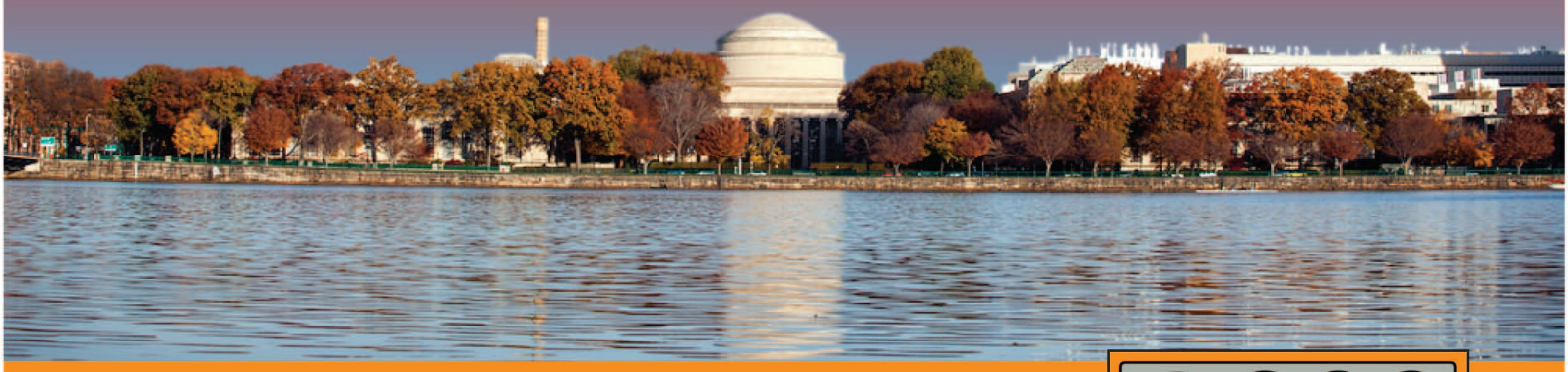
MITOPENCOURSEWARE

Lectures on Dynamic Systems and Control

by Mohammed Dahleh,
Munther A. Dahleh, &
George Verghese



Lectures on Dynamic Systems and Control

Mohammed Dahleh Munther A. Dahleh George Verghese
Department of Electrical Engineering and Computer Science
Massachuasetts Institute of Technology[1]

Chapter 1

Linear Algebra Review

1.1 Introduction

Dynamic systems are systems that evolve with time. Our models for them will comprise coupled sets of ordinary differential equations (ode's). We will study how the internal variables and outputs of such systems respond to their inputs and initial conditions, how their internal behavior can be inferred from input/output (I/O) measurements, how the inputs can be controlled to produce desired behavior, and so on. Most of our attention will be focused on *linear* models (and within this class, on *time invariant* models, i.e. on LTI models), for reasons that include the following:

- linear models describe small perturbations from nominal operation, and most control design is aimed at regulating such perturbations;
- linear models are far more tractable than general nonlinear models, so systematic and detailed control design approaches can be developed;
- engineered systems are often made up of modules that are designed to operate in essentially linear fashion, with any nonlinearities introduced in carefully selected locations and forms.

To describe the interactions of coupled variables in linear models, the tools of *linear algebra* are essential. In the first part of this course (4 or 5 lectures), we shall come up to speed with the "$Ax = y$" or linear equations part of linear algebra, by studying a variety of *least squares* problems. This will also serve to introduce ideas related to dynamic systems — e.g., recursive processing of I/O measurements from a finite-impulse-response (FIR) discrete-time (DT) LTI system, to produce estimates of its impulse response coefficients.

Later parts of the course will treat in considerable detail the representation, structure, and behavior of multi-input, multi-output (MIMO) LTI systems. The "$Av = \lambda v$"

- Show that the *intersection* of two subspaces of a vector space is itself a subspace.
- Show that the *union* of two subspaces is in general *not* a subspace. Also determine under what condition the union of subspaces will be a subspace.
- Show that the (Minkowski or) *direct sum* of subspaces, which by definition comprises vectors that can be written as the sum of vectors drawn from each of the subspaces, is a subspace.

Get in the habit of working up small (in $\mathbf{R}^2$ or $\mathbf{R}^3$, for instance) concrete examples for yourself, as you tackle problems such as the above. This will help you develop a feel for what is being stated — perhaps suggesting a strategy for a proof of a claim, or suggesting a counterexample to disprove a claim.
Review what it means for a set of vectors to be **(linearly) dependent** or **(linearly) independent**. A space is *n-dimensional* if every set of more than n vectors is dependent, but there is some set of n vectors that is independent; any such set of n independent vectors is referred to as a **basis** for the space.

- Show that any vector in an n-dimensional space can be written as a *unique* linear combination of the vectors in a basis set; we therefore say that any basis set *spans* the space.
- Show that a basis for a *sub*space can always be augmented to form a basis for the entire space.

If a space has a set of n independent vectors for every nonnegative n, then the space is called *infinite dimensional*.

- Show that the set of functions $f(t) = t^{n-1}$, $n = 1, 2, 3, \cdots$ forms a basis for an infinite dimensional space. (One route to proving this uses a key property of *Vandermonde* matrices, which you may have encountered somewhere.)

Norms

The "lengths" of vectors are measured by introducing the idea of a **norm**. A norm for a vector space $\mathcal{V}$ over the field of real numbers $\mathbf{R}$ or complex numbers $\mathbf{C}$ is defined to be a function that maps vectors x to nonnegative real numbers $\|x\|$, and that satisfies the following properties:

1. Positivity: $\|x\| > 0$ for $x \neq 0$
2. Homogeneity: $\|ax\| = |a| \, \|x\|$, scalar a.
3. Triangle inequality: $\|x + y\| \leq \|x\| + \|y\|$, $\forall x, y \in \mathcal{V}$.

- Verify that the usual Euclidean norm on $\mathbf{R}^n$ or $\mathbf{C}^n$ (namely $\sqrt{x'x}$ with $'$ denoting the complex conjugate of the transpose) satisfies these conditions.

- A complex matrix Q is termed **Hermitian** if $Q' = Q$; if Q is real, then this condition simply states that Q is symmetric. Verify that $x'Qx$ is always real, if Q is Hermitian. A matrix is termed **positive definite** if $x'Qx$ is real and positive for $x \neq 0$. Verify that $\sqrt{x'Qx}$ constitutes a norm if Q is Hermitian and positive definite.

- Verify that in $\mathbf{R}^n$ both $\|x\|_1 = \sum_1^n |x_i|$ and $\|x\|_\infty = \max_i |x_i|$ constitute norms. These are referred to as the 1-norm and ∞-norm respectively, while the examples of norms mentioned earlier are all instances of (weighted or unweighted) 2-norms. Describe the sets of vectors that have unit norm in each of these cases.

- The space of continuous fucntions on the interval [0, 1] clearly forms a vector space. One possible norm defined on this space is the ∞-norm defined as:

$$\|f\|_\infty = \sup_{t \in [0,1]} |f(t)|.$$

  This measures the peak value of the function in the interval $[0, 1]$. Another norm is the 2-norm defined as:

$$\|f\|_2 = \left(\int_0^1 |f(t)|^2 dt \right)^{\frac{1}{2}}.$$

  Verify that these measures satisfy the three properties of the norm.

Inner Product

The vector spaces that are most useful in practice are those on which one can define a notion of **inner product**. An inner product is a function of two vectors, usually denoted by $< x, y >$ where x and y are vectors, with the following properties:

1. Symmetry: $< x, y > = < y, x >'$.
2. Linearity: $< x, ay + bz > = a < x, y > + b < x, z >$ for all scalars a and b.
3. Positivity: $< x, x >$ positive for $x \neq 0$.

- Verify that $\sqrt{< x, x >}$ defines a norm.

- Verify that $x'Qy$ constitutes an inner product if Q is Hermitian and positive definite. The case of $Q = I$ corresponds to the usual Euclidean inner product.

- Verify that

$$\int_0^1 x(t)y(t)dt$$

  defines an inner product on the space of continuous functions. In this case, the norm generated from this inner product is the same as the 2-norm defined earlier.

- **Cauchy-Schwartz Inequality** Verify that for any x and y in an inner product space

$$| < x, y > | \leq \|x\|\|y\|$$

with equality if and only if $x = \alpha y$ for some scalar α. (Hint: Expand $< x+\alpha y, x+\alpha y >$).

Two vectors x, y are said to be **orthogonal** if $< x, y >= 0$; two *sets* of vectors $\mathcal{X}$ and $\mathcal{Y}$ are called orthogonal if *every* vector in one is orthogonal to *every* vector in the other. The **orthogonal complement** of a set of vectors $\mathcal{X}$ is the set of vectors orthogonal to $\mathcal{X}$, and is denoted by $\mathcal{X}^{\perp}$.

- Show that the orthogonal complement of any set is a subspace.

1.3 The Projection Theorem

Consider the following minimization problem:

$$\min_{m \in M} \|y - m\|$$

where the norm is defined through an inner product. The projection theorem (suggested by the figure below), states that the optimal solution $\hat{m}$ is characterized as follows:

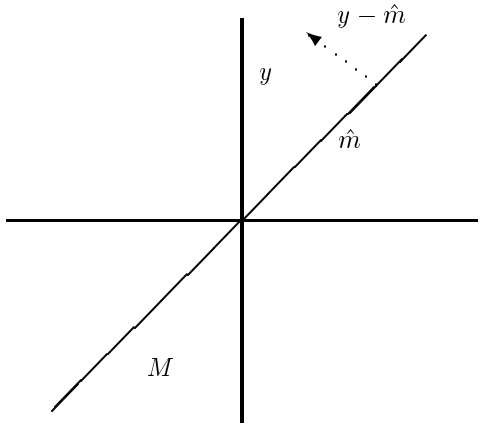
$$(y - \hat{m}) \perp M.$$

To verify this theorem, assume the converse. Then there exists an m_0, $\|m_0\| = 1$, such that $< y - \hat{m}, m_0 >= \delta \neq 0$. We now argue that $(\hat{m} + \delta m_0) \in M$ achieves a smaller value to the above minimization problem. In particular,

$$\begin{aligned} \|y - \hat{m} - \delta m_0\|^2 &= \|y - \hat{m}\|^2 - < y - \hat{m}, \delta m_0 > - < \delta m_0, y - \hat{m} > + |\delta|^2 \|m_0\|^2 \\ &= \|y - \hat{m}\|^2 - |\delta|^2 - |\delta|^2 + |\delta|^2 \\ &= \|y - \hat{m}\|^2 - |\delta|^2 \end{aligned}$$

This conradicts the optimality of $\hat{m}$.

- Given a subspace $\mathcal{S}$, show that any vector x can be *uniquely* written as $x = x_{\mathcal{S}} + x_{\mathcal{S}^{\perp}}$, where $x_{\mathcal{S}} \in \mathcal{S}$ and $x_{\mathcal{S}^{\perp}} \in \mathcal{S}^{\perp}$.



1.4 Matrices

Our usual notion of a matrix is that of a rectangular array of scalars. The definitions of matrix addition, multiplication, etc., are aimed at compactly representing and analyzing systems of equations of the form

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= y_1 \\ \cdots &\ \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= y_m \end{aligned}$$

This system of equations can be written as $Ax = y$ if we define

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ldots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}$$

The rules of matrix addition, matrix multiplication, and scalar multiplication of a matrix remain unchanged if the entries of the matrices we deal with are themselves (conformably dimensioned) *matrices* rather than scalars. A matrix with matrix entries is referred to as a **block** matrix or a **partitioned** matrix.

For example, the a_{ij}, x_j, and y_i in respectively A, x, and y above can be matrices, and the equation $Ax = y$ will still hold, as long as the dimensions of the various submatrices are conformable with the expressions $\sum a_{ij}x_j = y_i$ for $i = 1, \cdots, m$ and $j = 1, \cdots, n$. What this requires is that the number of rows in a_{ij} should equal the number of rows in y_i, the number of columns in a_{ij} should equal the number of rows in x_j, and the number of columns in the x_j and y_i should be the same.

- Verify that

$$\left(\begin{array}{cc|c} 1 & 2 & 2 \\ 0 & 1 & 3 \\ 1 & 1 & 7 \end{array}\right)\left(\begin{array}{cc} 4 & 5 \\ 8 & 9 \\ \hline & \\ 2 & 0 \end{array}\right) = \left(\begin{array}{cc} 1 & 2 \\ 0 & 1 \\ 1 & 1 \end{array}\right)\quad\left(\begin{array}{cc} 4 & 5 \\ 8 & 9 \end{array}\right) + \left(\begin{array}{c} 2 \\ 3 \\ 7 \end{array}\right)\quad\begin{array}{cc} 2 & 0 \end{array}$$

In addition to these simple rules for matrix addition, matrix multiplication, and scalar multiplication of partitioned matrices, there is a simple — and simply verified — rule for (complex conjugate) *transposition* of a partitioned matrix: if $[A]_{ij} = a_{ij}$, then $[A']_{ij} = a'_{ji}$, i.e., the (i,j)-th block element of A' is the *transpose* of the (j,i)-th block element of A.

For more involved matrix operations, one has to proceed with caution. For instance, the determinant of the square block-matrix

$$A = \left(\begin{array}{cc} A_1 & A_2 \\ A_3 & A_4 \end{array}\right)$$

is clearly *not* $A_1A_4 - A_3A_2$ unless all the blocks are actually scalar! We shall lead you to the correct expression (in the case where A_1 is square and invertible) in a future Homework.

Matrices as Linear Transformations

T is a transformation or mapping from X to Y, two vector spaces, if it associates to each $x \in X$ a unique element $y \in Y$. This transformation is linear if it satisfies

$$T(\alpha x + \beta y) = \alpha T(x) + \beta T(y).$$

- Verify that an $n \times m$ matrix A is a linear transformation from $\mathbf{R}^m$ to $\mathbf{R}^n$.

Does every linear transformation have a matrix representation? Assume that both X and Y are finite dimensional spaces with respective bases $\{x_1, \ldots x_m\}$ and $\{y_1, \ldots y_n\}$. Every $x \in X$ can be uniquely expressed as: $x = \sum_{i=1}^m a_i x_i$. Equivalently, every x is represented uniquely in terms of an element $a \in \mathbf{R}^m$. Similarly every element $y \in Y$ is uniquely represented in terms of an element $b \in \mathbf{R}^n$. Now: $T(x_j) = \sum_{i=1}^n b_{ij} y_i$ and hence

$$T(x) = \sum_{j=1}^m a_j T(x_j) = \sum_{i=1}^n y_i (\sum_{j=1}^m a_j b_{ij})$$

A matrix representation is then given by $B = (b_{ij})$. It is evident that a matrix representation is not unique and depends on the basis choice.

1.5 Linear Systems of Equations

Suppose that we have the following system of real or complex linear equations:

$$A^{m\times n}\, x^{n\times 1} = y^{m\times 1}$$

When does this system have a solution x for given A and y?

$$\exists \text{ a solution } x \iff y \in \mathcal{R}(A) \iff \mathcal{R}([A \,|\, y]) = \mathcal{R}(A)$$

We now analyze some possible cases:

(1) If $n = m$, then $\det(A) \neq 0 \Rightarrow x = A^{-1}y$, and x is the unique solution.

(2) If $m > n$, then there are more equations than unknowns, i.e. the system is "overconstrained". If A and/or y reflect actual experimental data, then it is quite likely that the n-component vector y does *not* lie in $\mathcal{R}(A)$, since this subspace is only n-dimensional (if A has full column rank) or less, but lives in an m-dimensional space. The system will then be *inconsistent*. This is the sort of situation encountered in estimation or identification problems, where x is a parameter vector of low dimension compared to the dimension of the measurements that are available. We then look for a choice of x that comes closest to achieving consistency, according to some error criterion. We shall say quite a bit more about this shortly.

(3) If $m < n$, then there are fewer equations than unknowns, and the system is "underconstrained". If the system has a particular solution x_p (and when $\text{rank}(A) = m$, there is guaranteed to be a solution for any y) then there exist an infinite number of solutions. More specifically, x is a solution iff (if and only if)

$$x = x_p + x_h \ , \qquad Ax_p = y \ , \qquad Ax_h = 0 \quad \text{i.e. } x_h \in \mathcal{N}(A)$$

Since the nullspace $\mathcal{N}(A)$ has dimension at least $n - m$, there are at least this many degrees of freedom in the solution. This is the sort of situation that occurs in many control problems, where the control objectives do not uniquely constrain or determine the control. We then typically search among the available solutions for ones that are optimal according to some criterion.

Exercises

Exercise 1.1 Partitioned Matrices

Suppose

$$A = \begin{matrix} A_1 & A_2 \\ 0 & A_4 \end{matrix}$$

with A_1 and A_4 *square*.

(a) Write the determinant $\det A$ in terms of $\det A_1$ and $\det A_4$. (Hint: Write A as the product

$$\begin{matrix} I & 0 \\ 0 & A_4 \end{matrix} \qquad \begin{matrix} A_1 & A_2 \\ 0 & I \end{matrix}$$

and use the fact that the determinant of the product of two *square* matrices is the product of the individual determinants — the individual determinants are easy to evaluate in this case.)

(b) Assume for this part that A_1 and A_4 are *nonsingular* (i.e., square and invertible). Now find A^{-1}. (Hint: Write $AB = I$ and partition B and I commensurably with the partitioning of A.)

Exercise 1.2 Partitioned Matrices

Suppose

$$A = \begin{matrix} A_1 & A_2 \\ A_3 & A_4 \end{matrix}$$

where the A_i are matrices of conformable dimension.

(a) What can A be premultiplied by to get the matrix

$$\begin{matrix} A_3 & A_4 \\ A_1 & A_2 \end{matrix} \quad ?$$

(b) Assume that A_1 is nonsingular. What can A be premultiplied by to get the matrix

$$\begin{matrix} A_1 & A_2 \\ 0 & C \end{matrix}$$

where $C = A_4 - A_3 A_1^{-1} A_2$?

(c) Suppose A is a square matrix. Use the result in (b) — and the fact mentioned in the hint to Problem 1(a) — to obtain an expression for $\det(A)$ in terms of determinants involving only the submatrices A_1, A_2, A_3, A_4.

Exercise 1.3 Matrix Identities

Prove the following *very useful* matrix identities. In proving identities such as these, see if you can obtain proofs that make as few assumptions as possible beyond those implied by the problem statement. For example, in (1) and (2) below, neither A nor B need be square, and in (3) neither B nor D need be square — so avoid assuming that any of these matrices is (square and) invertible!.

(a) $\det(I - AB) = \det(I - BA)$, if A is $p \times q$ and B is $q \times p$. (Hint: Evaluate the determinants of

$$\begin{matrix} I & A \\ B & I \end{matrix} \quad , \quad \begin{matrix} I & -A \\ 0 & I \end{matrix} \quad , \quad \begin{matrix} I & -A \\ 0 & I \end{matrix} \quad \begin{matrix} I & A \\ B & I \end{matrix}$$

to obtain the desired result). One common situation in which the above result is useful is when $p > q$; why is this so?

(b) Show that $(I - AB)^{-1}A = A(I - BA)^{-1}$.

(c) Show that $(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$. (Hint: Multiply the right side by $A + BCD$ and cleverly gather terms.) This is perhaps the most used of matrix identities, and is known by various names — the matrix inversion lemma, the $ABCD$ lemma (!), Woodbury's formula, etc. It is rediscovered from time to time in different guises. Its noteworthy feature is that, if A^{-1} is known, then the inverse of a modification of A is expressed as a modification of A^{-1} that may be simple to compute, e.g. when C is of small dimensions. Show, for instance, that evaluation of $(I - ab^T)^{-1}$, where a and b are column vectors, only requires inversion of a scalar quantity.

Exercise 1.4 Range and Rank

This is a practice problem in linear algebra (except that you have perhaps only seen such results stated for the case of real matrices and vectors, rather than *complex* ones — the extensions are routine).

Assume that $A \in \mathbf{C}^{m \times n}$ (i.e., A is a complex $m \times n$ matrix) and $B \in \mathbf{C}^{n \times p}$. We shall use the symbols $\mathcal{R}(A)$ and $\mathcal{N}(A)$ to respectively denote the range space and null space (or kernel) of the matrix A. Following the Matlab convention, we use the symbol A' to denote the transpose of the *complex conjugate of* the matrix A; $\mathcal{R}^{\perp}(A)$ denotes the subspace *orthogonal* to the subspace $\mathcal{R}(A)$, i.e. the set of vectors x such that $x'y = 0\,, \quad \forall y \in \mathcal{R}(A)$, etc.

(a) Show that $\mathcal{R}^{\perp}(A) = \mathcal{N}(A')$ and $\mathcal{N}^{\perp}(A) = \mathcal{R}(A')$.

(b) Show that

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

This result is referred to as *Sylvester's inequality.*

Exercise 1.5 Vandermonde Matrix

A matrix with the following structure is referred to as a *Vandermonde* matrix:

$$\begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^{n-1} \\ 1 & \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^{n-1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \lambda_n & \lambda_n^2 & \cdots & \lambda_n^{n-1} \end{pmatrix}$$

This matrix is clearly singular if the λ_i are *not* all distinct. Show the converse, namely that if all n of the λ_i are distinct, then the matrix is nonsingular. One way to do this — although **not** the easiest! — is to show by induction that the determinant of the Vandermonde matrix is

$$\prod_{i=1\,;\,j>i}^{i,j=n} (\lambda_j - \lambda_i)$$

Look for an easier argument first.

Exercise 1.6 Matrix Derivatives

(a) Suppose $A(t)$ and $B(t)$ are matrices whose entries are differentiable functions of t, and assume the product $A(t)B(t)$ is well-defined. Show that

$$\frac{d}{dt}\ A(t)B(t)\ = \frac{dA(t)}{dt}B(t)\ +\ A(t)\frac{dB(t)}{dt}$$

where the derivative of a matrix is, by definition, the matrix of derivatives — i.e., to obtain the derivative of a matrix, simply replace each entry of the matrix by its derivative. (Note: The ordering of the matrices in the above result is important!).

(b) Use the result of (a) to evaluate the derivative of the *inverse* of a matrix $A(t)$, i.e. evaluate the derivative of $A^{-1}(t)$.

Exercise 1.7 Suppose T is a linear transformation from X to itself. Verify that any two matrix representations, A and B, of T are related by a nonsingular transformation; i.e., $A = R^{-1}BR$ for some R. Show that as R varies over all nonsingular matrices, we get all possible representations.

Exercise 1.8 Let X be the vector space of polynomials of order less than or equal to M.

(a) Show that the set $B = \{1,\, x,\, \ldots\, x^M\}$ is a basis for this vector space.

(b) Consider the mapping T from X to X defined as:

$$f(x) = Tg(x) = \frac{d}{dx}g(x)$$

1. Show that T is linear.
2. Derive a matrix representation for T in terms of the basis B.
3. What are the eigenvalues of T.
4. Compute one eigenvector associated with one of the eighenvalues.

Chapter 2

Least Squares Estimation

2.1 Introduction

If the criterion used to measure the error $e = y - Ax$ in the case of inconsistent system of equations is the sum of squared magnitudes of the error components, i.e. $e'e$, or equivalently the square root of this, which is the usual Euclidean norm or 2-norm $\|e\|_2$, then the problem is called a *least squares* problem. Formally it can be written as

$$\min_x \|y - Ax\|_2. \tag{2.1}$$

The x that minimizes this criterion is called the least square error estimate, or more simply, the *least squares estimate*. The choice of this criterion and the solution of the problem go back to Legendre (1805) and Gauss (around the same time).

Example 2.1 Suppose we make some measurements y_i of an unknown function $f(t)$ at discrete points $t_i,\ i = 1, \ldots, N$:

$$y_i = f(t_i)\ , \qquad i = 1, \ldots, N.$$

We want to find the function $g(t)$ in the space χ of polynomials of order $m - 1 < N - 1$ that best approximates $f(t)$ at the measured points t_i, where

$$\chi = \left\{ g(t) = \sum_{i=0}^{m-1} \alpha_i t^i,\ \alpha_i \text{ real} \right\}$$

For any $g(t) \in \chi$, we will have $y_i = g(t_i) + e_i$ for $i = 1, \ldots, N$. Writing this in

matrix form for the available data, we have

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^{m-1} \\ \vdots & & & \vdots & \\ 1 & t_N & t_N^2 & \cdots & t_N^{m-1} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \alpha_0 \\ \vdots \\ \alpha_{m-1} \end{bmatrix}}_{x} + \underbrace{\begin{bmatrix} e_1 \\ \vdots \\ e_N \end{bmatrix}}_{e}$$

The problem is to find $\alpha_0, \ldots, \alpha_{m-1}$ such that $e'e = \sum_{i-1}^{N} e_i^2$ is minimized.

