. If (A, B) is stabilizable, $Q \succeq 0$, and $R \succ 0$, then $\overline{\Pi} = \lim_{N\to\infty} \Pi_N$ exists, and it is a positive semidefinite solution of the Discrete Algebraic Riccati Equation

$$\Pi = Q + A^\top \Pi A - A^\top \Pi B (R + B^\top \Pi B)^{-1} B^\top \Pi A.$$

The optimal control problem is then solved by the time-invariant state feedback

$$u_k^* = -(R + B^\top \overline{\Pi} B)^{-1} B^\top \overline{\Pi} A x_k,$$

which achieves the minimum value $J_\infty(x_0; u^*) = x_0^\top \overline{\Pi} x_0$. In addition, if[6] $Q \succ 0$, then u^* guarantees $x_k \to 0$ as $k \to \infty$. We omit derivations, as they are similar to continuous time.

[6] $Q \succ 0$ is again only sufficient for $x_k \to 0$ and can relaxed to $Q = C^\top C$ when (C, A) is detectable.

Kameshwar Poolla and Murat Arcak. Licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.