2.2 Computing the Estimate

The solution, $\hat{x}$, of Equation 2.1 is characterized by:

$$(y - A\hat{x}) \perp \mathcal{R}(A).$$

All elements in a basis of $\mathcal{R}(A)$ must be orthogonal to $(y - A\hat{x})$. Equivalently this is true for the set of columns of A, $[a_1, \ldots, a_n]$. Thus

$$\begin{aligned} (y - A\hat{x}) \perp \mathcal{R}(A) \quad &\Leftrightarrow a_i' (y - A\hat{x}) = 0 \qquad \text{for } i = 1, \ldots, n \\ &\Leftrightarrow A'(y - A\hat{x}) = 0 \\ &\Leftrightarrow A'A\hat{x} = A'y \end{aligned}$$

This system of m equations in the m unknowns of interest is referred to as the **normal equations**. We can solve for the unique $\hat{x}$ iff $A'A$ is invertible. Conditions for this will be derived shortly. In the sequel, we will present the generalization of the above ideas for infinite dimensional vector spaces.

2.3 Preliminary: The Gram Product

Given the array of n_A vectors $A = [a_1 \mid \cdots \mid a_{n_A}]$ and the array of n_B vectors $B = [b_1 \mid \cdots \mid b_{n_B}]$ from a given inner product space, let $\prec A, B \succ$ denote the $n_A \times n_B$ *matrix* whose (i, j)-th element is $< a_i, b_j >$. We shall refer to this object as the *Gram product* (but note that this terminology is not standard!).

If the vector space under consideration is $\mathbf{R}^m$ or $\mathbf{C}^m$, then both A and B are matrices with m rows, but our definition of $\prec A, B \succ$ can actually handle more general A, B. In fact, the vector space can be infinite dimensional, as long as we are only examining finite collections of vectors from this space. For instance, we could use the same notation to treat finite collections of vectors chosen from the infinite-dimensional vector space $\mathcal{L}^2$ of *square*

integrable functions, i.e. functions $a(t)$ for which $\int_{-\infty}^{\infty} a^2(t)\,dt < \infty$. The inner product in $\mathcal{L}^2$ is $< a(t), b(t) >= \int_{-\infty}^{\infty} a^*(t)b(t)\,dt$. (The space $\mathcal{L}^2$ is an example of an infinite dimensional *Hilbert space*, and most of what we know for finite dimensional spaces — which are also Hilbert spaces! — has natural generalizations to infinite dimensional Hilbert spaces. Many of these generalizations involve introducing notions of topology and measure, so we shall *not* venture too far there. It is worth also mentioning here another important infinite dimensional Hilbert space that is central to the *probabilistic* treatment of least squares estimation: the space of zero-mean *random variables*, with the expected value $E(ab)$ serving as the inner product $< a, b >$.)

For the usual Euclidean inner product in an m-dimensional space, where $< a_i, b_j >= a_i' b_j$, we simply have $\prec A, B \succ = A'B$. For the inner product defined by $< a_i, b_j >= a_i' S b_j$ for a positive definite, Hermitian matrix S, we have $\prec A, B \succ = A'SB$.

- Verify that the symmetry and linearity of the inner product imply the same for the Gram product, so $\prec AF, BG + CH \succ = F' \prec A, B \succ G + F' \prec A, C \succ H$, for any constant matrices F, G, H (a constant matrix is a matrix of scalars), with A, B, C denoting arrays whose columns are vectors.

2.4 The Least Squares Estimation Problem

The problem of interest is to find the least square error (LSE) estimate of the parameter vector x that arises in the linear model $y \approx Ax$, where A is an array of n vectors, $A = [a_1\,, \cdots, a_n]$. Defining the *error* e by

$$e = y - Ax$$

what we want to determine is

$$\widehat{x} = \arg\min_x \|e\| = \arg\min_x \|y - Ax\|\,, \qquad y,\ A \ \text{ given}$$

(where "arg $\min_x$" should be read as "the value of the argument x that minimizes"). To state this yet another way, note that as x is varied, Ax ranges over the subspace $\mathcal{R}(A)$, so we are looking for the point

$$\widehat{y} = A\widehat{x}$$

in $\mathcal{R}(A)$ that comes closest to y, as measured by whatever norm we are using.

Rather than restricting the norm in the above expression to be the Euclidean 2-norm used in Lecture 1, we shall now actually permit it to be any norm induced by an inner product, so $\|e\| = \sqrt{< e, e >}$. This will allow us to solve the so-called *weighted* least squares problem in a finite dimensional space with no additional work, because error criteria of the form $e'Se$ for positive definite Hermitian S are thereby included. Also, our problem formulation then applies to infinite dimensional spaces that have an inner product defined on them, with the restriction that our model Ax be confined to a finite dimensional subspace. This actually covers the cases of most interest to us; treatment of the more general case involves introducing further topological notions (*closed* subspaces, etc.), and we avoid doing this.

We shall also assume that the vectors a_i, $i = 1, \ldots, n$ in A are *independent*. This assumption is satisfied by any reasonably parametrized model, for otherwise there would be an infinite number of choices of x that attained any achievable value of the error $y - Ax$. If the vectors in A are discovered to be dependent, then a re-parametrization of the model is needed to yield a well-parametrized model with independent vectors in the new A. (A subtler problem — and one that we shall say something more about in the context of ill-conditioning and the singular value decomposition — is that the vectors in A can be *nearly* dependent, causing practical difficulties in numerical estimation of the parameters.)

Gram Matrix Lemma

An important route to verifying the independence of the vectors that make up the columns of A is a lemma that we shall refer to as the *Gram Matrix Lemma*. This states that the vectors in A are independent iff the associated Gram matrix (or *Gramian*) $\prec A, A \succ = [< a_i, a_j >]$ is invertible; all norms are equivalent, as far as this result is concerned — one can pick any norm. As noted above, for the case of the usual Euclidean inner product, $\prec A, A \succ = A'A$. For an inner product of the form $< a_i, a_j > = a_i' S a_j$, where S is Hermitian and positive definite, we have $\prec A, A \succ = A'SA$. The lemma applies to the infinite dimensional setting as well (e.g. $\mathcal{L}^2$), provided we are only considering the independence of a finite subset of vectors.

Proof: If the vectors in A are dependent, there is some nonzero vector η such that $A\eta = \sum_j a_j \eta_j = 0$. But then $\sum_j < a_i, a_j > \eta_j = 0$, by the linearity of the inner product; in matrix form, we can write $\prec A, A \succ \eta = 0$ — so $\prec A, A \succ$ is not invertible.

Conversely, if $\prec A, A \succ$ is not invertible, then $\prec A, A \succ \eta = 0$ for some nonzero η. But then $\eta' \prec A, A \succ \eta = 0$, so by the linearity of inner products $< \sum \eta_i a_i , \sum a_j \eta_j > = 0$, i.e. the norm of the vector $\sum a_j \eta_j = A\eta$ is zero, so the vectors in A are dependent.

2.5 The Projection Theorem and the Least Squares Estimate

The solution to our least squares problem is now given by the *Projection Theorem*, also referred to as the **Orthogonality Principle**, which states that

$$\widehat{e} = (y - A\widehat{x}) \quad \perp \quad \mathcal{R}(A)$$

from which - — as we shall see — $\widehat{x}$ can be determined. In words, the theorem/"principle" states that the point $\widehat{y} = A\widehat{x}$ in the subspace $\mathcal{R}(A)$ that comes closest to y is characterized by the fact that the associated error $\widehat{e} = y - \widehat{y}$ is orthogonal to $\mathcal{R}(A)$, i.e., orthogonal to the space spanned by the vectors in A. This principle was presented and proved in the previous chapter. We repeat the proof here in the context of the above problem.

Proof: We first show that y has a unique decomposition of the form $y = y_1 + y_2$, where $y_1 \in \mathcal{R}(A)$ and $y_2 \in \mathcal{R}^{\perp}(A)$. We can write any $y_1 \in \mathcal{R}(A)$ in the form $y_1 = A\alpha$ for some vector α.

If we want $(y - y_1) \in \mathcal{R}^{\perp}(A)$, we must see if there is an α that satisfies

$$< a_i, (y - A\alpha) > = 0 \ , \quad i = 1, \ldots, n$$

or, using our Gram product notation,

$$\prec A, (y - A\alpha) \succ = 0$$

Rearranging this equation and using the linearity of the Gram product, we get

$$\prec A, A \succ \alpha = \prec A, y \succ$$

which is in the form of the normal equations that we encountered in Lecture 1. Under our assumption that the vectors making up the columns of A are independent, the Gram matrix lemma shows that $\prec A, A \succ$ is invertible, so the unique solution of the preceding equation is

$$\alpha = \prec A, A \succ^{-1} \prec A, y \succ$$

We now have the decomposition that we sought.

To show that the preceding decomposition is unique, let $y = y_{1a} + y_{2a}$ be another such decomposition, with $y_{1a} \in \mathcal{R}(A)$ and $y_{2a} \in \mathcal{R}^{\perp}(A)$. Then

$$y_1 - y_{1a} = y_2 - y_{2a}$$

and the left side is in $\mathcal{R}(A)$ while the right side is in its orthogonal complement. It is easy to show that the only vector common to a subspace and its orthogonal complement is the zero vector, so $y_1 - y_{1a} = 0$ and $y_2 - y_{2a} = 0$, i.e., the decomposition of y is unique.

To proceed, decompose the error $e = y - Ax$ similarly (and uniquely) into the sum of $e_1 \in \mathcal{R}(A)$ and $e_2 \in \mathcal{R}^{\perp}(A)$. Note that

$$\|e\|^2 = \|e_1\|^2 + \|e_2\|^2$$

Now we can rewrite $e = y - Ax$ as

$$e_1 + e_2 = y_1 + y_2 - Ax$$

or

$$e_2 - y_2 = y_1 - e_1 - Ax$$

Since the right side of the above equation lies in $\mathcal{R}(A)$ and the left side lies in $\mathcal{R}^{\perp}(A)$, each side separately must equal 0 — again because this is the only vector common to a subspace and its orthogonal complement. We thus have $e_2 = y_2$, and the choice of x can do nothing to affect e_2. On the other hand, $e_1 = y_1 - Ax = A(\alpha - x)$, and the best we can do as far as minimizing $\|e\|^2$ is to make $e_1 = 0$ by choosing $x = \alpha$, so $\widehat{x} = \alpha$, i.e.,

$$\boxed{\widehat{x} = \prec A, A \succ^{-1} \prec A, y \succ}$$

This solves the least squares estimation problem that we have posed.

The above result, though rather abstractly developed, is immediately applicable to many concrete cases of interest.

- Specializing to the case of $\mathbf{R}^m$ or $\mathbf{C}^m$, and choosing x to minimize the usual Euclidean norm,

  $$\|e\|^2 = e'e = \sum_{i=1}^{m} |e_i|^2$$

  we have

  $$\widehat{x} = (A'A)^{-1}A'y$$

  Note that if the columns of A form a mutually orthogonal set (i.e. an orthogonal basis for $\mathcal{R}(A)$), then $A'A$ is diagonal, and its inversion is trivial.

- If instead we choose to minimize $e'Se$ for some positive definite Hermitian S ($\neq I$), we have a **weighted least squares** problem, with solution given by

  $$\widehat{x} = (A'SA)^{-1}A'Sy$$

  For instance, with a *diagonal* S, the criterion that we are trying to minimize becomes

  $$\sum_{i=1}^{m} s_{ii}|e_i|^2$$

  where the s_{ii} are all positive. We can thereby preferentially weight those equations in our linear system for which we want a smaller error in the final solution; a larger value of s_{ii} will encourage a smaller e_i.

  Such weighting is important in any practical situation, where different measurements y_i may have been subjected to different levels of noise or uncertainty. One might expect that s_{ii} should be inversely proportional to the noise intensity on the ith equation. In fact, a probabilistic derivation, assuming zero-mean noise on each equation in the system but noise that is uncorrelated across equations, shows that s_{ii} should vary inversely with the *variance* of e_i.

  A full matrix S rather than a diagonal one would make sense if the errors were correlated across measurements. A probabilistic treatment shows that the proper weighting matrix is $S = (E[ee'])^{-1}$, the inverse of the *covariance matrix* of e. In the deterministic setting, one has far less guidance on picking a good S.

- The boxed result also allows us to immediately write down the choice of coefficients x_i that minimizes the integral

$$\int [\, y(t) - a_1(t)x_1 - a_2(t)x_2 - \cdots - a_n(t)x_n \,]^2 \, dt$$

for specified functions $y(t)$ and $a_i(t)$. If, for instance, $y(t)$ is of finite extent (or finite "support") T, and the $a_i(t)$ are sinusoids whose frequencies are integral multiples of $2\pi/T$, then the formulas that we obtain for the x_i are just the familiar Fourier series expressions. A simplification in this example is that the vectors in A are orthogonal, so $\prec A, A \succ$ is diagonal.

2.6 Recursive Least Squares (optional)

What if the data is coming in sequentially? Do we have to recompute everything each time a new data point comes in, or can we write our new, updated estimate in terms of our old estimate?

Consider the model

$$y_i = A_i x + e_i \quad , \quad i = 0, 1, \ldots , \tag{2.2}$$

where $y_i \in \mathbf{C}^{m\times 1}$, $A_i \in \mathbf{C}^{m\times n}$, $x \in \mathbf{C}^{n\times 1}$, and $e_i \in \mathbf{C}^{m\times 1}$. The vector e_k represents the mismatch between the measurement y_k and the model for it, $A_k x$, where A_k is known and x is the vector of parameters to be estimated. At each time k, we wish to find

$$\widehat{x}_k = \arg\min_x \left(\sum_{i=1}^{k} (y_i - A_i x)_i' S_i (y_i - A_i x) \right) = \arg\min_x \left(\sum_{i=1}^{k} e_i' S_i e_i \right) \quad , \tag{2.3}$$

where $S_i \in \mathbf{C}^{m\times m}$ is a positive definite Hermitian matrix of weights, so that we can vary the importance of the e_i's and components of the e_i's in determining $\widehat{x}_k$.

To compute $\widehat{x}_{k+1}$, let:

$$\overline{y}_{k+1} = \begin{bmatrix} y_0 \\ y_1 \\ . \\ . \\ y_{k+1} \end{bmatrix} ; \qquad \overline{A}_{k+1} = \begin{bmatrix} A_0 \\ A_1 \\ . \\ . \\ A_{k+1} \end{bmatrix} ; \qquad \overline{e}_{k+1} = \begin{bmatrix} e_0 \\ e_1 \\ . \\ . \\ e_{k+1} \end{bmatrix} ;$$

and

$$\overline{S}_{k+1} = \text{diag } (S_0 \, , \, S_1 \, , \, \ldots \, , \, S_{k+1})$$

where S_i is the weighting matrix for e_i.

Our problem is then equivalent to

$$\min(\overline{e}'_{k+1}\overline{S}_{k+1}\overline{e}_{k+1})$$

$$\text{subject to:} \quad \overline{y}_{k+1} = \overline{A}_{k+1}x_{k+1} + \overline{e}_{k+1}$$

The solution can thus be written as

$$(\overline{A}'_{k+1}\overline{S}_{k+1}\overline{A}_{k+1})\widehat{x}_{k+1} = \overline{A}'_{k+1}\overline{S}_{k+1}\overline{y}_{k+1}$$

or in summation form as

$$\left(\sum_{i=0}^{k+1} A_i' S_i A_i\right) \widehat{x}_{k+1} = \sum_{i=0}^{k+1} A_i' S_i y_i$$

Defining

$$Q_{k+1} = \sum_{i=0}^{k+1} A_i' S_i A_i.$$

we can write a recursion for Q_{k+1} as follows:

$$Q_{k+1} = Q_k + A'_{k+1}S_{k+1}A_{k+1}.$$

Rearranging the summation form equation for $\widehat{x}_{k+1}$, we get

$$\begin{aligned}\widehat{x}_{k+1} &= Q_{k+1}^{-1}\left[\left(\textstyle\sum_{i=0}^{k} A_i' S_i A_i\right)\widehat{x}_k + A'_{k+1}S_{k+1}y_{k+1}\right] \\ &= Q_{k+1}^{-1}\left[Q_k\widehat{x}_k + A'_{k+1}S_{k+1}y_{k+1}\right]\end{aligned}$$

This clearly displays the new estimate as a weighted combination of the old estimate and the new data, so we have the desired recursion. Another useful form of this result is obtained by substituting from the recursion for Q_{k+1} above to get

$$\widehat{x}_{k+1} = \widehat{x}_k - Q_{k+1}^{-1}\left(A'_{k+1}S_{k+1}A_{k+1}\widehat{x}_k - A'_{k+1}S_{k+1}y_{k+1}\right) \quad ,$$

which finally reduces to

$$\widehat{x}_{k+1} = \widehat{x}_k + \underbrace{Q_{k+1}^{-1}A'_{k+1}S_{k+1}}_{\text{Kalman Filter Gain}}\underbrace{(y_{k+1} - A_{k+1}\widehat{x}_k)}_{\text{innovations}}$$

The quantity $Q_{k+1}^{-1}A'_{k+1}S_{k+1}$ is called the *Kalman gain*, and $y_{k+1} - A_{k+1}\widehat{x}_k$ is called the *innovations*, since it compares the difference between a data update and the prediction given the last estimate.

Unfortunately, as one acquires more and more data, i.e. as k grows large, the Kalman gain goes to zero. One data point cannot make much headway against the mass of previous data which has 'hardened' the estimate. If we leave this estimator as is—without modification—the estimator 'goes to sleep' after a while, and thus doesn't adapt well to parameter changes. The homework investigates the concept of a 'fading memory' so that the estimator doesn't go to sleep.

An Implementation Issue

Another concept which is important in the implementation of the RLS algorithm is the computation of Q_{k+1}^{-1}. If the dimension of Q_k is very large, computation of its inverse can be computationally expensive, so one would like to have a recursion for Q_{k+1}^{-1}.

This recursion is easy to obtain. Applying the handy matrix identity

$$(A + BCD)^{-1} = A^{-1} - A^{-1}B \ \ DA^{-1}B + C^{-1} \ \ ^{-1}DA^{-1}$$

to the recursion for Q_{k+1} yields

$$Q_{k+1}^{-1} = Q_k^{-1} - Q_k^{-1}A_{k+1}' \ \ A_{k+1}Q_k^{-1}A_{k+1}' + S_{k+1}^{-1} \ \ ^{-1}A_{k+1}Q_k^{-1} \ \ .$$

Upon defining

$$P_{k+1} = Q_{k+1}^{-1} \ \ ,$$

this becomes

$$P_{k+1} = P_k - P_k A_{k+1}' \ \ S_{k+1}^{-1} + A_{k+1}P_k A_{k+1}' \ \ ^{-1}A_{k+1}P_k \ \ .$$

which is called the (discrete-time) *Riccati equation.*

Interpretation

We have $\widehat{x}_k$ and y_{k+1} available for computing our updated estimate. Interpreting $\widehat{x}_k$ as a measurement, we see our model becomes

$$\left[\begin{array}{c} \widehat{x}_k \\ y_{k+1} \end{array} \right] = \left[\begin{array}{c} I \\ A_{k+1} \end{array} \right] x + \left[\begin{array}{c} e_k \\ e_{k+1} \end{array} \right].$$

The criterion, then, by which we choose $\widehat{x}_{k+1}$ is thus

$$\widehat{x}_{k+1} = \text{argmin} \, (e_k' Q_k e_k + e_{k+1}' S_{k+1} e_{k+1} \quad .$$

In this context, one interprets Q_k as the weighting factor for the previous estimate.

Exercises

Exercise 2.1 Least Squares Fit of an Ellipse

Suppose a particular object is modeled as moving in an elliptical orbit centered at the origin. Its nominal trajectory is described in rectangular coordinates (r, s) by the constraint equation $x_1 r^2 + x_2 s^2 + x_3 rs = 1$, where x_1, x_2, and x_3 are unknown parameters that specify the orbit. We have available the following noisy measurements of the object's coordinates (r, s) at ten different points on its orbit:

(0.6728, 0.0589) (0.3380, 0.4093) (0.2510, 0.3559) (-0.0684, 0.5449)
(-0.4329, 0.3657) (-0.6921, 0.0252) (-0.3681, -0.2020) (0.0019, -0.3769)
(0.0825, -0.3508) (0.5294, -0.2918)

The ten measurements are believed to be equally reliable. For your convenience, these ten pairs of measured (r, s) values have been stored in column vectors named r and s that you can access through the 6.241 locker on Athena.* After `add 6.241`, and once in the directory in which you are running Matlab, you can copy the data using `cp /mit/6.241/Public/fall95/hw1rs.mat hw1rs.mat`. Then, in Matlab, type `load hw1rs` to load the desired data; type `who` to confirm that the vectors r and s are indeed available.

Using the assumed constraint equation, we can arrange the given information in the form of the linear system of (approximate) equations $Ax \approx b$, where A is a known 10×3 matrix, b is a known 10×1 vector, and $x = (\, x_1, x_2, x_3 \,)^T$. This system of 10 equations in 3 unknowns is inconsistent. We wish to find the solution x that minimizes the Euclidean norm (or length) of the error $Ax - b$. Compare the solutions obtained by using the following four Matlab invocations, each of which in principle gives the desired least-square-error solution:

(a) $x = A \backslash b$

(b) $x = \mathrm{pinv}(A) * b$

(c) $x = \mathrm{inv}(A' * A) * A' * b$

(d) $[q, r] = qr(A)$, followed by implementation of the approach described in Exercise 3.1.

For more information on these commands, try `help slash`, `help qr`, `help pinv`, `help inv`, etc. [Incidentally, the prime, $'$, in Matlab takes the transpose of the *complex conjugate* of a matrix; if you want the ordinary transpose of a complex matrix C, you have to write $C.'$ or `transp`(C).]

You should include in your solutions a plot the ellipse that corresponds to your estimate of x. If you create the following function file in your Matlab directory, with the name `ellipse.m`, you can obtain the polar coordinates `theta, rho` of n points on the ellipse specified by the parameter vector x. To do this, enter `[theta,rho]=ellipse(x,n);` at the Matlab prompt. You can then plot the ellipse by using the `polar(theta,rho)` command.

```
function [theta,rho]=ellipse(x,n)
% [theta,rho]=ellipse(x,n)
%
% The vector x = [x(1),x(2),x(3)]', defines an ellipse centered at the origin
% via the equation x(1)*r^ 2 + x(2)*s^ 2 +x(3)*r*s = 1.
% This routine generates the polar coordinates of points on the ellipse,
% to send to a plot command. It does this by solving for the radial
% distance in n equally spaced angular directions.
% Use polar(theta,rho) to actually plot the ellipse.
```

* Athena is MIT's UNIX-based computing environment. OCW does not provide access to it.

```
theta = 0:(2*pi/n):(2*pi);
a = x(1)*cos(theta).^ 2 + x(2)*sin(theta).^ 2 + x(3)*(cos(theta).*sin(theta));
rho = ones(size(a))./sqrt(a);
```

Exercise 2.2 Approximation by a Polynomial

Let $f(t) = 0.5e^{0.8t}$, $t \in [0, 2]$.

(a) Suppose 16 exact measurements of $f(t)$ are available to you, taken at the times t_i listed in the array T below:

$$T = [2 \cdot 10^{-3}, \quad 0.136, \quad 0.268, \quad 0.402, \quad 0.536, \quad 0.668, \quad 0.802, \quad 0.936,$$
$$1.068, \quad 1.202, \quad 1.336, \quad 1.468, \quad 1.602, \quad 1.736, \quad 1.868, \quad 2.000]$$

Use Matlab to generate these measurements:

$$y_i = f(t_i) \qquad i = 1, \ldots, 16 \qquad t_i \in T$$

Now determine the coefficients of the least square error polynomial approximation of the measurements, for

1. a polynomial of degree 15, $p_{15}(t)$;
2. a polynomial of degree 2, $p_2(t)$.

Compare the quality of the two approximations by plotting $y(t_i)$, $p_{15}(t_i)$ and $p_2(t_i)$ for all t_i in T. To see how well we are approximating the function on the whole interval, also plot $f(t)$, $p_{15}(t)$ and $p_2(t)$ on the interval $[0, 2]$. (Pick a very fine grid for the interval, e.g. t=[0:1000]'/500.) Report your observations and comments.

(b) Now suppose that your measurements are affected by some noise. Generate the measurements using

$$y_i = f(t_i) + e(t_i) \qquad i = 1, \ldots, 16 \qquad t_i \in T$$

where the vector of noise values can be generated in the following way:

$$randn('seed', 0);$$
$$e = randn(size(T));$$

Again determine the coefficients of the least square error polynomial approximation of the measurements for

1. a polynomial of degree 15, $p_{15}(t)$;
2. a polynomial of degree 2, $p_2(t)$.

Compare the two approximations as in part (a). Report your observations and comments. Explain any surprising results.

(c) So far we have obtained polynomial approximations of $f(t)$, $t \in [0,2]$, by approximating the measurements at $t_i \in T$. We are now interested in minimizing the square error of the polynomial approximation over the whole interval $[0,2]$:

$$\min \|f(t) - p_n(t)\|_2^2 = \min \int_0^2 |f(t) - p_n(t)|^2 \, dt$$

where $p_n(t)$ is some polynomial of degree n. Find the polynomial $p_2(t)$ of degree 2 that solves the above problem. Are the optimal $p_2(t)$ in this case and the optimal $p_2(t)$ of parts (a) and (b) very different from each other? Elaborate.

Exercise 2.3 Combining Estimates

Suppose $y_1 = C_1 x + e_1$ and $y_2 = C_2 x + e_2$, where x is an n-vector, and C_1, C_2 have full column rank. Let $\hat{x}_1$ denote the value of x that minimizes $e_1^T S_1 e_1$, and $\hat{x}_2$ denote the value that minimizes $e_2^T S_2 e_2$, where S_1 and S_2 are positive definite matrices. Show that the value $\hat{x}$ of x that minimizes $e_1^T S_1 e_1 + e_2^T S_2 e_2$ can be written entirely in terms of $\hat{x}_1$, $\hat{x}_2$, and the $n \times n$ matrices $Q_1 = C_1^T S_1 C_1$ and $Q_2 = C_2^T S_2 C_2$. What is the significance of this result?

Exercise 2.4 Exponentially Windowed Estimates

Suppose we observe the *scalar* measurements

$$y_i = c_i x + e_i \ , \qquad i = 1, 2, \ldots$$

where c_i and x are possibly vectors (row- and column-vectors respectively).

(a) Show (by reducing this to a problem that we already know how to solve — don't start from scratch!) that the value $\hat{x}_k$ of x that minimizes the criterion

$$\sum_{i=1}^{k} f^{k-i} e_i^2, \qquad \text{some fixed } f, \quad 0 < f \leq 1$$

is given by

$$\hat{x}_k = \left(\sum_{i=1}^{k} f^{k-i} c_i^T c_i \right)^{-1} \sum_{i=1}^{k} f^{k-i} c_i^T y_i$$

The so-called *fade* or *forgetting* factor f allows us to preferentially weight the more recent measurements by picking $0 < f < 1$, so that old data is discounted at an exponential rate. We then say that the data has been subjected to exponential fading or forgetting or weighting or windowing or tapering or ... . This is usually desirable, in order to keep the filter adaptive to changes that may occur in x. Otherwise the filter becomes progressively less attentive to new data and falls asleep, with its gain approaching 0.

(b) Now show that

$$\hat{x}_k = \hat{x}_{k-1} + Q_k^{-1} c_k^T (y_k - c_k \hat{x}_{k-1})$$

where

$$Q_k = f Q_{k-1} + c_k^T c_k, \qquad Q_0 = 0$$

The vector $g_k = Q_k^{-1} c_k^T$ is termed the *gain* of the estimator.

(c) If x and c_i are scalars, and c_i is a constant c, determine g_k as a function of k. What is the *steady-state gain* g_∞? Does g_∞ increase or decrease as f increases — and why do you expect this?

Exercise 2.5 Suppose our model for some waveform $y(t)$ is $y(t) = \alpha \sin(\omega t)$, where α is a scalar, and suppose we have measurements $y(t_1), \ldots, y(t_p)$. Because of modeling errors and the presence of measurement noise, we will generally not find any choice of model parameters that allows us to precisely account for all p measurements.

(a) If ω is known, find the value of α that minimizes

$$\sum_{i=1}^{p} [y(t_i) - \alpha \sin(\omega t_i)]^2$$

(b) Determine this value of α if $\omega = 2$ and if the measured values of $y(t)$ are:

$$y(1) = +2.31 \quad y(2) = -2.01 \quad y(3) = -1.33 \quad y(4) = +3.23$$

$$y(5) = -1.28 \quad y(6) = -1.66 \quad y(7) = +3.28 \quad y(8) = -0.88$$

(I generated this data using the equation $y(t) = 3\sin(2t) + e(t)$ evaluated at the integer values $t = 1, \ldots, 8$, and with $e(t)$ for each t being a random number uniformly distributed in the interval - 0.5 to +0.5.)

(c) Suppose that α *and* ω are unknown, and that we wish to determine the values of these two variables that minimize the above criterion. Assume you are given initial estimates α_0 and ω_0 for the minimizing values of these variables. Using the Gauss-Newton algorithm for this nonlinear least squares problem, i.e. applying LLSE to the problem obtained by linearizing about the initial estimates, determine explicitly the estimates α_1 and ω_1 obtained after one iteration of this algorithm. Use the following notation to help you write out the solution in a condensed form:

$$a = \sum \sin^2(\omega_0 t_i) \ , \quad b = \sum t_i^2 \cos^2(\omega_0 t_i) \ , \quad c = \sum t_i [\sin(w_0 t_i)][\cos(w_0 t_i)]$$

(d) What values do you get for α_1 and ω_1 with the data given in (b) above if the initial guesses are $\alpha_0 = 3.2$ and $\omega_0 = 1.8$? Continue the iterative estimation a few more steps. Repeat the procedure when the initial guesses are $\alpha_0 = 3.5$ and $\omega_0 = 2.5$, verifying that the algorithm does not converge.

(e) Since only ω enters the model nonlinearly, we might think of a decomposed algorithm, in which α is estimated using *linear* least squares and ω is estimated via nonlinear least squares. Suppose, for example, that our initial estimate of ω is $\omega_0 = 1.8$. Now obtain an estimate α_1 of α using the linear least squares method that you used in (b). Then obtain an (improved?) estimate ω_1 of ω, using one iteration of a Gauss-Newton algorithm (similar to what is needed in (c), except that now you are only trying to estimate ω). Next obtain the estimate α_2 via linear least squares, and so on. Compare your results with what you obtain via this decomposed procedure when your initial estimate is $\omega_0 = 2.5$ instead of 1.8.

Exercise 2.6 Comparing Different Estimators

This problem asks you to compare the behavior of different parameter estimation algorithms by fitting a model of the type $y(t) = a\sin(2\pi t) + b\cos(4\pi t)$ to noisy data taken at values of t that are .02 apart in the interval (0,2].

First synthesize the data on which you will test the algorithms. Even though your estimation algorithms will assume that a and b are constant, we are interested in seeing how they track parameter changes as well. Accordingly, let $a = 2$, $b = 2$ for the first 50 points, and $a = 1$, $b = 3$ for the next 50 points. To get (approximately) normally distributed random variables, we use the function *randn* to produce variables with mean 0 and variance 1.

An elegant way to generate the data in Matlab, exploiting Matlab's facility with vectors, is to define the vectors $t1 = 0.02 : 0.02 : 1.0$ and $t2 = 1.02 : 0.02 : 2.0$, then set

$$y1 = 2 * \sin(2 * \text{pi} * t1) + 2 * \cos(4 * \text{pi} * t1) + s * randn(size(t1))$$

and

$$y2 = \sin(2 * \text{pi} * t2) + 3 * \cos(4 * \text{pi} * t2) + s * randn(size(t2))$$

where s determines the standard deviation of the noise. Pick $s = 1$ for this problem. Finally, set $y = [y1, y2]$. No loops, no counters, no fuss!!

Now estimate a and b from y using the following algorithms. Assume prior estimates $\hat{a}_0 = 3$ and $\hat{b}_0 = 1$, weighted equally with the measurements (so all weights can be taken as 1 without loss of generality). Plot your results to aid comparison.

(i) Recursive least squares.

(ii) Recursive least squares with exponentially fading memory, as in Problem 3. Use $f = .96$.

(iii) The algorithm in (ii), but with Q_k of Problem 3 replaced by $q_k = (1/n)\times\text{trace}(Q_k)$, where n is the number of parameters, so $n = 2$ in this case. (Recall that the trace of a matrix is the sum of its diagonal elements. Note that q_k itself satisfies a recursion, which you should write down.)

(iv) An algorithm of the form

$$\hat{x}_k = \hat{x}_{k-1} + \frac{.04}{c_k c_k^T} c_k^T (y_k - c_k \hat{x}_{k-1})$$

where $c_k = [\sin(2\pi t), \cos(4\pi t)]$ evaluated at the kth sampling instant, so $t = .02k$.

Exercise 2.7 Recursive Estimation of a State Vector

This course will soon begin to consider *state-space models* of the form

$$x_\ell = Ax_{\ell-1} \tag{2.4}$$

where x_ℓ is an n-vector denoting the state at time ℓ of our model of some system, and A is a known $n \times n$ matrix. For example, suppose the system of interest is a rotating machine, with angular position d_ℓ and angular velocity ω_ℓ at time $t = \ell T$, where T is some fixed sampling interval. If we believed the machine to be rotating at constant speed, we would be led to the model

$$\begin{pmatrix} d_\ell \\ \omega_\ell \end{pmatrix} = \begin{pmatrix} 1 & T \\ 0 & 1 \end{pmatrix} \begin{pmatrix} d_{\ell-1} \\ \omega_{\ell-1} \end{pmatrix}$$

Assume A to be nonsingular throughout this problem.

For the rotating machine example above, it is often of interest to obtain least-square-error estimates of the position and (constant) velocity, using noisy measurements of the angular position d_j at the sampling instants. More generally, it is of interest to obtain a least-square-error estimate of the state vector x_i in the model (2.4) from noisy p-component measurements y_j that are related to x_j by a linear equation of the form

$$y_j = Cx_j + e_j \ , \qquad j = 1, \ldots, i$$

where C is a $p \times n$ matrix. We shall also assume that a prior estimate $\hat{x}_0$ of x_0 is available:

$$\hat{x}_0 = x_0 + e_0$$

Let $\hat{x}_{i|i}$ denote the value of x_i that minimizes

$$\sum_{j=0}^{i} \|e_j\|^2$$

This is the estimate of x_i given the prior estimate and measurements up to time i, or the "filtered estimate" of x_i. Similarly, let $\hat{x}_{i|i-1}$ denote the value of x_i that minimizes

$$\sum_{j=0}^{i-1} \|e_j\|^2$$

This is the least-square-error estimate of x_i given the prior estimate and measurements up to time $i-1$, and is termed the "one-step prediction" of x_i.

a) Set up the linear system of equations whose least square error solution would be $\hat{x}_{i|i}$. Similarly, set up the linear system of equations whose least square error solution would be $\hat{x}_{i|i-1}$.

b) Show that $\hat{x}_{i|i-1} = A\hat{x}_{i-1|i-1}$.

c) Determine a recursion that expresses $\hat{x}_{i|i}$ in terms of $\hat{x}_{i-1|i-1}$ and y_i. This is the prototype of what is known as the *Kalman filter*. A more elaborate version of the Kalman filter would include additive noise driving the state-space model, and other embellishments, all in a stochastic context (rather than the deterministic one given here).

Exercise 2.8 Let $\hat{x}$ denote the value of x that minimizes $\|y - Ax\|^2$, where A has full column rank. Let $\overline{x}$ denote the value of x that minimizes this same criterion, but now subject to the constraint that $z = Dx$, where D has full *row* rank. Show that

$$\overline{x} = \hat{x} + (A^TA)^{-1}D^T \left[D(A^TA)^{-1}D^T \right]^{-1}(z - D\hat{x})$$

(Hint: One approach to solving this is to use our recursive least squares formulation, but modified for the limiting case where one of the measurement sets — namely $z = Dx$ in this case — is known to have no error. You may have to use some of the matrix identities from the previous chapter).

Chapter 3

Least Squares Solution of $y = \prec A, x \succ$

3.1 Introduction

We turn to a problem that is dual to the overconstrained estimation problems considered so far. Let A denote an array of m vectors, $A = [a_1|\cdots|a_m]$, where the a_i are vectors from any space on which an inner product is defined. The space is allowed to be infinite dimensional, e.g. the space $\mathcal{L}^2$ of square integrable functions mentioned in Chapter 2 . We are interested in the vector x, *of minimum length*, that satisfy the equation

$$y = \prec A, x \succ \tag{1a}$$

where we have used the Gram product notation introduced in Chapter 2.

Example 3.1 Let $y[0]$ denote the output at time 0 of a noncausal FIR filter whose input is the sequence $x[k]$, with

$$y[0] = \sum_{i=-N}^{N} h_i x[-i].$$

Describe the set of input values that yield $y[0] = 0$; repeat for $y[0] = 7$. The solution of minimum energy (or RMS value) is the one that minimizes $\sum_{i=-N}^{N} x^2[i]$.

3.2 Constructing all Solutions

When the a_i's are drawn from ordinary (real or complex) Euclidean n-space, with the usual (unweighted) inner product, A is an $n \times m$ matrix of full column rank, and the equation (1a) is simply

$$y = A' x \quad , \tag{1b}$$

where A' has full row rank. Since the m rows of A' in (1b) are independent, this matrix has m independent columns as well. It follows that the system (1b), which can be read as expressing y in terms of a linear combination of the columns of A' (with weights given by the components of x) has solutions x for any y.

If A' were square and therefore (under our rank assumption) invertible, (1b) would have a unique solution, obtained simply by premultiplying (1b) by the inverse of A'. The closest we come to having an invertible matrix in the non-square case is by invoking the Gram matrix lemma, which tells us that $A'A$ is invertible under our rank assumption. This fact, and inspection of (1b), allow us to explicitly write down one particular solution of (1b), which we denote by $\check{x}$:

$$\check{x} = A\,(A'\,A)^{-1}\,y \tag{2a}$$

Simple substitution of this expression in (1b) verifies that it is indeed a solution. We shall shortly see that this solution actually has minimum length (norm) among all solutions of (1b).

For the more general equation in (1a), we can establish the existence of a solution by demonstrating that the appropriate generalization of the expression in (2a) does indeed satisfy (1a). For this, pick

$$\check{x} = A\ {\prec}A, A{\succ}^{-1}\,y \tag{2b}$$

It is easy to see that this satisfies (1a), if we use the fact that ${\prec}A, A\alpha{\succ}={\prec}A, A{\succ}\ \alpha$ for any array α of scalars; in our case α is the $m \times 1$ array ${\prec}A, A{\succ}^{-1}\ y$.

Any other x is a solution of (1a) iff it differs from the particular solution above (or any other particular solution) by a solution of the homogeneous equation ${\prec}A, x{\succ}= 0$; the same statement can be made for solutions of (1b). The proof is easy, and presented below for (1b), with x denoting any solution, x_p denoting a particular solution, and x_h denoting a solution of the homogeneous equation:

$$y = A'x_p = A'x \quad \Rightarrow \quad A'\underbrace{(x - x_p)}_{x_h} = 0 \quad \Rightarrow \quad x = x_p + x_h$$

Conversely,

$$y = A'x_p \qquad A'x_h = 0 \quad \Rightarrow \quad y = A'\underbrace{(x_p + x_h)}_{x} \quad \Rightarrow \quad x = x_p + x_h.$$

Equations of the form (1a), (1b) commonly arise in situations where x represents a vector of control inputs and y represents a vector of objectives or targets. The problem is then to use some appropriate criterion and/or constraints to narrow down the set of controls.

Example 3.2 Let $m = 1$, so that A' is a single nonzero row, which we shall denote by a'. If $y = 0$, the set of solutions corresponds to vectors x that are orthogonal to the vector a, i.e. to vectors in the orthogonal complement of a, namely in the subspace $\mathcal{R}a^{\perp}(a)$. Use this to costruct all solutions to Example 3.1.

There are several different criteria and constraints that may reasonably be used to select among the different possible solutions. For example, in some problems it may be natural to restrict the components x_i of x to be nonnegative, and to ask for the control that minimizes $\sum s_i x_i$, where s_i represents the cost of control component x_i. This is the prototypical form of what is termed the *linear programming* problem. (You should geometrically characterize the solution to this problem for the case given in the above example.) The general linear programming problem arises in a host of applications.

We shall focus on the problem of determining the solution x of (1a) for which $\|x\|^2 =< x, x >$ is minimized; in the case of (1b), we are looking to minimize $x'x$. For the situation depicted in the above example, the optimum x is immediately seen to be the solution vector that is aligned with a. It can be found by projecting any particular solution of (1b) onto the space spanned by the vector a. (This fact is related to the Cauchy-Schwartz inequality: For x of a specified length, the inner product $< a, x >$ is maximized by aligning x with a, and for specified $< a, x >$ the length of x is minimized by again aligning x with a.) The generalization to $m > 1$ and to the broader setting of (1a) is direct, and is presented next. You should note the similarity to the proof of the orthogonality principle.

3.3 Least Squares Solution

Let x be a particular solution of (1a). Denote by x_A its unique projection onto the range of A (i.e. onto the space spanned by the vectors a_i) and let $x_{A^\perp}$ denote the projection onto the space orthogonal to this. Following the same development as in the proof of the orthogonality principle in Lecture 2, we find

$$x_A = A \prec A, A\succ^{-1} \prec A, x\succ \tag{3a}$$

with $x_{A^\perp} = x - x_A$. Now (1a) allows us to make the substitution $y =\prec A, x\succ$ in (3a), so

$$x_A = A \prec A, A\succ^{-1} y \tag{3b}$$

which is exactly the expression we had for the solution $\check{x}$ that we determined earlier by inspection, see (2b).

Now note from (3b) that x_A is the same for all solutions x, because it is determined entirely by A and y. Hence it is only $x_{A^\perp}$ that is varied by varying x. The orthogonality of x_A and $x_{A^\perp}$ allows us to write

$$< x, x >=< x_A \ x_A > \ + < x_{A^\perp} \ x_{A^\perp} >$$

so the best we can do as far as minimizing $< x, x >$ is concerned is to make $x_{A^\perp} = 0$. In other words, the optimum solution is $x = x_A = \check{x}$.

Example 3.3 For the FIR filter mentioned in Example 3.1, and considering all input sequences $x[k]$ that result in $y[0] = 7$, find the sequence for which $\sum_{i=-N}^{N} x^2[i]$ is minimized. (Work out this example for yourself!)

Example 3.4 Consider a unit mass moving in a straight line under the action of a force $x(t)$, with position at time t given by $p(t)$. Assume $p(0) = 0$, $\dot{p}(0) = 0$, and suppose we wish to have $p(T) = y$ (with no constraint on $\dot{p}(T)$). Then

$$y = p(T) = \int_0^T (T - t)x(t)\, dt =< a(t)\ \ x(t) > \tag{4}$$

This is a typical underconstrained problem, with many choices of $x(t)$ for $0 \leq t \leq T$ that will result in $p(T) = y$. Let us find the solution $x(t)$ for which

$$\int_0^T x^2(t)\, dt\ =< x(t)\ \ x(t) > \tag{5}$$

is minimized. Evaluating the expression in (2a), we find

$$\check{x}(t) = (T - t)y/(T^3/3) \tag{6}$$

How does your solution change if there is the additional constraint that the mass should be brought to rest at time T, so that $\dot{p}(T) = 0$?

We leave you to consider how *weighted* norms can be minimized.

Exercises

Exercise 3.1 Least Square Error Solution We begin with a mini-tutorial on orthogonal and unitary matrices. An *orthogonal matrix* may be defined as a *square* real matrix whose columns are of unit length and mutually orthogonal to each other — i.e., its columns form an *orthonormal* set. It follows quite easily (as you should try and verify for yourself) that:

- the *inverse* of an orthogonal matrix is just its transpose;
- the *rows* of an orthogonal matrix form an orthonormal set as well;
- the usual Euclidean *inner product* of two real vectors v and w, namely the scalar $v'w$, equals the inner product of Uv and Uw, if U is an orthogonal matrix — and therefore the length of v, namely $\sqrt{v'v}$, equals that of Uv.

A *unitary matrix* is similarly defined, except that its entries are allowed to be complex — so its inverse is the *complex conjugate* of its transpose. A fact about orthogonal matrices that turns out to be important in several numerical algorithms is the following: Given a real $m \times n$ matrix A of full column rank, it is possible (in many ways) to find an orthogonal matrix U such that

$$UA = \begin{pmatrix} R \\ 0 \end{pmatrix}$$

where R is a nonsingular, upper-triangular matrix. (If A is *complex*, then we can find a *unitary* matrix U that leads to the same equation.) To see how to compute U in Matlab, read the comments obtained by typing `help qr`; the matrix Q that is referred to in the comments is just U'.

We now turn to the problem of interest. Given a real $m \times n$ matrix A of full column rank, and a real m-vector y, we wish to approximately satisfy the equation $y = Ax$. Specifically, let us choose the vector x to minimize $\|y - Ax\|^2 = (y - Ax)'(y - Ax)$, the squared Euclidean length of the "error" $y - Ax$. By invoking the above results on orthogonal matrices, **show that** (in the notation introduced earlier) the minimizing x is

$$\hat{x} = R^{-1} y_1$$

where y_1 denotes the vector formed from the first n components of Uy. (In practice, we would not bother to find R^{-1} explicitly. Instead, taking advantage of the upper-triangular structure of R, we would solve the system of equations $R\hat{x} = y_1$ by back substitution, starting from the last equation.)

The above way of solving a least-squares problem (proposed by Householder in 1958, but sometimes referred to as Golub's algorithm) is numerically preferable in most cases to solving the "normal equations" in the form $\hat{x} = (A'A)^{-1}A'y$, and is essentially what Matlab does when you write $\hat{x} = A\backslash y$. An (oversimplified!) explanation of the trouble with the normal equation solution is that it implicitly evaluates the product $(R'R)^{-1}R'$, whereas the Householder/Golub method recognizes that this product simply equals R^{-1}, and thereby avoids unnecessary and error prone steps.

Exercise 3.2 Suppose the input sequence $\{u_j\}$ and the output sequence $\{y_j\}$ of a particular system are related by

$$y_k = \sum_{i=1}^{n} h_i u_{k-i}$$

where all quantities are scalar.
(i) Assume we want to have y_n equal to some specified number $\bar{y}$. Determine $u_0, \ldots, u_{n-1}$ so as to achieve this while minimizing $u_0^2 + \ldots + u_{n-1}^2$.
(ii) Suppose now that we are willing to relax our objective of exactly attaining $y_n = \bar{y}$. This leads us to the following modified problem. Determine $u_0, \ldots, u_{n-1}$ so as to minimize

$$r(\bar{y} - y_n)^2 + u_0^2 + \ldots + u_{n-1}^2$$

where r is a positive weighting parameter.

(a) Solve the modified problem.

(b) What do you expect the answer to be in the limiting cases of $r = 0$ and $r = \infty$? Show that your answer in (a) indeed gives you these expected limiting results.

Exercise 3.3 Return to the problem considered in Example 3.4. Suppose that, in addition to requiring $p(T) = y$ for a specified y, we also want $\dot{p}(T) = 0$. In other words, we want to bring the mass *to rest* at the position y at time T. Of all the force functions $x(t)$ that can accomplish this, determine the one that minimizes $< x(t), x(t) >= \int_0^T x^2(t)\, dt$.

Exercise 3.4 **(a)** Given $y = A'x$, with A' of full row rank, find the solution vector x for which $x'Wx$ is minimum, where $W = L'L$ and L is nonsingular (i.e. where W is Hermitian and positive definite).

(b) A specified current I_0 is to be sent through the fixed voltage source V_0 in the figure. Find what values v_1, v_2, v_3 and v_4 must take so that the total power dissipation in the resistors is minimized.

Chapter 4

Matrix Norms and Singular Value Decomposition

4.1 Introduction

In this lecture, we introduce the notion of a *norm* for matrices. The *singular value decomposition* or SVD of a matrix is then presented. The SVD exposes the 2-norm of a matrix, but its value to us goes much further: it enables the solution of a class of *matrix perturbation problems* that form the basis for the stability robustness concepts introduced later; it solves the so-called *total least squares* problem, which is a generalization of the least squares estimation problem considered earlier; and it allows us to clarify the notion of *conditioning*, in the context of matrix inversion. These applications of the SVD are presented at greater length in the next lecture.

Example 4.1 To provide some immediate motivation for the study and application of matrix norms, we begin with an example that clearly brings out the issue of matrix conditioning with respect to inversion. The question of interest is how sensitive the inverse of a matrix is to perturbations of the matrix.

Consider inverting the matrix

$$A = \begin{pmatrix} 100 & 100 \\ 100.2 & 100 \end{pmatrix} \tag{4.1}$$

A quick calculation shows that

$$A^{-1} = \begin{pmatrix} -5 & 5 \\ 5.01 & -5 \end{pmatrix} \tag{4.2}$$

Now suppose we invert the perturbed matrix

$$A + \Delta A = \begin{pmatrix} 100 & 100 \\ 100.1 & 100 \end{pmatrix} \tag{4.3}$$

The result now is

$$(A + \Delta A)^{-1} = A^{-1} + \Delta(A^{-1}) = \begin{pmatrix} -10 & 10 \\ 10.01 & -10 \end{pmatrix} \tag{4.4}$$

Here ΔA denotes the perturbation in A and $\Delta(A^{-1})$ denotes the resulting perturbation in A^{-1}. Evidently a 0.1% change in one entry of A has resulted in a 100% change in the entries of A^{-1}. If we want to solve the problem $Ax = b$ where $b = [1 \ \ -1]^T$, then $x = A^{-1}b = [-10 \ \ 10.01]^T$, while after perturbation of A we get $x + \Delta x = (A + \Delta A)^{-1}b = [-20 \ \ 20.01]^T$. Again, we see a 100% change in the entries of the solution with only a 0.1% change in the starting data.

The situation seen in the above example is much worse than what can ever arise in the scalar case. If a is a scalar, then $d(a^{-1})/(a^{-1}) = -da/a$, so the fractional change in the inverse of a has the same maginitude as the fractional change in a itself. What is seen in the above example, therefore, is a purely matrix phenomenon. It would seem to be related to the fact that A is nearly singular — in the sense that its columns are nearly dependent, its determinant is much smaller than its largest element, and so on. In what follows (see next lecture), we shall develop a sound way to measure nearness to singularity, and show how this measure relates to sensitivity under inversion.

Before understanding such sensitivity to perturbations in more detail, we need ways to measure the "magnitudes" of vectors and matrices. We have already introduced the notion of vector norms in Lecture 1, so we now turn to the definition of matrix norms.

4.2 Matrix Norms

An $m \times n$ complex matrix may be viewed as an operator on the (finite dimensional) normed vector space $\mathbb{C}^n$:

$$A^{m \times n} \ : \ (\ \mathbb{C}^n, \|\cdot\|_2\ ) \longrightarrow (\ \mathbb{C}^m, \|\cdot\|_2\ ) \tag{4.5}$$

where the norm here is taken to be the standard Euclidean norm. Define the **induced 2-norm** of A as follows:

$$\|A\|_2 \ \stackrel{\triangle}{=} \ \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \tag{4.6}$$

$$= \ \max_{\|x\|_2 = 1} \|Ax\|_2 \ . \tag{4.7}$$

The term "induced" refers to the fact that the definition of a norm for *vectors* such as Ax and x is what enables the above definition of a *matrix* norm. From this definition, it follows that the induced norm measures the amount of "amplification" the matrix A provides to vectors on the unit sphere in $\mathbb{C}^n$, i.e. it measures the "gain" of the matrix.

Rather than measuring the vectors x and Ax using the 2-norm, we could use any p-norm, the interesting cases being $p = 1, 2, \infty$. Our notation for this is

$$\|A\|_p = \max_{\|x\|_p = 1} \|Ax\|_p \ . \tag{4.8}$$

An important question to consider is whether or not the induced norm is actually a norm, in the sense defined for vectors in Lecture 1. Recall the three conditions that define a norm:

1. $\|x\| \geq 0$, and $\|x\| = 0 \Longleftrightarrow x = 0$;
2. $\|\alpha x\| = |\alpha| \ \|x\|$;
3. $\|x + y\| \leq \|x\| + \|y\|$.

Now let us verify that $\|A\|_p$ is a norm on $\mathbb{C}^{m\times n}$ — using the preceding definition:

1. $\|A\|_p \geq 0$ since $\|Ax\|_p \geq 0$ for any x. Futhermore, $\|A\|_p = 0 \Longleftrightarrow A = 0$, since $\|A\|_p$ is calculated from the *maximum* of $\|Ax\|_p$ evaluated on the unit sphere.
2. $\|\alpha A\|_p = |\alpha| \ \|A\|_p$ follows from $\|\alpha y\|_p = |\alpha| \ \|y\|_p$ (for any y).
3. The triangle inequality holds since:

$$\begin{aligned} \|A + B\|_p &= \max_{\|x\|_p=1} \|(A + B)x\|_p \\ &\leq \max_{\|x\|_p=1} \left(\|Ax\|_p + \|Bx\|_p\right) \\ &\leq \|A\|_p + \|B\|_p \ . \end{aligned}$$

Induced norms have two additional properties that are very important:

1. $\|Ax\|_p \leq \|A\|_p \|x\|_p$, which is a direct consequence of the definition of an induced norm;
2. For $A^{m\times n}$, $B^{n\times r}$,

$$\|AB\|_p \leq \|A\|_p \|B\|_p \tag{4.9}$$

which is called the *submultiplicative property*. This also follows directly from the definition:

$$\begin{aligned} \|ABx\|_p &\leq \|A\|_p \|Bx\|_p \\ &\leq \|A\|_p \|B\|_p \|x\|_p \ \text{ for any } x. \end{aligned}$$

Dividing by $\|x\|_p$:

$$\frac{\|ABx\|_p}{\|x\|_p} \leq \|A\|_p \|B\|_p \ ,$$

from which the result follows.

Before we turn to a more detailed study of ideas surrounding the induced 2-norm, which will be the focus of this lecture and the next, we make some remarks about the other induced norms of practical interest, namely the induced 1-norm and induced ∞-norm. We shall also

say something about an important matrix norm that is *not* an induced norm, namely the *Frobenius* norm.

It is a fairly simple exercise to prove that

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}| \quad \text{(max of absolute column sums of } A\text{)} \tag{4.10}$$

and

$$\|A\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| \quad \text{(max of absolute row sums of } A\text{)} \ . \tag{4.11}$$

(Note that these definitions reduce to the familiar ones for the 1-norm and ∞-norm of *column vectors* in the case $n = 1$.)

The proof for the induced ∞-norm involves two stages, namely:

1. Prove that the quantity in Equation (4.11) provides an upper bound γ:

$$\|Ax\|_\infty \le \gamma \|x\|_\infty \quad \forall x \quad ;$$

2. Show that this bound is achievable for some $x = \hat{x}$:

$$\|A\hat{x}\|_\infty = \gamma \|\hat{x}\|_\infty \quad \text{for some } \hat{x} \quad .$$

In order to show how these steps can be implemented, we give the details for the ∞-norm case. Let $x \in \mathbb{C}^n$ and consider

$$\begin{aligned}
\|Ax\|_\infty &= \max_{1\le i\le m} |\sum_{j=1}^{n} a_{ij}x_j| \\
&\le \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}||x_j| \\
&\le \left(\max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| \right) \max_{1\le j\le n} |x_j| \\
&= \left(\max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| \right) \|x\|_\infty
\end{aligned}$$

The above inequalities show that an upper bound γ is given by

$$\max_{\|x\|_\infty = 1} \|Ax\|_\infty \le \gamma = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|.$$

Now in order to show that this upper bound is achieved by some vector $\hat{x}$, let $\bar{i}$ be an index at which the expression of γ achieves a maximum, that is $\gamma = \sum_{j=1}^{n} |a_{\bar{i}j}|$. Define the vector $\hat{x}$ as

$$\hat{x} = \begin{bmatrix} sgn(a_{\bar{i}1}) \\ sgn(a_{\bar{i}2}) \\ \vdots \\ sgn(a_{\bar{i}n}) \end{bmatrix}.$$

Clearly $\|\hat{x}\|_\infty = 1$ and

$$\|A\hat{x}\|_\infty = \sum_{j=1}^{n} |a_{\bar{i}j}| = \gamma.$$

The proof for the 1-norm proceeds in exactly the same way, and is left to the reader.

There are matrix norms — i.e. functions that satisfy the three defining conditions stated earlier — that are *not* induced norms. The most important example of this for us is the **Frobenius norm**:

$$\|A\|_F \triangleq \left(\sum_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}|^2 \right)^{\frac{1}{2}} \tag{4.12}$$

$$= (\text{trace}(A'A))^{\frac{1}{2}} \quad \text{(verify)} \tag{4.13}$$

In other words, the Frobenius norm is defined as the root sum of squares of the entries, i.e. the usual Euclidean 2-norm of the matrix when it is regarded simply as a vector in $\mathbb{C}^{mn}$. Although it can be shown that it is not an induced matrix norm, the Frobenius norm still has the submultiplicative property that was noted for induced norms. Yet other matrix norms may be defined (some of them without the submultiplicative property), but the ones above are the only ones of interest to us.

4.3 Singular Value Decomposition

Before we discuss the singular value decomposition of matrices, we begin with some matrix facts and definitions.

Some Matrix Facts:

- A matrix $U \in \mathbb{C}^{n \times n}$ is unitary if $U'U = UU' = I$. Here, as in Matlab, the superscript $'$ denotes the (entry-by-entry) *complex conjugate* of the *transpose*, which is also called the *Hermitian transpose* or *conjugate transpose*.
- A matrix $U \in \mathbb{R}^{n \times n}$ is orthogonal if $U^T U = UU^T = I$, where the superscript ^{T} denotes the transpose.
- Property: If U is unitary, then $\|Ux\|_2 = \|x\|_2$.

- If $S = S'$ (*i.e.* S equals its Hermitian transpose, in which case we say S is *Hermitian*), then there exists a unitary matrix such that $U'SU =$ [diagonal matrix].[1]
- For any matrix A, both $A'A$ and AA' are Hermitian, and thus can always be diagonalized by unitary matrices.
- For any matrix A, the eigenvalues of $A'A$ and AA' are always real and non-negative (proved easily by contradiction).

Theorem 4.1 (Singular Value Decomposition, or SVD) Given any matrix $A \in \mathbb{C}^{m\times n}$, A can be written as

$$A = \overset{m\times m}{U}\ \overset{m\times n}{\Sigma}\ \overset{n\times n}{V'} \tag{4.14}$$

where $U'U = I$, $V'V = I$,

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \tag{4.15}$$

and $\sigma_i = \sqrt{i\text{th nonzero eigenvalue of } A'A}$. The σ_i are termed the **singular values** of A, and are arranged in order of descending magnitude, *i.e.*,

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0 \ .$$

Proof: We will prove this theorem for the case $\text{rank}(A) = m$; the general case involves very little more than what is required for this case. The matrix AA' is Hermitian, and it can therefore be diagonalized by a unitary matrix $U \in \mathbb{C}^{m\times m}$, so that

$$U\Lambda_1 U' = AA'.$$

Note that $\Lambda_1 = diag(\lambda_1\ \lambda_2\ \ldots\ \lambda_m)$ has real positive diagonal entries λ_i due to the fact that AA' is positive definite. We can write $\Lambda_1 = \Sigma_1^2 = diag(\sigma_1^2\ \sigma_2^2\ \ldots\ \sigma_m^2)$. Define $V_1' \in \mathbb{C}^{m\times n}$ by $V_1' = \Sigma_1^{-1}U'A$. V_1' has orthonormal rows as can be seen from the following calculation: $V_1'V_1 = \Sigma_1^{-1}U'AA'U\Sigma_1^{-1} = I$. Choose the matrix V_2' in such a way that

$$V' = \left[\begin{array}{c} V_1' \\ V_2' \end{array}\right]$$

is in $\mathbb{C}^{n\times n}$ and unitary. Define the $m \times n$ matrix $\Sigma = [\Sigma_1|0]$. This implies that

$$\Sigma V' = \Sigma_1 V_1' = U'A.$$

In other words we have $A = U\Sigma V'$.

[1] One cannot always diagonalize an arbitrary matrix—cf the Jordan form.

Example 4.2 For the matrix A given at the beginning of this lecture, the SVD — computed easily in Matlab by writing $[u, s, v] = svd(A)$ — is

$$A = \begin{pmatrix} .7068 & .7075 \\ .7075 & .-7068 \end{pmatrix} \begin{pmatrix} 200.1 & 0 \\ 0 & 0.1 \end{pmatrix} \begin{pmatrix} .7075 & .7068 \\ -.7068 & .7075 \end{pmatrix} \tag{4.16}$$

Observations:

i)

$$\begin{aligned} AA' &= U\Sigma V' V \Sigma^T U' \\ &= U\Sigma\Sigma^T U' \\ &= U \left[\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right] U' \end{aligned} \tag{4.17}$$

which tells us U diagonalizes AA';

ii)

$$\begin{aligned} A'A &= V\Sigma^T U' U \Sigma V' \\ &= V\Sigma^T\Sigma V' \\ &= V \left[\begin{array}{ccc|c} \sigma_1^2 & & & \\ & \ddots & & 0 \\ & & \sigma_r^2 & \\ \hline & 0 & & 0 \end{array}\right] V' \end{aligned} \tag{4.18}$$

which tells us V diagonalizes $A'A$;

iii) If U and V are expressed in terms of their columns, *i.e.*,

$$U = \begin{bmatrix} u_1 & u_2 & \cdots & u_m \end{bmatrix}$$

and

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

then

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i' \tag{4.19}$$

which is another way to write the SVD. The u_i are termed the **left singular vectors** of A, and the v_i are its **right singular vectors**. From this we see that we can alternately interpret Ax as

$$Ax = \sum_{i=1}^{r} \sigma_i \, u_i \underbrace{(v_i' x)}_{\text{projection}} \tag{4.20}$$

which is a weighted sum of the u_i, where the weights are the products of the singular values and the projections of x onto the v_i.

Observation (iii) tells us that $\mathcal{R}a(A) = \text{span}\{u_1 \ \ldots \ u_r\}$ (because $Ax = \sum_{i=1}^{r} c_i u_i$ — where the c_i are scalar weights). Since the columns of U are independent, $\dim \mathcal{R}a(A) = r = \text{rank}\ (A)$, and $\{u_1 \ \ldots \ u_r\}$ constitute a *basis* for the range space of A. The null space of A is given by $\text{span}\{v_{r+1} \ \ldots \ v_n\}$. To see this:

$$\begin{aligned}
U\Sigma V'x = 0 \quad &\Longleftrightarrow \quad \Sigma V'x = 0 \\
&\Longleftrightarrow \quad \begin{bmatrix} \sigma_1 v_1' x \\ \vdots \\ \sigma_r v_r' x \end{bmatrix} = 0 \\
&\Longleftrightarrow \quad v_i' x = 0 \qquad i = 1 \ \ldots \ r \\
&\Longleftrightarrow \quad x \in \text{span}\{v_{r+1} \ \ldots \ v_n\}.
\end{aligned}$$

Example 4.3 One application of singular value decomposition is to the solution of a system of algebraic equations. Suppose A is an $m \times n$ complex matrix and b is a vector in $\mathbb{C}^m$. Assume that the rank of A is equal to k, with $k < m$. We are looking for a solution of the linear system $Ax = b$. By applying the singular value decomposition procedure to A, we get

$$\begin{aligned}
A &= U\Sigma V' \\
&= U \left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & 0 \end{array}\right] V'
\end{aligned}$$

where Σ_1 is a $k \times k$ non-singular diagonal matrix. We will express the unitary matrices U and V columnwise as

$$\begin{aligned}
U &= \begin{bmatrix} u_1 & u_2 & \ldots & u_m \end{bmatrix} \\
V &= \begin{bmatrix} v_1 & v_2 & \ldots & v_n \end{bmatrix}.
\end{aligned}$$

A necessary and sufficient condition for the solvability of this system of equations is that $u_i' b = 0$ for all i satisfying $k < i \leq m$. Otherwise, the system of equations is inconsistent. This condition means that the vector b must be orthogonal to the

last $m-k$ columns of U. Therefore the system of linear equations can be written as

$$\left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & 0 \end{array}\right] V'x = U'b$$

$$\left[\begin{array}{c|c} \Sigma_1 & 0 \\ \hline 0 & 0 \end{array}\right] V'x = \begin{bmatrix} u_1'b \\ u_2'b \\ \vdots \\ \vdots \\ u_m'b \end{bmatrix} = \begin{bmatrix} u_1'b \\ \vdots \\ u_k'b \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Using the above equation and the invertibility of Σ_1, we can rewrite the system of equations as

$$\begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_k' \end{bmatrix} x = \begin{bmatrix} \frac{1}{\sigma_1}u_1'b \\ \frac{1}{\sigma_2}u_2'b \\ \cdots \\ \frac{1}{\sigma_k}u_k'b \end{bmatrix}$$

By using the fact that

$$\begin{bmatrix} v_1' \\ v_2' \\ \vdots \\ v_k' \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \ldots & v_k \end{bmatrix} = I$$

we obtain a solution of the form

$$x = \sum_{i=1}^{k} \frac{1}{\sigma_i} u_i'b \; v_i.$$

From the observations that were made earlier, we know that the vectors $v_{k+1} \; v_{k+2} \; \ldots \; v_n$ span the kernel of A, and therefore a general solution of the system of linear equations is given by

$$x = \sum_{i=1}^{k} \frac{1}{\sigma_i}(u_i'b) \; v_i + \sum_{i=k+1}^{n} \beta_i v_i$$

where the coefficients β_i, with i in the interval $k+1 \leq i \leq n$, are arbitrary complex numbers.

4.4 Relationship to Matrix Norms

The singular value decomposition can be used to compute the induced 2-norm of a matrix A.

Theorem 4.2

$$\begin{aligned} \|A\|_2 &\triangleq \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} \\ &= \sigma_1 \\ &= \sigma_{max}(A) \end{aligned} \tag{4.21}$$

which tells us that the maximum amplification is given by the maximum singular value.

Proof:

$$\begin{aligned} \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} &= \sup_{x\neq 0} \frac{\|U\Sigma V'x\|_2}{\|x\|_2} \\ &= \sup_{x\neq 0} \frac{\|\Sigma V'x\|_2}{\|x\|_2} \\ &= \sup_{y\neq 0} \frac{\|\Sigma y\|_2}{\|Vy\|_2} \\ &= \sup_{y\neq 0} \frac{\sum_{i=1}^r \sigma_i^2 |y_i|^2 \ ^{\frac{1}{2}}}{\sum_{i=1}^r |y_i|^2 \ ^{\frac{1}{2}}} \\ &\leq \sigma_1 \ . \end{aligned}$$

For $y = [1\ 0\ \cdots\ 0]^T$, $\|\Sigma y\|_2 = \sigma_1$, and the supremum is attained. (Notice that this correponds to $x = v_1$. Hence, $Av_1 = \sigma_1 u_1$.)

Another application of the singular value decomposition is in computing the *minimal* amplification a full rank matrix exerts on elements with 2-norm equal to 1.

Theorem 4.3 Given $A \in \mathbb{C}^{m\times n}$, suppose $\text{rank}(A) = n$. Then

$$\min_{\|x\|_2=1} \|Ax\|_2 = \sigma_n(A) \ . \tag{4.22}$$

Note that if $\text{rank}(A) < n$, then there is an x such that the minimum is zero (rewrite A in terms of its SVD to see this).

Proof: For any $\|x\|_2 = 1$,

$$\begin{aligned} \|Ax\|_2 &= \|U\Sigma V'x\|_2 \\ &= \|\Sigma V'x\|_2 \ \text{(invariant under multiplication by unitary matrices)} \\ &= \|\Sigma y\|_2 \end{aligned}$$

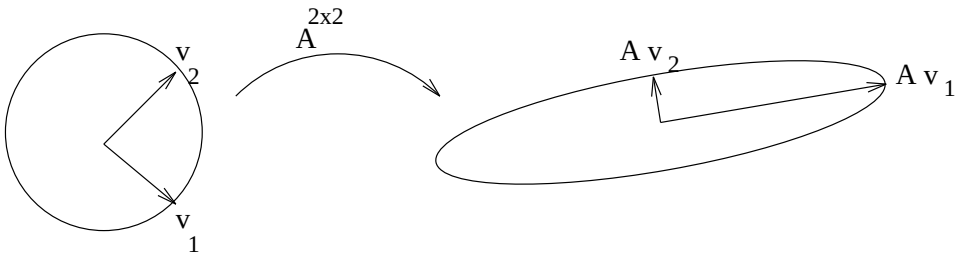


Figure 4.1: Graphical depiction of the mapping involving $A^{2\times 2}$. Note that $Av_1 = \sigma_1 u_1$ and that $Av_2 = \sigma_2 u_2$.

for $y = V'x$. Now

$$\begin{aligned} \|\Sigma y\|_2 &= \left(\sum_{i=1}^{n} |\sigma_i y_i|^2\right)^{\frac{1}{2}} \\ &\geq \sigma_n \ . \end{aligned}$$

Note that the minimum is achieved for $y = [0\ 0\ \cdots\ 0\ 1]^T$; thus the proof is complete.

The Frobenius norm can also be expressed quite simply in terms of the singular values. We leave you to verify that

$$\begin{aligned} \|A\|_F &\triangleq \left(\sum_{j=1}^{n}\sum_{i=1}^{m} |a_{ij}|^2\right)^{\frac{1}{2}} \\ &= (\text{trace}(A'A)\ )^{\frac{1}{2}} \\ &= \left(\sum_{i=1}^{r} \sigma_i^2\right)^{\frac{1}{2}} \end{aligned} \tag{4.23}$$

Example 4.4 Matrix Inequality

We say $A \leq B$, two square matrices, if

$$x'Ax \leq x'Bx \qquad \text{for all } x \neq 0.$$

It follows that for any matrix A, not necessarily square,

$$\|A\|_2 \leq \gamma \leftrightarrow A'A \leq \gamma^2 I.$$

Exercises

Exercise 4.1 Verify that for any A, an $m \times n$ matrix, the following holds:

$$\frac{1}{\sqrt{n}}\|A\|_1 \leq \|A\|_2 \leq \sqrt{m}\|A\|_\infty.$$

Exercise 4.2 Suppose $A' = A$. Find the exact relation between the eigenvalues and singular values of A. Does this hold if A is not conjugate symmetric?

Exercise 4.3 Show that if $\text{rank}(A) = 1$, then, $\|A\|_F = \|A\|_2$.

Exercise 4.4 This problem leads you through the argument for the existence of the SVD, using an iterative construction. Showing that $A = U\Sigma V'$, where U and V are unitary matrices is equivalent to showing that $U'AV = \Sigma$.

a) Argue from the definition of $\|A\|_2$ that there exist unit vectors (measured in the 2-norm) $x \in \mathbb{C}^n$ and $y \in \mathbb{C}^m$ such that $Ax = \sigma y$, where $\sigma = \|A\|_2$.

b) We can extend both x and y above to orthonormal bases, i.e. we can find *unitary matrices* V_1 and U_1 whose first columns are x and y respectively:

$$V_1 = [x \;\; \tilde{V}_1] \;, \qquad U_1 = [y \;\; \tilde{U}_1]$$

Show that one way to do this is via Householder transformations, as follows:

$$V_1 = I - 2\frac{hh'}{h'h} \;, \qquad h = x - [1, 0, \ldots, 0]'$$

and likewise for U_1.

c) Now define $A_1 = U_1'AV_1$. Why is $\|A_1\|_2 = \|A\|_2$?

d) Note that

$$A_1 = \begin{pmatrix} y'Ax & y'A\tilde{V}_1 \\ \tilde{U}_1'Ax & \tilde{U}_1'A\tilde{V}_1 \end{pmatrix} = \begin{pmatrix} \sigma & w' \\ 0 & B \end{pmatrix}$$

What is the justification for claiming that the lower left element in the above matrix is 0?

e) Now show that

$$\|A_1 \begin{pmatrix} \sigma \\ w \end{pmatrix} \|_2 \geq \sigma^2 + w'w$$

and combine this with the fact that $\|A_1\|_2 = \|A\|_2 = \sigma$ to deduce that $w = 0$, so

$$A_1 = \begin{pmatrix} \sigma & 0 \\ 0 & B \end{pmatrix}$$

At the next iteration, we apply the above procedure to B, and so on. When the iterations terminate, we have the SVD.

[The reason that this is only an existence proof and not an algorithm is that it begins by invoking the existence of x and y, but does not show how to compute them. Very good algorithms do exist for computing the SVD — see Golub and Van Loan's classic, *Matrix Computations*, Johns Hopkins Press, 1989. The SVD is a cornerstone of numerical computations in a host of applications.]

Exercise 4.5 Suppose the $m \times n$ matrix A is decomposed in the form

$$A = U \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} V'$$

where U and V are unitary matrices, and Σ is an invertible $r \times r$ matrix (— the SVD could be used to produce such a decomposition). Then the "Moore-Penrose inverse", or *pseudo-inverse* of A, denoted by A^+, can be defined as the $n \times m$ matrix

$$A^+ = V \begin{pmatrix} \Sigma^{-1} & 0 \\ 0 & 0 \end{pmatrix} U'$$

(You can invoke it in Matlab with `pinv(A)`.)

a) Show that A^+A and AA^+ are symmetric, and that $AA^+A = A$ and $A^+AA^+ = A^+$. (These four conditions actually constitute an alternative definition of the pseudo-inverse.)

b) Show that when A has full column rank then $A^+ = (A'A)^{-1}A'$, and that when A has full row rank then $A^+ = A'(AA')^{-1}$.

c) Show that, of all x that minimize $\|y - Ax\|_2$ (and there will be many, if A does not have full column rank), the one with smallest length $\|x\|_2$ is given by $\hat{x} = A^+y$.

Exercise 4.6 All the matrices in this problem are real. Suppose

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix}$$

with Q being an $m \times m$ orthogonal matrix and R an $n \times n$ invertible matrix. (Recall that such a decomposition exists for any matrix A that has full column rank.) Also let Y be an $m \times p$ matrix of the form

$$Y = Q \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$

where the partitioning in the expression for Y is conformable with the partitioning for A.

(a) What choice $\hat{X}$ of the $n \times p$ matrix X minimizes the Frobenius norm, or equivalently the squared Frobenius norm, of $Y - AX$? In other words, find

$$\hat{X} = \text{argmin } \|Y - AX\|_F^2$$

Also determine the value of $\|Y - A\hat{X}\|_F^2$. (Your answers should be expressed in terms of the matrices Q, R, Y_1 and Y_2.)

(b) Can your $\hat{X}$ in (a) also be written as $(A'A)^{-1}A'Y$? Can it be written as A^+Y, where A^+ denotes the (Moore-Penrose) pseudo-inverse of A ?

(c) Now obtain an expression for the choice $\overline{X}$ of X that minimizes

$$\|Y - AX\|_F^2 + \|Z - BX\|_F^2$$

where Z and B are given matrices of appropriate dimensions. (Your answer can be expressed in terms of A, B, Y, and Z.)

Exercise 4.7 Structured Singular Values

Given a complex square matrix A, define the *structured singular value function* as follows.

$$\mu_{\underline{\Delta}}(A) = \frac{1}{\min_{\Delta \in \underline{\Delta}}\{\sigma_{max}(\Delta) \mid \det(I - \Delta A) = 0\}}$$

where $\underline{\Delta}$ is some set of matrices.

a) If $\underline{\Delta} = \{\alpha I : \alpha \in \mathbb{C}\}$, show that $\mu_{\underline{\Delta}}(A) = \rho(A)$, where ρ is the *spectral radius* of A, defined as: $\rho(A) = \max_i |\lambda_i|$ and the λ_i's are the eigenvalues of A.

b) If $\underline{\Delta} = \{\Delta \in \mathbb{C}^{n \times n}\}$, show that $\mu_{\underline{\Delta}}(A) = \sigma_{max}(A)$

c) If $\underline{\Delta} = \{diag(\alpha_1, \cdots, \alpha_n) \mid \alpha_i \in \mathbb{C}\}$, show that

$$\rho(A) \leq \mu_{\underline{\Delta}}(A) = \mu_{\underline{\Delta}}(D^{-1}AD) \leq \sigma_{max}(D^{-1}AD)$$

where

$$D \in \{\text{diag}(d_1, \cdots, d_n) \mid d_i > 0\}$$

Exercise 4.8 Consider again the *structured singular value function* of a complex square matrix A defined in the preceding problem. If A has more structure, it is sometimes possible to compute $\mu_{\underline{\Delta}}(A)$ exactly. In this problem, assume A is a rank-one matrix, so that we can write $A = uv'$ where u, v are complex vectors of dimension n. Compute $\mu_{\underline{\Delta}}(A)$ when

(a) $\underline{\Delta} = diag(\delta_1, \ldots, \delta_n), \quad \delta_i \in \mathbb{C}.$

(b) $\underline{\Delta} = diag(\delta_1, \ldots, \delta_n), \quad \delta_i \in \mathbb{R}.$

To simplify the computation, minimize the Frobenius norm of Δ in the defintion of $\mu_{\underline{\Delta}}(A)$.

Chapter 5

Matrix Perturbations

5.1 Introduction

The following question arises frequently in matrix theory: What is the smallest possible perturbation of a matrix that causes it to lose rank? We discuss two cases next, with perturbations measured in the 2-norm, and then discuss the measurement of perturbations in the Frobenius norm. This provides us with a new formulation to the least squares estimation problem in which uncertainty is present in the matrix A as well as the vector y. This is known as *total least squares*.

5.2 Additive Perturbation

Theorem 5.1 Suppose $A \in \mathbb{C}^{m \times n}$ has full column rank $(= n)$. Then

$$\min_{\Delta \in \mathbb{C}^{m \times n}} \{\|\Delta\|_2 \mid A + \Delta \text{ has rank } < n\} = \sigma_n(A) \ . \tag{5.1}$$

Proof: Suppose $A + \Delta$ has rank $< n$. Then there exists $x \neq 0$ such that $\|x\|_2 = 1$ and

$$(A + \Delta)\, x = 0 \ .$$

Since $\Delta x = -Ax$,

$$\begin{aligned} \|\Delta x\|_2 &= \|Ax\|_2 \\ &\geq \sigma_n(A) \ . \end{aligned} \tag{5.2}$$

From the properties of induced norms (see Section 3.1), we also know that

$$\|\Delta\|_2 \|x\|_2 \geq \|\Delta x\|_2.$$

Using Equation (24.3) and the fact that $\|x\|_2 = 1$, we arrive at the following:

$$\begin{aligned} \|\Delta\|_2 &\geq \|\Delta x\|_2 \\ &\geq \sigma_n(A) \end{aligned} \tag{5.3}$$

To complete the proof, we must show that the lower bound from Equation (5.3) can be achieved. Thus, we must construct a Δ so that $A + \Delta$ has rank $< n$ and $\|\Delta\|_2 = \sigma_n(A)$; such a Δ will be a minimizing solution. For this, choose

$$\Delta = -\sigma_n u_n v_n'$$

where u_n, v_n are the left and right singular vectors associated with the smallest singular value σ_n of A. Notice that $\|\Delta\|_2 = \sigma_n(A)$. This choice yields

$$\begin{aligned} (A + \Delta)\, v_n &= \sigma_n u_n - \sigma_n u_n v_n^* v_n \\ &= \sigma_n u_n - \sigma_n u_n \\ &= 0 \ . \end{aligned}$$

That is, $A + \Delta$ has rank $< n$. This completes the proof.

5.3 Multiplicative Perturbation

Theorem 5.2 (Small Gain) Given $A \in \mathbb{C}^{m \times n}$,

$$\min_{\Delta \in \mathbb{C}^{n \times m}} \{ \|\Delta\|_2 \mid I - A\Delta \text{ is singular } \} = \frac{1}{\sigma_1(A)} \ . \tag{5.4}$$

Proof: Suppose $I - A\Delta$ is singular. Then there exists $x \neq 0$ such that

$$(I - A\Delta)\, x = 0$$

so

$$\|A\Delta x\|_2 = \|x\|_2 \ . \tag{5.5}$$

From the properties of induced norms (see Lecture 4 notes),

$$\begin{aligned} \|A\Delta x\|_2 &\leq \|A\|_2 \|\Delta x\|_2 \\ &= \sigma_1(A) \|\Delta x\|_2 \ . \end{aligned}$$

Upon substituting the result in Equation (5.5) for $\|A\Delta x\|_2$, we find

$$\|x\|_2 \leq \sigma_1(A) \|\Delta x\|_2 \ .$$

Dividing through by $\sigma_1(A)\|x\|_2$ yields

$$\frac{\|\Delta x\|_2}{\|x\|_2} \geq \frac{1}{\sigma_1(A)} \ ,$$

which implies

$$\|\Delta\|_2 \geq \frac{1}{\sigma_1(A)} \ . \tag{5.6}$$

To conclude the proof, we must show that this lower bound can be achieved. Thus, we construct a Δ which satisfies Equation (5.6) with equality and also causes $(I - A\Delta)$ to be singular. For this, choose

$$\Delta = \frac{1}{\sigma_1(A)} v_1 u_1' \ .$$

Notice that the lower bound (Equation (5.6)) is satisfied with equality, *i.e.*, $\|\Delta\|_2 = 1/\sigma_1(A)$. Now choose $x = u_1$. Then:

$$\begin{aligned}
(I - A\Delta)\, x &= (I - A\Delta)\, u_1 \\
&= \left(I - \frac{A v_1 u_1'}{\sigma_1}\right) u_1 \\
&= u_1 - \underbrace{\frac{A v_1}{\sigma_1}}_{u_1} \\
&= u_1 - u_1 \quad (\text{since } A v_1 = \sigma_1 u_1) \\
&= 0 \ .
\end{aligned}$$

This completes the proof.

The theorem just proved is called the **small gain theorem**. The reason for this is that it guarantees $(I - A\Delta)$ is nonsingular provided

$$\|\Delta\|_2 < \frac{1}{\|A\|_2} \ .$$

This condition is most often written as

$$\|\Delta\|_2 \|A\|_2 < 1 \ , \tag{5.7}$$

i.e., the product of the gains is less than one.

Remark: We can actually obtain the additive perturbation result from multiplicative perturbation methods. Assume A is invertible, and Δ is a matrix which makes its sum with A singular. Since

$$A + \Delta = A\left(I + A^{-1}\Delta\right) \ ,$$

and A is nonsingular, then $(I + A^{-1} \quad)$ must be singular. By our work with multiplicative perturbations, we know that the $\quad$ associated with the smallest $\| \quad \|_2$ that makes this quantity singular satisfies

$$\| \quad \|_2 = \frac{1}{\sigma_1(A^{-1})} = \sigma_n(A) \quad .$$

5.4 Perturbations Measured in the Frobenius Norm

We will now demonstrate that, for the multiplicative and additive perturbation cases where we minimized the induced 2-norm, we also minimized the Frobenius norm.

Let $A \in \mathbb{C}^{m \times n}$, and let $\text{rank}(A) = r$.

$$\begin{aligned}
\|A\|_F &\stackrel{\triangle}{=} \left(\sum_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}|^2 \right)^{\frac{1}{2}} && (5.8)\\
&= \left(\text{trace}(A'A) \right)^{\frac{1}{2}} && (5.9)\\
&= \left(\sum_{i=1}^{r} \sigma_i^2 \right)^{\frac{1}{2}} \quad \text{(the trace of a matrix is the sum of its eigenvalues)} && (5.10)\\
&\quad \sigma_1(A) \quad . && (5.11)
\end{aligned}$$

Therefore,

$$\|A\|_F \qquad \|A\|_2 \tag{5.12}$$

which is a useful inequality.

In both the perturbation problems that we considered earlier, we found a rank-one solution, or dyad, for Δ:

$$= \alpha u v' \tag{5.13}$$

where $\alpha \in \mathbb{C}$, $u \in \mathbb{C}^m$, $v \in \mathbb{C}^n$ such that $\|u\|_2 = \|v\|_2 = 1$. It is easy to show that the Frobenius norm and induced 2-norm are *equal* for rank one matrices of the form in Equation (5.13). It follows from this that the $\quad$ which minimizes the induced 2-norm also minimizes the Frobenius norm, for the additive and multiplicative perturbation cases we have examined. In general, however, minimizing the induced 2-norm of a matrix does not imply the Frobenius norm is minimized (or vice versa.)

Example 5.1 This example is intended to illustrate the use of the singular value decomposition and Frobenius norms in the solution of a minimum distance problem. Suppose we have a matrix $A \in \mathbb{C}^{n \times n}$, and we are interested in finding the closest matrix to A of the form cW where c is a complex number and W is a

unitary matrix. The distance is to be measured by the Frobenius norm. This problem can be formulated as

$$\min_{c\in\mathbb{C},W\in\mathbb{C}^{n\times n}} \|A - cW\|_F$$

where $W'W = I$. We can write

$$\begin{aligned}\|A - cW\|_F^2 &= \mathrm{Tr}\left((A - cW)'(A - cW)\right) \\ &= \mathrm{Tr}(A'A) - c'\mathrm{Tr}(W'A) - c\mathrm{Tr}(A'W) + |c|^2\mathrm{Tr}(W'W).\end{aligned}$$

Note that $\mathrm{Tr}(W'W) = \mathrm{Tr}(I) = n$. Therefore, we have

$$\|A - cW\|_F^2 = \|A\|_F^2 - 2Re\left(c'\mathrm{Tr}(W'A)\right) + n|c|^2 \tag{5.14}$$

and by taking

$$c = \frac{1}{n}\mathrm{Tr}(W'A)$$

the right hand side of Equation (5.14) will be minimized. Therefore we have that

$$\|A - cW\|_F^2 \quad \|A\|_F^2 - \frac{1}{n}|\mathrm{Tr}(W'A)|^2.$$

Now we must minimize the right hand side with respect to W, which is equivalent to maximizing $|\mathrm{Tr}(W'A)|$. In order to achieve this we employ the singular value decomposition of A as $U\Sigma V'$, which gives

$$\begin{aligned}|\mathrm{Tr}(W'A)|^2 &= |\mathrm{Tr}(W'U\Sigma V')|^2 \\ &= |\mathrm{Tr}(V'W'U\Sigma)|^2.\end{aligned}$$

The matrix $Z = V'W'U$ satisfies

$$\begin{aligned}ZZ' &= V'W'UU'WV \\ &= I.\end{aligned}$$

Therefore,

$$|\mathrm{Tr}(Z\Sigma)|^2 = |\sum_{i=1}^{n}\sigma_i z_{ii}|^2 \leq \left(\sum_{i=1}^{n}\sigma_i\right)^2$$

implies that

$$\min_{c,W}\|A - cW\|_F^2 \quad \|A\|_F^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\sigma_i\right)^2. \tag{5.15}$$

In order to complete this example we show that the lower bound in Equation (5.15) can actually be achieved with a specific choice of W. Observe that

$$\mathrm{Tr}(W'U\Sigma V') = \mathrm{Tr}(W'UV'\Sigma)$$

and by letting $W' = VU'$ we obtain

$$\mathrm{Tr}(W'A) = \mathrm{Tr}(\Sigma) = \sum_{i=1}^{n} \sigma_i$$

and

$$c = \frac{1}{n} \sum_{i=1}^{n} \sigma_i.$$

Putting all the pieces together, we get that

$$\min_{c,W} \|A - cW\|_F^2 = \sum_{i=1}^{n} \sigma_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} \sigma_i^2 \right)^2$$

and the minimizing unitary matrix is given by

$$cW = \frac{1}{n} \left(\sum_{i=1}^{n} \sigma_i \right) UV'.$$

It is clear also that, in order for a matrix to be exactly represented as a complex multiple of a unitary matrix, all of its singular values must be equal.

5.5 Total Least Squares

We have previously examined solving least squares problems of the form $y = Ax + e$. An interpretation of the problem we solved there is that we perturbed y as little as possible — in the least squares sense — to make the resulting equation $y - e = Ax$ consistent. It is natural to ask what happens if we allow A to be perturbed as well, in addition to perturbing y. This makes sense in situations where the uncertainty in our model and the noise in our measurements cannot or should not be attributed entirely to y, but also to A. The simplest least squares problem of this type is one that allows a perturbed model of the form

$$y = (A + \Delta)\, x + e \quad . \tag{5.16}$$

The so-called *total least squares* estimation problem can now be stated as

$$\min_{\Delta,e} \left(\sum_{i,j} |\quad_{ij}|^2 + \sum_i |e_i|^2 \right)^{\frac{1}{2}} = \min_{\Delta,e} \| \quad \vdots\, e \|_F \tag{5.17}$$

$$= \min_{\Delta,e} \| \hat{\ } \|_F \tag{5.18}$$

where

$$\hat{\ } = \left[\quad \vdots\, e \right] \quad . \tag{5.19}$$

Weighted versions of this problem can also be posed, but we omit these generalizations.

Note that no constraints have been imposed on in the above problem statement, and this can often limit the direct usefulness of the total least squares formulation in practical problems. In practice, the expected or allowed perturbations of A are often quite structured; however, the solution of the total least squares problem under such structural constraints is much harder than that of the unconstrained problem that we present the solution of next. Nevertheless, the total least squares formulation can provide a useful benchmark. (The same sorts of comments can of course be made about the conventional least squares formulation: it is often not the criterion that we would want to use, but its tractability compared to other criteria makes it a useful point of departure.)

If we make the definitions

$$\hat{A} = \left[\begin{array}{c|c} A & -y \end{array} \right] \qquad \hat{x} = \left[\begin{array}{c} x \\ 1 \end{array} \right] \tag{5.20}$$

then the perturbed model in Equation (5.16) can be rewritten as

$$\hat{A} + \hat{} \quad \hat{x} = 0 \ . \tag{5.21}$$

This equation makes evident that what we seek is the $\hat{}$ with minimal Frobenius norm that satisfies Equation (5.21)—the smallest $\hat{}$ that makes $\hat{A} + \hat{}$ singular.

Let us suppose that A has full column rank (n), and that it has more rows than columns (which is normally the case, since in least squares estimation we typically have many more measurements than parameters to estimate). In addition, let us assume that $\hat{A}$ has rank $(n+1)$, which is also generally true. From what we've learned about additive perturbations, we now see that a minimal (in a Frobenius sense) $\hat{}$ that satisfies Equation (5.21) is

$$\hat{} = -\sigma_{n+1} u_{n+1} v'_{n+1} \tag{5.22}$$

where the σ_{n+1}, u_{n+1} and v_{n+1} are derived from the SVD of $\hat{A}$ (i.e. σ_{n+1} is the smallest singular value of $\hat{A}$, etc.). Given that we now know $\hat{A}$ and $\hat{\Delta}$, choosing $\hat{x} = v_{n+1}$, and rescaling $\hat{x}$, we have

$$\hat{A} + \hat{} \quad \left[\begin{array}{c} x \\ 1 \end{array} \right] = 0$$

which gives us x, the total least squares solution. This solution is due to Golub and Van Loan (see their classic text on *Matrix Computations*, Second Edition, Johns Hopkins University Press, 1989).

5.6 Conditioning of Matrix Inversion

We are now in a position to address some of the issues that came up in Example 1 of Lecture 4, regarding the sensitivity of the inverse A^{-1} and of the solution $x = A^{-1}b$ to perturbations

in A (and/or b, for that matter). We first consider the case where A is invertible, and examine the sensitivity of A^{-1}. Taking differentials in the defining equation $A^{-1}A = I$, we find

$$d(A^{-1})\,A + A^{-1}\,dA = 0 \tag{5.23}$$

where the order of the terms in each half of the sum is important, of course. (Rather than working with differentials, we could equivalently work with perturbations of the form $A + \epsilon P$, etc., where ϵ is vanishingly small, but this really amounts to the same thing.) Rearranging the preceding expression, we find

$$d(A^{-1}) = -A^{-1}\;dA\;A^{-1} \tag{5.24}$$

Taking norms, the result is

$$\|d(A^{-1})\| \leq \|A^{-1}\|^2\|dA\| \tag{5.25}$$

or equivalently

$$\frac{\|d(A^{-1})\|}{\|A^{-1}\|} \leq \|A\|\|A^{-1}\|\frac{\|dA\|}{\|A\|} \tag{5.26}$$

This derivation holds for any submultiplicative norm. The product $\|A\|\|A^{-1}\|$ is termed the *condition number* of A with respect to inversion (or simply the condition number of A) and denoted by $K(A)$:

$$K(A) = \|A\|\|A^{-1}\| \tag{5.27}$$

When we wish to specify which norm is being used, a subscript is attached to $K(A)$. Our earlier results on the SVD show, for example, that

$$K_2(A) = \sigma_{max}/\sigma_{min} \tag{5.28}$$

The condition number in this 2-norm tells us how slender the ellipsoid Ax for $\|x\|_2 = 1$ is — see Figure 5.1. In what follows, we shall focus on the 2-norm condition number (but will omit the subscript unless essential).

Some properties of the 2-norm condition number (all of which are easy to show, and some of which extend to the condition number in other norms) are

- $K(A) \quad 1$;
- $K(A) = K(A^{-1})$;
- $K(AB) \leq K(A)K(B)$;
- Given $U'U = I$, $K(UA) = K(A)$.

The importance of (5.26) is that the bound can actually be attained for some choice of the perturbation dA and of the matrix norm, so the situation can get as bad as the bound allows: *the fractional change in the inverse can be $K(A)$ times as large as the fractional change in the original.* In the case of 2-norms, a particular perturbation that attains the bound

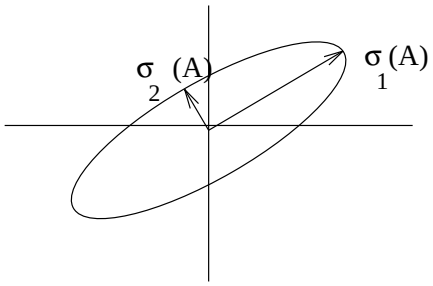


Figure 5.1: Depiction of how A (a real 2×2 matrix) maps the unit circle. The major axis of the ellipse corresponds to the largest singular value, the minor axis to the smallest.

can be derived from the of Theorem 5.1, by simply replacing $-\sigma_n$ in by a differential perturbation:

$$dA = -d\sigma \, u_n v_n' \tag{5.29}$$

We have established that a large condition number corresponds to a matrix whose inverse is very sensitive to relatively small perturbations in the matrix. Such a matrix is termed *ill conditioned* or poorly conditioned with respect to inversion. A perfectly conditioned matrix is one whose condition number takes the minimum possible value, namely 1.

A high condition number also indicates that a matrix is close to losing rank, in the following sense: There is a perturbation of small norm $(= \sigma_{min})$ relative to $\|A\|$ $(= \sigma_{max})$ such that $A +$ has lower rank than A. This follows from our additive perturbation result in Theorem 5.1. This interpretation extends to non-square matrices as well. We shall term the ratio in (5.28) the condition number of A even when A is non-square, and think of it as a measure of *nearness to a rank loss*.

Turning now to the sensitivity of the solution $x = A^{-1}b$ of a linear system of equations in the form $Ax = b$, we can proceed similarly. Taking differentials, we find that

$$dx = -A^{-1} \, dA \; A^{-1}b + A^{-1} \, db = -A^{-1} \, dA \; x + A^{-1}b \tag{5.30}$$

Taking norms then yields

$$\|dx\| \leq \|A^{-1}\| \|dA\| \|x\| + \|A^{-1}\| \|db\| \tag{5.31}$$

Dividing both sides of this by $\|x\|$, and using the fact that $\|x\|$ $(\|b\|/\|A\|)$, we get

$$\frac{\|dx\|}{\|x\|} \leq K(A) \; \frac{\|dA\|}{\|A\|} + \frac{\|db\|}{\|b\|} \tag{5.32}$$

We can come close to attaining this bound if, for example, b happens to be nearly collinear with the column of U in the SVD of A that is associated with σ_{min}, and if appropriate perturbations occur. Once again, therefore, the fractional change in the answer can be close to $K(A)$ times as large as the fractional changes in the given matrices.

Example 5.2 For the matrix A given in Example 1 of Lecture 4, the SVD is

$$A = \begin{matrix} 100 & 100 \\ 100.2 & 100 \end{matrix} = \begin{matrix} .7068 & .7075 \\ .7075 & .-7068 \end{matrix} \begin{matrix} 200.1 & 0 \\ 0 & 0.1 \end{matrix} \begin{matrix} .7075 & .7068 \\ -.7068 & .7075 \end{matrix} \tag{5.33}$$

The condition number of A is seen to be 2001, which accounts for the 1000-fold magnification of error in the inverse for the perturbation we used in that example. The perturbation of smallest 2-norm that causes $A +$ to become singular is

$$= \begin{matrix} .7068 & .7075 \\ .7075 & .-7068 \end{matrix} \begin{matrix} 0 & 0 \\ 0 & -0.1 \end{matrix} \begin{matrix} .7075 & .7068 \\ -.7068 & .7075 \end{matrix}$$

whose norm is 0.1. Carrying out the multiplication gives

$$\approx \begin{matrix} .05 & -.05 \\ -.05 & .05 \end{matrix}$$

With $b = [1 \ -1]^T$, we saw large sensitivity of the solution x to perturbations in A. Note that this b is indeed nearly collinear with the second column of U. If, on the other hand, we had $b = [1 \ 1]$, which is more closely aligned with the first column of U, then the solution would have been hardly affected by the perturbation in A — a claim that we leave you to verify.

Thus $K(A)$ serves as a bound on the magnification factor that relates fractional changes in A or b to fractional changes in our solution x.

Conditioning of Least Squares Estimation

Our objective in the least-square-error estimation problem was to find the value $\widehat{x}$ of x that minimizes $\|y - Ax\|_2^2$, under the assumption that A has full column rank. A detailed analysis of the conditioning of this case is beyond our scope (see *Matrix Computations* by Golub and Van Loan, cited above, for a detailed treatment). We shall make do here with a statement of the main result in the case that the fractional residual is much less than 1, i.e.

$$\frac{\|y - A\widehat{x}\|_2}{\|y\|_2} \ll 1 \tag{5.34}$$

This low-residual case is certainly of interest in practice, assuming that one is fitting a reasonably good model to the data. In this case, it can be shown that the fractional change $\|d\widehat{x}\|_2/\|\widehat{x}\|_2$ in the solution $\widehat{x}$ can approach $K(A)$ times the sum of the fractional changes in A and y, where $K(A) = \sigma_{max}(A)/\sigma_{min}(A)$. In the light of our earlier results for the case of invertible A, this result is perhaps not surprising.

Given this result, it is easy to explain why solving the normal equations

$$(A'A)\widehat{x} = A'y$$

to determine $\widehat{x}$ is numerically unattractive (in the low-residual case). The numerical inversion of $A'A$ is governed by the condition number of $A'A$, and this is the *square* of the condition number of A:

$$K(A'A) = K^2(A)$$

You should confirm this using the SVD of A. The process of directly solving the normal equations will thus introduce errors that are not intrinsic to the least-square-error problem, because this problem is governed by the condition number $K(A)$, according to the result quoted above. Fortunately, there are other algorithms for computing $\widehat{x}$ that are governed by the condition number $K(A)$ rather than the square of this (and Matlab uses one such algorithm to compute $\widehat{x}$ when you invoke its least squares solution command).

Exercises

Exercise 5.1 Suppose the complex $m \times n$ matrix A is perturbed to the matrix $A + E$.

(a) Show that

$$| \sigma_{max}(A + E) - \sigma_{max}(A) | \leq \sigma_{max}(E)$$

Also find an E that results in the inequality being achieved with equality.

(Hint: To show the inequality, write $(A + E) = A + E$ and $A = (A + E) - E$, take the 2-norm on both sides of each equation, and use the triangle inequality.)

It turns out that the result in (a) actually applies to *all* the singular values of A and $A + E$, not just the largest one. Part (b) below is one version of the result for the smallest singular value.

(b) Suppose A has *less than* full column rank, i.e. has rank$< n$, but $A + E$ has full column rank. Show (following a procedure similar to part (a) — but looking at $\min \|(A + E)x\|_2$ rather than the norm of $A + E$, etc.) that

$$\sigma_{min}(A + E) \leq \sigma_{max}(E)$$

Again find an E that results in the inequality being achieved with equality.

[The result in (b), and some extensions of it, give rise to the following sound (and widely used) procedure for estimating the rank of some underlying matrix A, given only the matrix $A + E$ and knowledge of $\|E\|_2$: Compute the SVD of $A + E$, then declare the "numerical rank" of A to be the number of singular values of $A + E$ that are larger than the threshold $\|E\|_2$. The given information is consistent with having an A of this rank.]

(c) Verify the above results using your own examples in MATLAB. You might also find it interesting to verify numerically that for large m, n, the norm of the matrix $E = s * randn(m, n)$ — which is a matrix whose entries are independent, zero-mean, Gaussian, with standard deviation s — is close to $s * (\sqrt{m} + \sqrt{n})$. So if A is perturbed by such a matrix, then a reasonable value to use as a threshold when determining the numerical rank of A is this number.

Exercise 5.2 Let A and E be $m \times n$ matrices. Show that

$$\min_{\text{rank } E \leq r} \|A - E\|_2 = \sigma_{r+1}(A).$$

To prove this, notice that the rank constraint on E can be interpreted as follows: If $v_1, \ldots, v_{r+1}$ are linearly independent vectors, then there exists a nonzero vector z, expressed as a linear combination of such vectors, that belongs to the nullspace of E. Proceed as follows:

1. Select the v_i's from the SVD of A.
2. Select a candidate element z with $\|z\|_2 = 1$.
3. Show that $\|(A - E)z\|_2 \geq \sigma_{r+1}$. This implies that $\|A - E\|_2 \geq \sigma_{r+1}$.
4. Construct an E that achieves the above bound.

Exercise 5.3 Consider the real, square system of equations $Ax = (U\Sigma V^T)x = y$, where U and V are orthogonal matrices, with

$$\Sigma = \begin{matrix} 1 & 0 \\ 0 & 10^{-6} \end{matrix} \quad , \quad y = U \begin{matrix} 1 \\ 10^{-6} \end{matrix}$$

All norms in this problem are taken to be 2-norms.

(a) What is the norm of the exact solution x?

(b) Suppose y is perturbed to $y + \delta y$, and that correspondingly the solution changes from x in (a) to $x + \delta x$. Find a perturbation δy, with $\|\delta y\| = 10^{-6}$, such that

$$\frac{\|\delta x\|}{\|x\|} \approx \kappa(A) \frac{\|\delta y\|}{\|y\|}$$

where $\kappa(A)$ is the condition number of A.

(c) Suppose instead of perturbing y we perturb A, changing it to $A + \delta A$, with the solution correspondingly changing from x to $x + \delta x$ (for some δx that is different than in part (b)). Find a perturbation δA, with $\|\delta A\| = 10^{-7}$, such that

$$\frac{\|\delta x\|}{\|x\|} \approx \kappa(A) \frac{\|\delta A\|}{\|A\|}$$

Exercise 5.4 Positive Definite Matrices

A matrix A is positive semi-definite if $x'Ax \geq 0$ for all $x \neq 0$. We say Y is the square root of a Hermitian positive semi-definite matrix if $Y'Y = A$. Show that Y always exists and can be constructed from the SVD of A.

Exercise 5.5 Let A and B have compatible dimensions. Show that if

$$\|Ax\|_2 \leq \|Bx\|_2 \qquad \text{for all } x,$$

then there exists a matrix Y with $\|Y\|_2 \leq 1$ such that

$$A = YB.$$

Assume B has full rank to simplicity.

Exercise 5.6 **(a)** Suppose

$$\left\| \begin{matrix} X \\ A \end{matrix} \right\| \leq \gamma.$$

Show that there exists a matrix Y with $\|Y\|_2 \leq 1$ such that

$$X = Y(\gamma^2 I - A'A)^{\frac{1}{2}}$$

(b) Suppose

$$\|\begin{pmatrix} X & A \end{pmatrix}\| \leq \gamma.$$

Show that there exists a matrix Z with $\|Z\| \leq 1$ such that $X = (\gamma^2 I - AA^*)^{\frac{1}{2}} Z$.

Exercise 5.7 Matrix Dilation

The problems above can help us prove the following important result:

$$\gamma_0 := \min_X \left\| \begin{matrix} X & B \\ C & A \end{matrix} \right\| = \max \left\{ \|\begin{pmatrix} C & A \end{pmatrix}\|, \left\| \begin{matrix} B \\ A \end{matrix} \right\| \right\}.$$

This is known as the matrix dilation theorem. Notice that the left hand side is always greater than or equal to the right hand side irrespective of the choice of X. Below, we outline the steps necessary to prove that this lower bound is tight. Matrix dilations play an important role in systems theory particularly in model reduction problems.

1. Let γ_1 be defined as

$$\gamma_1 = \max \left\{ \|\begin{pmatrix} C & A \end{pmatrix}\|, \left\| \begin{matrix} B \\ A \end{matrix} \right\| \right\}.$$

Show that:

$$\gamma_0 \geq \gamma_1.$$

2. Use the previous exercise to show that there exists two matrices Y and Z with norms less than or equal to one such that

$$B = Y(\gamma_1^2 I - A^* A)^{\frac{1}{2}}, \quad C = (\gamma_1^2 I - AA^*)^{\frac{1}{2}} Z.$$

3. Define a candidate solution to be $\tilde{X} = -YA^*Z$. Show by direct substitution that

$$\begin{aligned} \left\| \begin{matrix} \tilde{X} & B \\ C & A \end{matrix} \right\| &= \left\| \begin{matrix} -YA^*Z & Y(\gamma_1^2 I - A^*A)^{\frac{1}{2}} \\ C = (\gamma_1^2 I - AA^*)^{\frac{1}{2}} Z & A \end{matrix} \right\| \\ &= \left\| \begin{matrix} Y & 0 & -A^* & (\gamma_1^2 I - A^*A)^{\frac{1}{2}} & Z & 0 \\ 0 & I & C = (\gamma_1^2 I - AA^*)^{\frac{1}{2}} & A & 0 & I \end{matrix} \right\| \end{aligned}$$

4. Show that

$$\left\| \begin{matrix} \tilde{X} & B \\ C & A \end{matrix} \right\| \leq \gamma_1.$$

This implies that $\gamma_0 \leq \gamma_1$ which proves the assertion.

Exercise 5.8 Prove or disprove (through a counter example) the following singular values inequalities.

1. $\sigma_{min}(A + B) \leq \sigma_{min}(A) + \sigma_{min}(B)$ for any A and B.
2. $\sigma_{min}(A + E) \leq \sigma_{max}(E)$ whenever A does not have column rank, and E is any matrix.

3. If $\sigma_{max}(A) < 1$, then

$$\sigma_{max}(I - A)^{-1} \leq \frac{1}{1 - \sigma_{max}(A)}$$

4. $\sigma_i(I + A) \leq \sigma_i(A) + 1$.

Chapter 6

Dynamic Models

6.1 Introduction: Signals, Systems and Models

A **system** may be thought of as something that imposes *constraints* on — or enforces relationships among — a set of variables. This "system as constraints" point of view is very general and powerful. Rather more restricted, but still very useful and common, is the view of a system as a *mapping* from a set of *input* variables to a set of *output* variables; a mapping is evidently a very particular form of constraint.

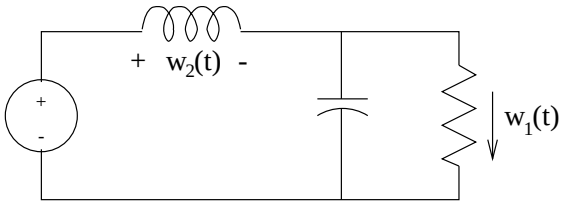
A **(behavioral) model** lists the variables of interest (the *"manifest"* variables) and the constraints that they must satisfy. Any combination of variables that satisfies the constraints is possible or allowed, and is termed *a behavior* of the model.

To facilitate the specification of the constraints, one may introduce auxiliary (*"latent"*) variables. One might then distinguish among the manifest behavior, latent behavior, and *full* behavior (manifest as well as latent).

For a **dynamic model**, the "variables" referred to above are actually *signals* that evolve as a function of time (and/or a function of other independent variables, e.g. space). We first need to specify a *time axis* $\mathbb{T}$ (discrete, continuous, infinite, semi-infinite ...) and a *signal space* $\mathbb{W}$, *i.e.* the space of values the signals live in at each time instant. A dynamic model for a set of signals $\{w_i(t)\}$ is then completed by listing the constraints that the $w_i(t)$ must satisfy. Any combination $w(t) = [\ w_1(t),\ \cdots,\ w_\ell(t)\ ]$ of signals that satisfies the constraints is *a behavior* of the model, $w(t) \in \mathbb{B}$, where $\mathbb{B}$ denotes *the behavior.*

We now present some examples of dynamic models, to highlight various possible model representations.

Example 6.1 (Circuit)



Suppose the signals (variables) of interest — the manifest signals — in the above circuit diagram are $w_1(t)$, $w_2(t)$ and $w_3(t)$ for $t \geq 0$, so the signal space $\mathbb{W}$ is $\mathbb{R}^3$ and the time axis $\mathbb{T}$ is $\mathbb{R}^+$ (i.e. the interval $[0, \infty]$). Picking all other component voltages and currents as latent signals, we can write the constraints that define the model as:

$$\begin{cases} \text{2 Kirchhoff's voltage law (KVL) equations} \\ \text{2 Kirchhoff's current law (KCL) equations} \\ \text{4 defining equations for the components} \end{cases}$$

Any set of manifest and latent signals that simultaneously satisfies (or solves) the preceding constraint equations constitutes *a* behavior, and *the* behavior $\mathbb{B}$ of the model is the space of all such solutions.

The same behavior may equivalently be described by a model written entirely in terms of the manifest variables, by eliminating all the other variables in the above equations to obtain

$$0 = \frac{w_1}{R} + C\dot{w}_1 - w_2 \tag{6.1}$$

$$0 = -w_3 + L\dot{w}_2 + w_1 \tag{6.2}$$

Still further reduction to a single second-order differential equation is possible, by taking the derivative of one of these equations and eliminating one variable.

Example 6.2 (Mass-Spring System)

An object of mass M moves on a horizontal frictionless slide, and is attached to one end of it by a linear spring with spring constant k. A horizontal force $u(t)$ is applied to the mass. Assume that the variable z measures the change in the spring length from its natural length. From Newton's law we obtain the model

$$M\ddot{z} = -kz + u.$$

Example 6.3 (Inverted Pendulum)

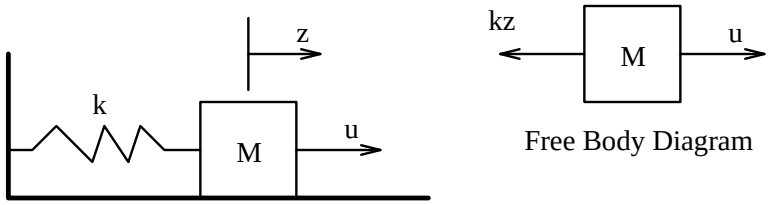


Figure 6.1: Mass Spring System.

A cart of mass M slides on a horizontal frictionless track, and is pulled by a horizontal force $u(t)$. On the cart an inverted pendulum of mass m is attached via a frictionless hinge, as shown in Figure 28.1. The pendulum's center of mass is located at a distance l from its two ends, and the pendulum's moment of inertia about its center of mass is denoted by I. The point of support of the pendulum is a distance $s(t)$ from some reference point. The angle $\theta(t)$ is the angle that the pendulum makes with respect to the vertical axis. The vertical force exerted by the cart on the base of the pendulum is denoted by P, and the horizontal force by N. What we wish to model are the constraints governing the (manifest) signals $u(t)$, $s(t)$ and $\theta(t)$.

First let us write the equations of motion that result from the free-body diagram of the cart. The vertical forces P, R and Mg balance out. For the horizontal forces we have the following equation:

$$M\ddot{s} = u - N. \tag{6.3}$$

From the free-body diagram of the pendulum, the balance of forces in the horizontal direction gives the equation

$$\begin{aligned} m\frac{d^2}{dt^2}\left(s + l\sin(\theta)\right) &= N, \qquad \text{or} \\ m\left(\ddot{s} - l\sin(\theta)(\dot{\theta})^2 + l\cos(\theta)\ddot{\theta}\right) &= N, \end{aligned} \tag{6.4}$$

and the balance of forces in the vertical direction gives the equation

$$\begin{aligned} m\frac{d^2}{dt^2}\left(l\cos(\theta)\right) &= P - mg, \qquad \text{or} \\ m\left(-l\cos(\theta)(\dot{\theta})^2 - l\sin(\theta)\ddot{\theta}\right) &= P - mg. \end{aligned} \tag{6.5}$$

From equations (28.16) and (28.17) we can eliminate the force N to obtain

$$(M+m)\ddot{s} + m\left(l\cos(\theta)\ddot{\theta} - l\sin(\theta)(\dot{\theta})^2\right) = u. \tag{6.6}$$

By balancing the moments around the center of mass, we get the equation

$$I\ddot{\theta} = Pl\sin(\theta) - Nl\cos(\theta). \tag{6.7}$$

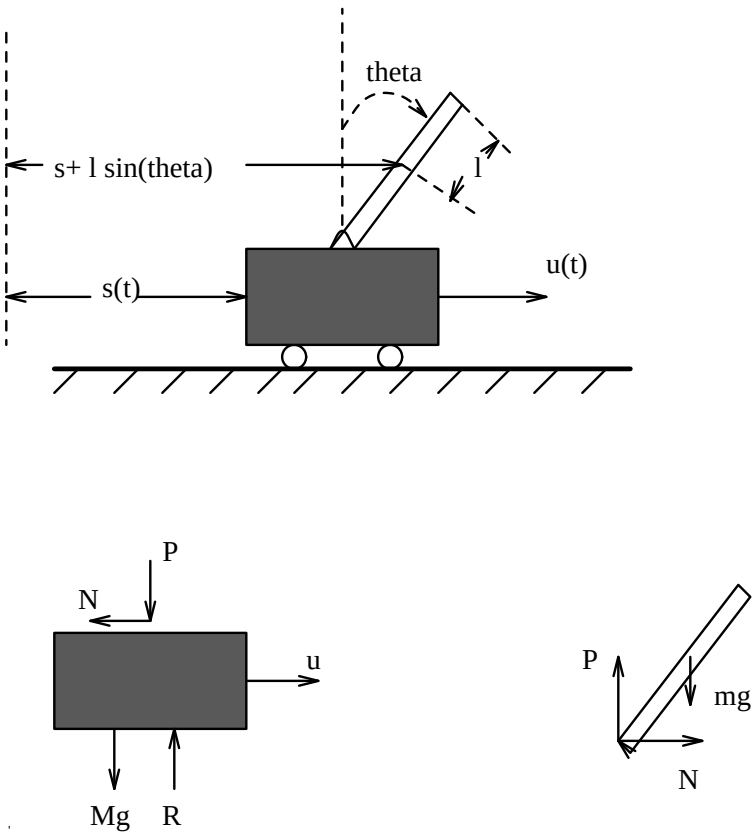


Figure 6.2: Inverted Pendulum

Substituting (28.17) and (28.18) into (28.19) gives us

$$\begin{aligned} I\ddot{\theta} &= l\left(mg - ml\cos(\theta)(\dot{\theta})^2 - ml\sin(\theta)\ddot{\theta}\right)\sin(\theta) \\ &- l\left(m\ddot{s} - ml\sin(\theta)(\dot{\theta})^2 + ml\cos(\theta)\ddot{\theta}\right)\cos(\theta). \end{aligned}$$

Simplifying the above expression gives us the equation

$$(I + ml^2)\ddot{\theta} = mgl\sin(\theta) - ml\ddot{s}\cos(\theta). \tag{6.8}$$

The equations that comprise our model for the system are (28.20) and (28.21).

We can have a further simplification of the system of equations by removing the term $\ddot{\theta}$ from equation (28.20), and the term $\ddot{s}$ from equation (28.21). Define the constants

$$\begin{aligned} \mathcal{M} &= M + m \\ L &= \frac{I + ml^2}{ml}. \end{aligned}$$

Substituting $\ddot{\theta}$ from (28.21) into (28.20), we get

$$\left(1 - \frac{ml}{\mathcal{M}L}\cos(\theta)^2\right)\ddot{s} + \frac{ml}{\mathcal{M}L}g\sin(\theta)\cos(\theta) - \frac{ml}{\mathcal{M}}\sin(\theta)(\dot{\theta})^2 = \frac{1}{\mathcal{M}}u. \tag{6.9}$$

Similarly we can substitute $\ddot{s}$ from (28.20) into (28.21) to get

$$\left(1 - \frac{ml}{\mathcal{M}L}\cos(\theta)^2\right)\ddot{\theta} - \frac{g}{L}\sin(\theta) + \frac{ml}{\mathcal{M}L}\sin(\theta)\cos(\theta)(\dot{\theta})^2 = -\frac{1}{\mathcal{M}L}\cos(\theta)u. \tag{6.10}$$

Example 6.4 (Predator-Prey Model)

While the previous examples are physically based, there are many examples of dynamic models that are hypothesized on the basis of a behavioral pattern. For a classical illustration, consider an island populated primarily by goats and foxes. Goats survive on the island's vegetation while foxes survive by eating goats.

To build a model of the population growth of these two interacting animals, define:

$$N_1(t) = \text{number of goats at time t} \tag{6.11}$$
$$N_2(t) = \text{number of foxes at time t} \tag{6.12}$$

where t refers to (discrete) time measured in multiples of months. Volterra proposed the following model:

$$N_1(t+1) = aN_1(t) - bN_1(t)N_2(t) \tag{6.13}$$
$$N_2(t+1) = cN_2(t) + dN_1(t)N_2(t) \tag{6.14}$$

The constants a, b, c, and d are all positive, with $a > 1$, $c < 1$. If there were no goats on the island, $N_1(0) = 0$, then — according to this model — the foxes' population would decrease geometrically (i.e. as a discrete-time exponential). If there were no foxes on the island, then the goat population would grow geometrically (presumably there is an unlimited supply of vegetation, water and space). On the other hand, if both species existed on the island, then the frequency of their encounters, which is modeled as being proportional to the product N_1N_2, determines at what rate goats are eaten and foxes are well-fed. Among the questions that might now be asked are: What sorts of qualitative behavioral characteristics are associated with such a model, and what predictions follow from this behavior? What choices of the parameters a, b, c, d best match the behavior observed in practice?

Example 6.5 (Smearing in an Imaging System)

Consider a model that describes the relationship between a two-dimensional object and its image on a planar film in a camera. Due to limited aperture, lens imperfections and focusing errors, the image of a unit point source at the origin

in the object, represented by the unit impulse $\delta(x, y)$ in the object plane, will be smeared. The intensity of the light at the image may be modeled by some function $h(x, y)$, $x, y \in \mathbb{R}$, for example $h(x, y) = e^{-a(x^2+y^2)}$. An object $u(x, y)$ can be viewed as the superposition of individual points distributed spatially, i.e.,

$$u(x, y) = \int \int_{-\infty}^{\infty} \delta(x - \lambda, y - \mu)\, u(\lambda, \mu) d\lambda\, d\mu \ .$$

Assuming that the effect of the lens is linear and translation invariant, the image of such an object is given by the following intensity function:

$$m(x, y) = \int \int_{-\infty}^{\infty} h(x - \lambda, y - \mu)\, u(\lambda, d\mu) d\lambda\, d\mu$$

We can view u as the input to this system, m as the output.

6.2 System Representations

There are two general representations of a dynamic model that we shall be interested in, namely behavioral and input-output description.

6.2.1 Behavioral Models

This a very general representation, which we have actually taken as the basis for our initial definition of a dynamic model. In this representation, the system is described as a collection of constraints on designated signals, w_i. Any combination $w(t) = [\ w_1(t),\ \cdots,\ w_\ell(t)\ ]$ of signals that satisfies the constraints is *a behavior* of the model, $w(t) \in \mathbb{B}$, where $\mathbb{B}$ denotes *the behavior*. An example of such a representation is Example 6.1.

Linearity

We call a model **linear** if its behavior constitutes a vector space, i.e. if *superposition* applies:

$$w_a(t),\ w_b(t) \in \mathbb{B} \implies \alpha\, w_a(t) + \beta\, w_b(t) \in \mathbb{B} \tag{6.15}$$

where α and β are arbitrary scalars. Example 6.1 is evidently linear.

Time-Invariance

We call a model **time-invariant** (or translation-invariant, or shift-invariant) if every possible time shift of a behavior — in which each of the signals is shifted by the same amount — yields a behavior:

$$w(t) \in \mathbb{B} \implies \sigma_\tau w(t) = w(t - \tau) \in \mathbb{B}, \tag{6.16}$$

for all valid τ, i.e. τ for which $\mathbb{T} - \tau \subset \mathbb{T}$, with σ_τ denoting the τ-*shift operator*. Example 6.1 is evidently time-invariant.

Memoryless Models

A model is **memoryless** if the constraints that describe the associated signals $w(\cdot)$ are purely *algebraic*, i.e., they only involve constraints on $w(t_0)$ for each $t_0 \in \mathbb{T}$ (and so do *not* involve derivatives, integrals, etc.). More interesting to us are non-memoryless, or *dynamic* systems, where the constraints involve signal values at different times.

6.2.2 Input-Output Models

For this class of models, the system is modeled as a *mapping* from a set of input signals $u(t)$ to a set of output signals, $y(t)$. We may represent this map as

$$y(t) = (S\ u)\,(t) \tag{6.17}$$

(i.e., the result of operating on the entire signal $u(\cdot)$ with the mapping S yields the signal $y(\cdot)$, and the particular value of the output at some time t is then denoted as above). The above mapping clearly also constitutes a constraint relating $u(t)$ and $y(t)$; this fact could be emphasized by trivially rewriting the equation in the form

$$y(t) - (S\ u)\,(t) = 0\ . \tag{6.18}$$

The definitions of linearity, time-invariance and memorylessness from the behavioral case therefore specialize easily to mappings. An example of a system representation in the form of a mapping is Example 6.5.

Linearity and Time-Invariance

From the behavioral point of view, the signals of interest are given by $w(t) = [u(t)\ y(t)]$. It then follows from the preceding discussion of behavioral models that the model is **linear** if and only if

$$\Big(S\ (\alpha u_a + \beta u_b)\Big)(t)\ = \alpha y_a(t) + \beta y_b(t) = \alpha(S\ u_a)(t) + \beta(S\ u_b)(t) \tag{6.19}$$

and the model is **time-invariant** if and only if

$$\Big(S\ \sigma_\tau u\Big)(t) = (\sigma_\tau y)(t) = y(t-\tau) \tag{6.20}$$

where σ_τ is again the τ-shift operator (so time-invariance of a mapping corresponds to requiring mapping to commute with the shift operator).

Memoryless Models

Again specializing the behavioral definition, we see that a mapping is **memoryless** if and only if $y(t_0)$ only depends on $u(t_0)$, for every $t_0 \in \mathbb{T}$:

$$y(t_0) = (S\ u)\,(t_0)\ = f\Big(u(t_0)\Big)\ \ . \tag{6.21}$$

Causality

We say the mapping is **causal** if the output does not depend on future values of the input. To describe causality conveniently in mathematical form, define the *truncation operator* P_T on a signal by the condition

$$(P_T\ u)\,(t) = \begin{cases} u(t) & \text{for } t \le T \\ 0 & \text{for } t > T \end{cases} \quad . \tag{6.22}$$

Thus, if u is a record of a function over all time, then $(P_T\ u)$ is a record of u up to time T, trivially extended by 0. Then the system S is said to be causal if

$$P_T S P_T = P_T S \quad . \tag{6.23}$$

In other words, the output up to time T depends only on the input up to time T.

Example 6.6 Example 6.5 shows a system represented as an input-output map. It is evident that the model is linear, translation-invariant, and not memoryless (unless $h(x, y) = \delta(x, y)$).

Notes

For much more on the behavioral approach to modeling and analysis of dynamic systems, see

J. C. Willems, "Paradigms and Puzzles in the Theory of Dynamic Systems," *IEEE Transactions on Automatic Control,* Vol. 36, pp. 259–294, March 1991.

Exercises

Exercise 6.1 Suppose the output $y(t)$ of a system is related to the input $u(t)$ via the following relation:

$$y(t) = \int_0^\infty e^{-(t-s)} u(s) ds.$$

Verify that the model is linear, time-varying, non-causal, and not memoryless.

Exercise 6.2 Suppose the input-output relation of a system is given by

$$y(t) = \begin{cases} u(t) & \text{if } |u(t)| \leq 1 \\ \frac{u(t)}{|u(t)|} & \text{if } |u(t)| > 1 \end{cases}.$$

This input-output relation represents a *saturation* element. Is this map nonlinear? Is it memoryless?

Exercise 6.3 Consider a system modeled as a map from $u(t)$ to $y(t)$, and assume you know that when

$$u(t) = \begin{cases} 1 & \text{for } 1 \leq t \leq 2 \\ 0 & \text{otherwise} \end{cases},$$

the corresponding output is

$$y(t) = \begin{cases} e^{t-1} - e^{t-2} & \text{for } t \leq 1 \\ 2 - e^{1-t} - e^{t-2} & \text{for } 1 \leq t \leq 2 \\ e^{2-t} - e^{1-t} & \text{for } t \geq 2 \end{cases}.$$

In addition, the system takes the zero input to the zero output. Is the system causal? Is it memoryless?

A particular mapping that is consistent with the above experiment is described by

$$y(t) = \int_{-\infty}^{\infty} e^{-|t-s|} u(s) ds. \tag{6.24}$$

Is the model linear? Is it time-invariant?

Exercise 6.4 For each of the following maps, determine whether the model is (a) linear, (b) time-invariant, (c) causal, (d) memoryless.

(i)

$$y(t) = \int_0^t (t-s)^3 u(s) ds$$

(ii)

$$y(t) = 1 + \int_0^t (t-s)^3 u(s) ds$$

(iii)

$$y(t) = u^3(t)$$

(iv)

$$y(t) = \int_0^t e^{-ts} u(s) ds$$

Chapter 7

State-Space Models

7.1 Introduction

A central question in dealing with a causal discrete-time (DT) system with input u, output y, is the following:

Given the input at some time n, i.e. given $u[n]$, how much information do we need about past inputs, i.e. about $u[k]$ for $k < n$, in order to determine the present output, namely $y[n]$?

The same question can be asked for continuous-time (CT) systems. This question addresses the issue of *memory* in the system. Why is this a central question? Some reasons:

- The answer gives us an idea of the complexity, or number of degrees of freedom, associated with the dynamic behavior of the system. The more information we need about past inputs in order to determine the present output, the richer the variety of possible output behaviors.
- In a control application, the answer to the above question suggests the required degree of complexity of the controller, because the controller has to remember enough about the past to determine the effects of present control actions on the response of the system.
- For a computer algorithm that acts causally on a data stream, the answer to the above question suggests how much memory will be needed to run the algorithm.

We now describe the general structure of *state-space models*, for which the preceding question has an immediate and transparent answer.

7.2 General Description

For a *causal* system with m inputs $u_j(t)$ and p outputs $y_i(t)$ (hence $m+p$ manifest variables), an nth-order state-space description is one that introduces n latent variables $x_\ell(t)$ called *state variables* in order to obtain a particular form for the constraints that define the model. Letting

$$u(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_m(t) \end{bmatrix} \quad , \quad y(t) = \begin{bmatrix} y_1(t) \\ \vdots \\ y_p(t) \end{bmatrix} \quad , \quad x(t) = \begin{bmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{bmatrix} \quad ,$$

an nth-order state-space description takes the form

$$\dot{x}(t) = f\left(x(t), u(t), t\right) \quad \text{(state evolution equations)} \tag{7.1}$$

$$y(t) = g\left(x(t), u(t), t\right) \quad \text{(instantaneous output equations)} \ . \tag{7.2}$$

To save writing the same equations over for both continuous and discrete time, we interpret

$$\dot{x}(t) = \frac{dx(t)}{dt} \quad , \quad t \in \mathbb{R} \text{ or } \mathbb{R}^+$$

for CT systems, and

$$\dot{x}(t) = x(t+1) \quad , \quad t \in \mathbb{Z} \text{ or } \mathbb{Z}^+$$

for DT systems. We will only consider finite-order (or finite-dimensional, or *lumped*) state-space models, although there is also a rather well developed (but much more subtle and technical) theory of infinite-order (or infinite-dimensional, or *distributed*) state-space models.

DT Models

The key feature of a state-space description is the following property, which we shall refer to as the *state property*. Given the present state vector (or "state") and present input at time t, we can compute: (i) the present output, using (7.2); and (ii) the next state using (7.1). It is easy to see that this puts us in a position to do the same thing at time $t+1$, and therefore to continue the process over any time interval. Extending this argument, we can make the following claim:

State Property of DT state-Space Models
Given the initial state $x(t_0)$
and input $u(t)$ for $t_0 \le t < t_f$
(with t_0 and t_f arbitrary),
we can compute the output $y(t)$ for $t_0 \le t < t_f$
and the state $x(t)$ for $t_0 < t \le t_f$.

Thus, the state at any time t_0 summarizes everything about the past that is relevant to the future. Keeping in mind this fact — that the state variables are the *memory variables* (or, in more physical situations, the *energy storage* variables) of a system — often guides us quickly to good choices of state variables in any given context.

CT Models

The same state property turns out to hold in the CT case, at least for $f(\,.\,)$ that are well behaved enough for the state evolution equations to have a unique solution for all inputs of interest and over the entire time axis — these will typically be the only sorts of CT systems of interest to us. A demonstration of this claim, and an elucidation of the precise conditions under which it holds, would require an excursion into the theory of differential equations beyond what is appropriate for this course. We can make this result plausible, however, by considering the Taylor series approximation

$$\begin{aligned} x(t_0+\epsilon) &\approx x(t_0) + \left(\frac{dx(t)}{dt}\right)_{t=t_0} \epsilon & (7.3)\\ &= x(t_0) + f\left(x(t_0), u(t_0), t_0\right) \epsilon & (7.4) \end{aligned}$$

where the second equation results from applying the state evolution equation (7.1). This suggests that we can approximately compute $x(t_0+\epsilon)$, given $x(t_0)$ and $u(t_0)$; the error in the approximation is of order ϵ^2, and can therefore be made smaller by making ϵ smaller. For sufficiently well behaved $f(\,\cdot\,)$, we can similarly step forwards from $t_0+\epsilon$ to $t_0+2\epsilon$, and so on, eventually arriving at the final time t_f, taking on the order of ϵ^{-1} steps in the process. The accumulated error at time t_f is then of order $\epsilon^{-1}.\epsilon^2 = \epsilon$, and can be made arbitrarily small by making ϵ sufficiently small. Also note that, once the state at any time is determined and the input at that time is known, then the output at that time is immediately given by (7.2), even in the CT case.

The simple-minded Taylor series approximation in (7.4) corresponds to the crudest of numerical schemes — the "forward Euler" method — for integrating a system of equations of the form (7.1). Far more sophisticated schemes exist (e.g. Runge-Kutta methods, Adams-Gear schemes for "stiff" systems that exhibit widely differing time scales, etc.), but the forward Euler scheme suffices to make plausible the fact that the state property highlighted above applies to CT systems as well as DT ones.

Example 7.1 RC Circuit

This example demonstrates a fine point in the definition of a state for CT systems. Consider an RC circuit in series with a voltage source u. Using KVL, we get the following equation describing the system:

$$-u + v_R + RC\dot{v}_C = 0.$$

It is clear that v_C defines a state for the system as we described before. Does v_R define a state? If $v_R(t_0)$ is given, and the input $u(t)$, $t_0 \le t < t_f$ is known, then

one can compute $v_C(t_0)$ and using the state property $v_C(t_f)$ can be computed from which $v_R(t_f)$ can be computed. This says that $v_R(t)$ defines a state which contradicts our intuition since it is not an energy storage component.

There is an easy fix of this problem if we assume that all inputs are piece-wise continuous functions. In that case we define the state property as the ability to compute future values of the state from the initial value $x(t_0)$ and the input $u(t)$, $t_0 < t < t_f$. Notice the strict inequality. We leave it to you to verify that this definition rules out v_R as a state variable.

Linearity and Time-Invariance

If in the state-space description (7.1), (7.2), we have

$$f\left(x(t), u(t), t\right) = f\left(x(t), u(t)\right) \tag{7.5}$$
$$g\left(x(t), u(t), t\right) = g\left(x(t), u(t)\right) \tag{7.6}$$

then the model is *time-invariant* (in the sense defined earlier, for behavioral models). This corresponds to requiring time-invariance of the functions that specify how the state variables and inputs are combined to determine the state evolution and outputs. The results of experiments on a time-invariant system depend only on the inputs and initial state, not on *when* the experiments are performed.

If, on the other hand, the functions $f(\,.\,)$ and $g(\,.\,)$ in the state-space description are linear functions of the state variables and inputs, i.e. if

$$f\left(x(t), u(t), t\right) = A(t)x(t) + B(t)u(t) \tag{7.7}$$
$$g\left(x(t), u(t), t\right) = C(t)x(t) + D(t)u(t) \tag{7.8}$$

then the model is *linear*, again in the behavioral sense. The case of a *linear and periodically varying* (LPV) model is often of interest; when $A(t) = A(t+T)$, $B(t) = B(t+T)$, $C(t) = C(t+T)$, and $D(t) = D(t+T)$ for all t, the model is LPV with period T.

Of even more importance to us is the case of a model that is **linear** *and* **time-invariant** (LTI). For an LTI model, the state-space description simplifies to

$$f\left(x(t), u(t), t\right) = Ax(t) + Bu(t) \tag{7.9}$$
$$g\left(x(t), u(t), t\right) = Cx(t) + Du(t)\ . \tag{7.10}$$

We will primarily study LTI models in this course. Note that LTI state-space models are sometimes designated as (A, B, C, D) or

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

as these four matrices completely specify the state-space model.

System	Type
$\dot{x}(t) = tx^2(t)$	NLTV
$\dot{x}(t) = x^2(t)$	NLTI
$\dot{x}(t) = tx(t)$	LTV
$\dot{x}(t) = (\cos t)x(t)$	LPV
$\dot{x}(t) = x(t)$	LTI

Table 7.1: Some examples of linear, nonlinear, time-varying, periodically-varying, and time-invariant state-space descriptions.

Some examples of the various classes of systems listed above are given in Table 7.1. More elaborate examples follow.

One might think that the state-space formulation is restrictive since it only involves first-order derivatives. However, by appropriately choosing the state variables, higher-order dynamics can be described. The examples in this section and on homework will make this clear.

Example 7.2 (Mass-Spring System)

For the mass-spring system in Example 6.2, we derived the following system representation:

$$M\ddot{z} = -kz + u.$$

To put this in state space form, choose position and velocity as state variables:

$$\begin{aligned} x_1 &= z \\ x_2 &= \dot{z}. \end{aligned} \tag{7.11}$$

Therefore,

$$\begin{aligned} \dot{x}_1 &= \dot{z} = x_2 \\ \dot{x}_2 &= -\frac{k}{M}z + \frac{1}{M}u = -\frac{k}{M}x_1 + \frac{1}{M}u \ . \end{aligned}$$

The input is the force u and let the output be the position of the mass. The resulting state space description of this system is

$$\begin{aligned} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} &= \begin{bmatrix} x_2 \\ -\frac{k}{M}x_1 + \frac{1}{M}u \end{bmatrix} \\ y &= x_1 \quad . \end{aligned}$$

The above example suggests something that is true in general for mechanical systems: the natural state variables are the position and velocity variables (associated with potential energy and kinetic energy respectively).

Example 7.3 (Nonlinear Circuit)

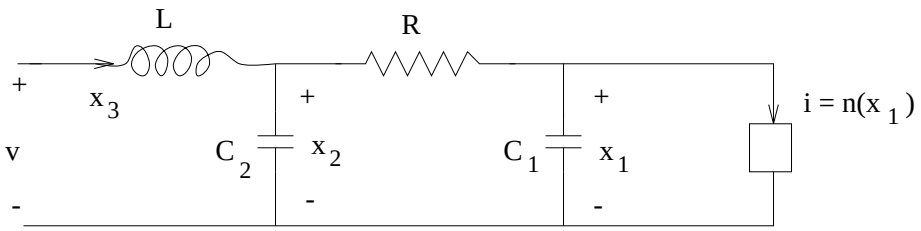


Figure 7.1: Nonlinear circuit.

We wish to put the relationships describing the above circuit's behavior in state-space form, taking the voltage v as an input, and choosing as output variables the voltage across the nonlinear element and the current through the inductor. The constituent relationship for the nonlinear admittance in the circuit diagram is $i_{nonlin} = \mathcal{N}(v_{nonlin})$, where $\mathcal{N}(\,.\,)$ denotes some nonlinear function.

Let us try taking as our state variables the capacitor voltages and inductor current, because these variables represent the energy storage mechanisms in the circuit. The corresponding state-space description will express the rates of change of these variables in terms of the instantaneous values of these variables and the instantaneous value of the input voltage v. It is natural, therefore, to look for expressions for $C_1\dot{x}_1$ (the current through C_1), for $C_2\dot{x}_2$ (the current through C_2), and for $L\dot{x}_3$ (the voltage across L).

Applying KCL to the node where R, C_1, and the nonlinear device meet, we get

$$C_1\dot{x}_1 = \frac{(x_2 - x_1)}{R} - \mathcal{N}(x_1)$$

Applying KCL to the node where R, C_2 and L meet, we find

$$C_2\dot{x}_2 = x_3 - \frac{(x_2 - x_1)}{R}$$

Finally, KVL applied to a loop containing L yields

$$L\dot{x}_3 = v - x_2$$

Now we can combine these three equations to obtain a state-space description of this system:

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{C_1}\left(\frac{x_2 - x_1}{R} - \mathcal{N}(x_1)\right) \\ \frac{1}{C_2}\left(x_3 - \frac{x_2 - x_1}{R}\right) \\ -\frac{1}{L}x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L}v \end{bmatrix} \tag{7.12}$$

$$y = \begin{bmatrix} x_1 \\ x_3 \end{bmatrix} . \tag{7.13}$$

Observe that the output variables are described by an instantaneous output equation of the form (7.2). This state-space description is time-invariant but nonlinear. This makes sense, because the circuit does contain a nonlinear element!

Example 7.4 (Discretization)

Assume we have a continuous-time system described in state-space form by

$$\begin{aligned} \frac{dx(t)}{dt} &= Ax(t) + Bu(t), \\ y(t) &= Cx(t) + Du(t). \end{aligned}$$

Let us now sample this system with a period of T, and approximate the derivative as a forward difference:

$$\frac{1}{T}\left(x\left((k+1)T\right) - x\left(kT\right)\right) = Ax\left(kT\right) + Bu\left(kT\right), \qquad k \in \mathbb{Z}. \tag{7.14}$$

It is convenient to change our notation, writing $x[k] \equiv x(kT)$, and similarly for u and y. Our sampled equation can thereby be rewritten as

$$\begin{aligned} x[k+1] &= \left(I + TA\right)x[k] + TBu[k] \\ &= \hat{A}\mathbf{x}[k] + \hat{B}\mathbf{u}[k] , \\ y[k] &= Cx[k] + Du[k] . \end{aligned} \tag{7.15}$$

which is in standard state-space form.

In many modern applications, control systems are implemented digitally. For that purpose, the control engineer must be able to analyze both discrete-time as well as continuous-time systems. In this example a crude sampling method was used to obtain a discrete-time model from a continuous-time one. We will discuss more refined discretization methods later on in this book.

It is also important to point out that there are physical phenomena that directly require or suggest discrete-time models; not all discrete-time models that one encounters in applications are discretizations of continuous-time ones.

7.3 Linearization

Much of our attention in this course will be focused on *linear* models. Linear models frequently arise as descriptions of small perturbations away from a nominal solution of the system. Consider, for example, the continuous-time (CT) state-space model

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t), t) \\ y(t) &= g(x(t), u(t), t) \end{aligned} \tag{7.16}$$

where $x(t)$ is the n-dimensional state-vector at time t, $u(t)$ is the m-dimensional vector of inputs, and $y(t)$ is the p-dimensional vector of outputs. Suppose $x_o(t)$, $u_o(t)$ and $y_o(t)$ constitute a *nominal solution* of the system, i.e. a collection of CT signals that jointly satisfy the equations in (7.16). Now let the control and initial condition be perturbed from their nominal values to $u(t) = u_o(t) + \delta u(t)$ and $x(0) = x_o(0) + \delta x(0)$ respectively, and let the state trajectory accordingly be perturbed to $x(t) = x_o(t) + \delta x(t)$. Substituting these new values into (7.16) and performing a (multivariable) Taylor series expansion to first-order terms, we find

$$\begin{aligned} \dot{\delta x}(t) &\approx \left[\frac{\partial f}{\partial x}\right]_o \delta x(t) + \left[\frac{\partial f}{\partial u}\right]_o \delta u(t) \\ \delta y(t) &\approx \left[\frac{\partial g}{\partial x}\right]_o \delta x(t) + \left[\frac{\partial g}{\partial u}\right]_o \delta u(t) \end{aligned} \tag{7.17}$$

where the $n \times n$ matrix $[\partial f/\partial x]_o$ denotes the Jacobian of $f(.,.,.)$ with respect to x, i.e. a matrix whose ij-th entry is the partial derivative of the ith component of $f(.,.,.)$ with respect to the jth component of x, and where the other Jacobian matrices in (7.17) are similarly defined. The subscript $_o$ indicates that the Jacobians are evaluated along the nominal trajectory, i.e. at $x(t) = x_o(t)$ and $u(t) = u_o(t)$. The linearized model (7.17) is evidently linear, of the form

$$\begin{aligned} \dot{\delta x}(t) &= A(t)\,\delta x(t) + B(t)\,\delta u(t) \\ \delta y(t) &= C(t)\,\delta x(t) + D(t)\,\delta u(t). \end{aligned} \tag{7.18}$$

When the original nonlinear model is time-invariant, the linearized model will also be time-invariant *if the nominal solution is constant* (i.e. if the nominal solution corresponds to a constant *equilibrium*); however, the linearized model *may be time varying* if the nominal solution is time varying (even if the original nonlinear model is time-invariant), and will be *periodic* — i.e., have periodically varying coefficients — if the nominal solution is periodic (as happens when the nominal solution corresponds to operation in some cyclic or *periodic steady state*).

The same development can be carried out for discrete-time (DT) systems, but we focus in this lecture on the CT case.

Example 7.5 (Linearizing a Nonlinear Circuit Model)

Consider linearizing the state-space model we obtained for the nonlinear circuit in Example 7.3. We ended up there with a nonlinear model of the form

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{C_1}\left(\frac{x_2 - x_1}{R} - \mathcal{N}(x_1)\right) \\ \frac{1}{C_2}\left(x_3 - \frac{x_2 - x_1}{R}\right) \\ -\frac{1}{L}x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \frac{1}{L}v \end{bmatrix} \quad . \tag{7.19}$$

For the linearization, all that happens is each x_j is replaced by δx_j, and $\mathcal{N}(x_1)$ is replaced by $[d\mathcal{N}(x_1)/dx_1]_o\,\delta x_1$, resulting in a linear state-space model of the form

$$\dot{\delta x}(t) = A\,\delta x(t) + B\,\delta v(t) \tag{7.20}$$

with

$$A = \begin{pmatrix} -\frac{1}{RC_1} - \frac{1}{C_1}\left[\frac{d\mathcal{N}}{dx_1}\right]_o & \frac{1}{RC_1} & 0 \\ \frac{1}{RC_2} & -\frac{1}{RC_2} & \frac{1}{C_2} \\ 0 & \frac{1}{L} & 0 \end{pmatrix} \quad , \quad B = \begin{pmatrix} 0 \\ 0 \\ \frac{1}{L} \end{pmatrix} \tag{7.21}$$

Example 7.6 (Linearizing the Inverted Pendulum)

Recall from Example 6.3 the equations that describe the dynamics of the inverted pendulum. Those equations are nonlinear due to the presence of the terms $\sin(\theta)$, $\cos(\theta)$, and $(\dot{\theta})^2$. We can linearize these equations around $\theta = 0$ and $\dot{\theta} = 0$, by assuming that $\theta(t)$ and $\dot{\theta}(t)$ remain small. Recall that for small θ

$$\begin{aligned} \sin(\theta) \qquad & \theta - \frac{1}{6}\theta^3 \\ \cos(\theta) \qquad & 1 - \frac{1}{2}\theta^2, \end{aligned}$$

and using the linear parts of these relations the linearized system of equations takes the form

$$\begin{aligned} \left(1 - \frac{ml}{ML}\right)\ddot{s} + \frac{ml}{M}\frac{g}{L}\theta &= \frac{1}{M}u, \\ \left(1 - \frac{ml}{ML}\right)\ddot{\theta} - \frac{g}{L}\theta &= -\frac{1}{ML}u\,. \end{aligned}$$

Using as state vector

$$x = \begin{bmatrix} s \\ \dot{s} \\ \theta \\ \dot{\theta} \end{bmatrix},$$

the following state-space model can be easily obtained:

$$\begin{aligned} \frac{d}{dt}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & -\alpha\frac{ml}{ML}g & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & \alpha\frac{g}{L} & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} + \begin{pmatrix} 0 \\ \frac{\alpha}{M} \\ 0 \\ -\frac{\alpha}{LM} \end{pmatrix} u \\ y &= \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} x, \end{aligned}$$

where the constant α is given by

$$\alpha = \frac{1}{\left(1 - \frac{ml}{ML}\right)}.$$

Exercises

Exercise 7.1 Consider the nonlinear difference equation

$$y(k+n) = F[y(k+n-1), \ldots, y(k), u(k+n-1), \ldots, u(k), k]$$

where n is a fixed integer, and k is the time index.

(a) Find a state-space representation of order $2n-1$ for this difference equation.

(b) Find an nth-order state-space representation in LTI case (what is the form of F in this case?), using z-transforms for guidance (natural state variables are the coefficients of the initial-condition terms in the z-transformed version of the difference equation — try a third-order difference equation — remind of forward shift theorem from z-transforms). This part will guide the solution of (c).

(c) Find an nth-order state-space representation for the nonlinear system in (a) for the case where $F[\,.\,]$ has the special form

$$F[\,.\,] = \sum_{i=1}^{n} f_i[y(k+n-i), u(k+n-i)]$$

(Hint: Note that the difference equation in part (b) has this form; use your definition of state variables in (b) to guide your choice here.)

Exercise 7.2 Consider a causal continuous-time system with input-output representation $y(t) = h * u(t)$, where $*$ denotes convolution and $h(t)$ is the impulse response of the system:

$$h(t) = 2e^{-t} - ce^{-2t} \quad \text{for } t \geq 0$$

Here c denotes a constant.

(a) Suppose $c = 2$. Use only the input-output representation of the system to show that the variables $x_1(t) = y(t)$ and $x_2(t) = \dot{y}(t)$ qualify as state variables of the system at time t.

(b) Compute the transfer function of the system, and use it to describe what may be special about the case $c = 2$.

Exercise 7.3 The input $u(t)$ and output $y(t)$ of a system are related by the equation

$$\frac{dy(t)}{dt} + a_0(t)y(t) = b_0(t)u(t) + b_1(t)\frac{du(t)}{dt}$$

Find a linear, time varying state-space representation of this system.

Exercise 7.4 Given the *periodically varying* system $x(k+1) = A(k)x(k) + B(k)u(k)$ of period N, with $A(k+N) = A(k)$ and $B(k+N) = B(k)$, define the *sampled state* $z[k]$ and the associated *extended input vector* $v[k]$ by

$$z[k] = x(kN) \ , \qquad v[k] = \begin{pmatrix} u(kN) \\ u(kN+1) \\ \vdots \\ u(kN+N-1) \end{pmatrix}$$

Now show that $z[k+1] = Fz[k] + Gv[k]$ for *constant* matrices F and G (i.e. matrices independent of k) by determining F and G explicitly.

Exercise 7.5 Let the state space representations of two given systems be

$$x_i(k+1) = A_i x_i(k) + B_i u_i(k) \ , \quad y_i(k) = C_i x_i(k) \ , \quad i = 1, 2$$

Determine a state-space representation in the form

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) \\ y(k) &= Cx(k) \end{aligned}$$

for the new system obtained when systems 1 and 2 are interconnected (a) in series, (b) in parallel, and in a feedback loop. Assume the size of the inputs and outputs of the two systems are consistent for each of the above configuration to make sense.

Exercise 7.6 Consider a pendulum comprising a mass m at the end of a light but rigid rod of length r. The angle of the pendulum from its equilibrium position is denoted by θ. Suppose a torque $u(t)$ can be applied about the axis of support of the pendulum (e.g. suppose the pendulum is attached to the axis of an electric motor, with the current through the motor being converted to torque). A simple model for this system takes the form

$$mr^2\ddot{\theta}(t) + f\dot{\theta}(t) + mgr\sin\theta(t) = u(t)$$

where the term $f\dot{\theta}$ represents a frictional torque, with f being a positive coefficient, and g is the acceleration due to gravity.

(a) Find a state-space representation for this model. Is your state-space model linear? time invariant?

(b) What nominal input $u_o(t)$ corresponds to the nominal motion $\theta_o(t) = \Omega t$ for all t, where Ω is some fixed constant?

(c) Linearize your state-space model in (a) around the nominal solution in (b). Is the resulting model linear? Is it time invariant or periodically varying?

Exercise 7.7 Consider the horizontal motion of a particle of unit mass sliding under the influence of gravity on a frictionless wire. It can be shown that, if the wire is bent so that its height h is given by $h(x) = V_\alpha(x)$, then a state-space model for the motion is given by

$$\begin{aligned} \dot{x} &= z \\ \dot{z} &= -\frac{d}{dx} V_\alpha(x), \end{aligned}$$

Suppose $V_\alpha(x) = x^4 - \alpha x^2$.

(a) Verify that the above model has $(z, x) = (0, 0)$ as equilibrium point for any α in the interval $-1 \leq \alpha \leq 1$, and it also has $(z, x) = \left(0, \pm\sqrt{\frac{\alpha}{2}}\right)$ as equilibrium points when α is in the interval $0 < \alpha \leq 1$.

(b) Derive the linearized system at each of these equilibrium points.

Chapter 8

Simulation/Realization

8.1 Introduction

Given an nth-order state-space description of the form

$$\dot{x}(t) = f\left(x(t), u(t), t\right) \quad \text{(state evolution equations)} \tag{8.1}$$

$$y(t) = g\left(x(t), u(t), t\right) \quad \text{(instantaneous output equations)} \ . \tag{8.2}$$

(which may be CT or DT, depending on how we interpret the symbol $\dot{x}$), how do we *simulate* the model, i.e., how do we implement it or *realize* it in hardware or software? In the DT case, where $\dot{x}(t) = x(t+1)$, this is easy if we have available: (i) storage registers that can be updated at each time step (or "clock cycle") — these will store the state variables; and (ii) a means of evaluating the functions $f(\,\cdot\,)$ and $g(\,\cdot\,)$ that appear in the state-space description — in the linear case, all that we need for this are multipliers and adders. A straightforward realization is then obtained as shown in the figure below. The storage registers are labeled D for (one-step) *delay*, because the output of the block represents the data currently stored in the register while the input of such a block represents the data waiting to be read into the register at the *next* clock pulse. In the CT case, where $\dot{x}(t) = dx(t)/dt$, the only difference is that the delay elements are replaced by integrators. The outputs of the integrators are then the state variables.

8.2 Realization from I/O Representations

In this section, we will describe how a state space realization can be obtained for a causal input-output dynamic system described in terms of convolution.

8.2.1 Convolution with an Exponential

Consider a causal DT LTI system with impulse response $h[n]$ (which is 0 for $n < 0$):

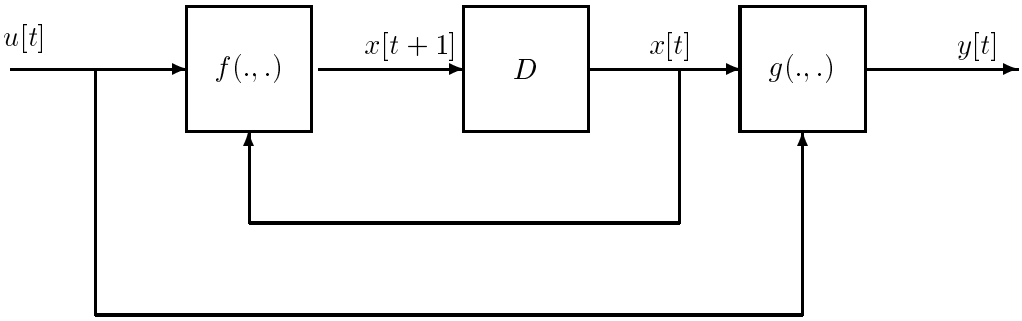


Figure 8.1: Simulation Diagram

$$\begin{aligned} y[n] &= \sum_{-\infty}^{n} h[n-k]u[k] \\ &= \Big(\sum_{-\infty}^{n-1} h[n-k]u[k]\Big) + h[0]u[n] \end{aligned} \tag{8.3}$$

The first term above, namely

$$x[n] = \sum_{-\infty}^{n-1} h[n-k]u[k] \tag{8.4}$$

represents the effect of the past on the present. This expression shows that, in general (i.e. if $h[n]$ has no special form), the number $x[n]$ has to be recomputed from scratch for each n. When we move from n to $n+1$, none of the past input, i.e. $u[k]$ for $k \leq n$, can be discarded, because all of the past will again be needed to compute $x[n+1]$. In other words, the memory of the system is infinite.

Now look at an instance where special structure in $h[n]$ makes the situation much better. Suppose

$$h[n] = \lambda^n \quad \text{for } n \geq 0, \text{ and 0 otherwise} \tag{8.5}$$

Then

$$x[n] = \sum_{-\infty}^{n-1} \lambda^{n-k} u[k] \tag{8.6}$$

and

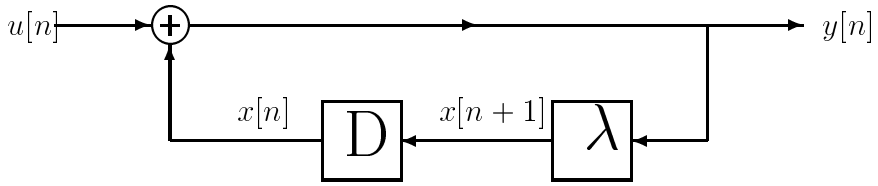
$$\begin{aligned} x[n+1] &= \sum_{-\infty}^{n} \lambda^{n+1-k} u[k] \\ &= \lambda\Big(\sum_{-\infty}^{n-1} \lambda^{n-k} u[k]\Big) + \lambda u[n] \\ &= \lambda x[n] + \lambda u[n] \end{aligned} \tag{8.7}$$

(You will find it instructive to graphically represent the convolutions that are involved here, in order to understand more visually why the relationship (8.7) holds.) Gathering (8.3) and (8.6) with (8.7), we obtain a pair of equations that together constitute a *state-space description* for this system:

$$x[n+1] = \lambda x[n] + \lambda u[n] \tag{8.8}$$

$$y[n] = x[n] + u[n] \tag{8.9}$$

To *realize* this model in hardware, or to *simulate* it, we can use a delay-adder-gain system that is obtained as follows. We start with a delay element, whose output will be $x[n]$ when its input is $x[n+1]$. Now the state evolution equation tells us how to combine the present output of the delay element, $x[n]$, with the present input to the system, $u[n]$, in order to obtain the present input to the delay element, $x[n+1]$. This leads to the following block diagram, in which we have used the output equation to determine how to obtain $y[n]$ from the present state and input of the system:

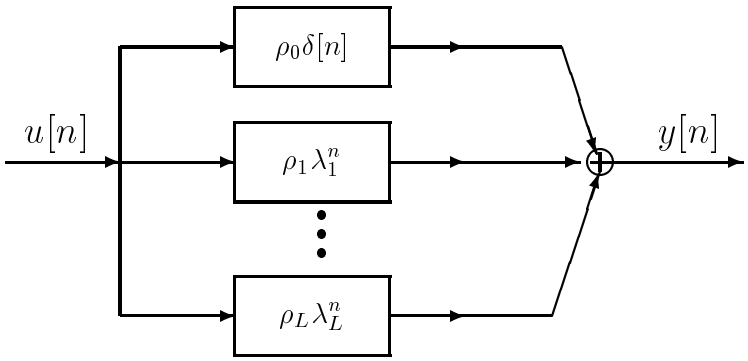


8.2.2 Convolution with a Sum of Exponentials

Consider a more complicated causal impulse response than the previous example, namely

$$h[n] = \rho_0 \delta[n] + (\rho_1 \lambda_1^n + \rho_2 \lambda_2^n + \cdots + \rho_L \lambda_L^n) \tag{8.10}$$

with the ρ_i being constants. The following block diagram shows that this system can be considered as being obtained through the parallel interconnection of causal subsystems that are as simple as the one treated earlier, plus a direct feedthrough of the input through the gain ρ_0 (each block is labeled with its impulse response, with causality implying that these responses are 0 for $n < 0$):



Motivated by the above structure and the treatment of the earlier, let us define a state variable for each of the L subsystems:

$$x_i[n] = \sum_{-\infty}^{n-1} \lambda_i^{n-k} u[k] \ , \quad i = 1, 2, \ldots, L \tag{8.11}$$

With this, we immediately obtain the following state-evolution equations for the subsystems:

$$x_i[n+1] = \lambda_i x_i[n] + \lambda_i u[n] \ , \quad i = 1, 2, \ldots, L \tag{8.12}$$

Also, after a little algebra, we directly find

$$y[n] = \rho_1 x_1[n] + \rho_2 x_2[n] + \cdots + \rho_L x_L[n] + (\sum_{0}^{L} \rho_i)\, u[n] \tag{8.13}$$

We have thus arrived at an Lth-order state-space description of the given system. To write the above state-space description in matrix form, define the state vector at time n to be

$$\mathbf{x}[n] = \begin{pmatrix} x_1[n] \\ x_2[n] \\ \vdots \\ x_L[n] \end{pmatrix} \tag{8.14}$$

Also define the diagonal matrix $\mathbf{A}$, column vector $\mathbf{b}$, and row vector $\mathbf{c}$ as follows:

$$\mathbf{A} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & \lambda_L \end{pmatrix} \ , \quad \mathbf{b} = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_L \end{pmatrix} \tag{8.15}$$

$$\mathbf{c} = \begin{pmatrix} \rho_1 & \rho_2 & \cdots & \cdots & \cdots & \rho_L \end{pmatrix} \tag{8.16}$$

Then our state-space model takes the desired matrix form, as you can easily verify:

$$\mathbf{x}[n+1] = \mathbf{A}\mathbf{x}[n] + \mathbf{b}u[n] \tag{8.17}$$

$$y[n] = \mathbf{c}\mathbf{x}[n] + \mathtt{d}u[n] \tag{8.18}$$

where

$$\mathtt{d} = \sum_{0}^{L} \rho_i \tag{8.19}$$

8.3 Realization from an LTI Differential/Difference equation

In this section, we describe how a realization can be obtained from a difference or a differential equation. We begin with an example.

Example 8.1 (State-Space Models for an LTI Difference Equation)

Let us examine some ways of representing the following input-output difference equation in state-space form:

$$y[n] + a_1 y[n-1] + a_2 y[n-2] = b_1 u[n-1] + b_2 u[n-2] \tag{8.20}$$

For a first attempt, consider using as state vector the quantity

$$\mathbf{x}[n] = \begin{pmatrix} y[n-1] \\ y[n-2] \\ u[n-1] \\ u[n-2] \end{pmatrix} \tag{8.21}$$

The corresponding 4th-order state-space model would take the form

$$\begin{aligned} \mathbf{x}[n+1] = \begin{pmatrix} y[n] \\ y[n-1] \\ u[n] \\ u[n-1] \end{pmatrix} &= \begin{pmatrix} -a_1 & -a_2 & b_1 & b_2 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} y[n-1] \\ y[n-2] \\ u[n-1] \\ u[n-2] \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} u[n] \\ y[n] &= \begin{pmatrix} -a_1 & -a_2 & b_1 & b_2 \end{pmatrix} \begin{pmatrix} y[n-1] \\ y[n-2] \\ u[n-1] \\ u[n-2] \end{pmatrix} + (0)\, u[n] \end{aligned} \tag{8.22}$$

If we are somewhat more careful about our choice of state variables, it is possible to get more economical models. For a 3rd-order model, suppose we pick as state vector

$$\mathbf{x}[n] = \begin{pmatrix} y[n] \\ y[n-1] \\ u[n-1] \end{pmatrix} \tag{8.23}$$

The corresponding 3rd-order state-space model takes the form

$$\begin{aligned} \mathbf{x}[n+1] = \begin{pmatrix} y[n+1] \\ y[n] \\ u[n] \end{pmatrix} &= \begin{pmatrix} -a_1 & -a_2 & b_2 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ y[n-1] \\ u[n-1] \end{pmatrix} + \begin{pmatrix} b_1 \\ 0 \\ 1 \end{pmatrix} u[n] \\ y[n] &= \begin{pmatrix} 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ y[n-1] \\ u[n-1] \end{pmatrix} + (0)\, u[n] \end{aligned} \tag{8.24}$$

A still more clever/devious choice of state variables yields a 2nd-order state-space model. For this, pick

$$\mathbf{x}[n] = \begin{pmatrix} y[n] \\ -a_2 y[n-1] + b_2 u[n-1] \end{pmatrix} \tag{8.25}$$

The corresponding 2nd-order state-space model takes the form

$$\begin{aligned} \begin{pmatrix} y[n+1] \\ -a_2 y[n] + b_2 u[n] \end{pmatrix} &= \begin{pmatrix} -a_1 & 1 \\ -a_2 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ -a_2 y[n-1] + b_2 u[n-1] \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} u[n] \\ y[n] &= \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} y[n] \\ -a_2 y[n-1] + b_2 u[n-1] \end{pmatrix} + (\,0\,)\, u[n] \end{aligned} \tag{8.26}$$

It turns out to be impossible in general to get a state-space description of order lower than 2 in this case. This should not be surprising, in view of the fact that we started with a 2nd-order difference equation, which we know (from earlier courses!) requires two initial conditions in order to solve forwards in time. Notice how, in each of the above cases, we have incorporated the information contained in the original difference equation that we started with.

This example was built around a second-order difference equation, but has natural generalizations to the nth-order case, and natural parallels in the case of CT differential equations.

Next, we will present two realizations of an nth-Order LTI differential equation. While realizations are not unique, these two have certain nice properties that will be discussed in the future.

8.3.1 Observability Canonical Form

Suppose we are given the LTI differential equation

$$y^{(n)} + a_{n-1} y^{(n-1)} + \cdots + a_0 y = b_0 u + b_1 \dot{u} + \cdots + b_{n-1} u^{(n-1)},$$

which can be rearranged as

$$y^{(n)} = (b_{n-1} u^{(n-1)} - b_{n-1} y^{(n-1)}) + (b_{n-2} u^{(n-2)} - a_{n-2} y^{(n-2)}) + \cdots + (b_0 u - a_0 y).$$

Integrated n times, this becomes

$$y = \int (b_{n-1} u - a_{n-1} y) + \int \int (b_{n-2} u - a_{n-2} y) + \cdots + \int \underset{n}{\cdots} \int (b_0 u - a_0 y). \tag{8.27}$$

The block diagram given in Figure 8.2 then follows directly from (8.27). This particular realization is called the *observability canonical form* realization — "canonical" in the sense of

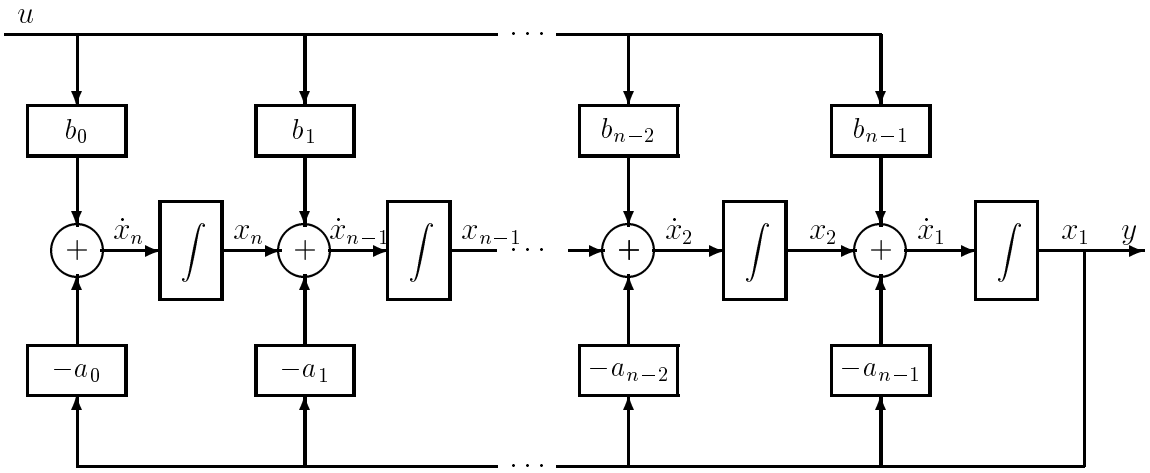


Figure 8.2: Observability Canonical Form

"simple" (but there is actually a strict mathematical definition as well), and "observability" for reasons that will emerge later in the course.

We can now read the state equations directly from Figure 8.2, once we recognize that the natural state variables are the outputs of the integrators:

$$\begin{aligned}
\dot{x}_1 &= -a_{n-1}x_1 + x_2 + b_{n-1}u \\
\dot{x}_2 &= -a_{n-2}x_1 + x_3 + b_{n-2}u \\
&\vdots \\
\dot{x}_n &= -a_0 x_1 + b_0 u \\
\\
y &= x_1.
\end{aligned}$$

If this is written in our usual matrix form, we would have

$$A = \begin{bmatrix} -a_{n-1} & 1 & 0 & \cdots & 0 \\ -a_{n-2} & 0 & 1 & \cdots & 0 \\ \vdots & & & \ddots & \\ & & & & 1 \\ -a_0 & 0 & & \cdots & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} b_{n-1} \\ b_{n-2} \\ \vdots \\ \vdots \\ b_0 \end{bmatrix}$$

$$c = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}.$$

The matrix A is said to be in *companion form*, a term used to refer to any one of four matrices whose pattern of 0's and 1's is, or resembles, the pattern seen above. The characteristic polynomial of such a matrix can be directly read off from the remaining coefficients, as we shall

see when we talk about these polynomials, so this matrix is a "companion" to its characteristic polynomial.

8.3.2 Reachability Canonical Form

There is a "dual" realization to the one presented in the previous section for the LTI differential equation

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_0 y = c_0 u + c_1 \dot{u} + \cdots + c_{n-1}u^{(n-1)}. \tag{8.28}$$

First, consider a special case of this, namely the differential equation

$$w^{(n)} + a_{n-1}w^{(n-1)} + \cdots + a_0 w = u \tag{8.29}$$

To obtain an nth-order state-space realization of the system in 8.29, define

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} w \\ \dot{w} \\ \ddot{w} \\ \vdots \\ \frac{d^{n-2}w}{dt^{n-2}} \\ \frac{d^{n-1}w}{dt^{n-1}} \end{bmatrix}.$$

Then it is easy to verify that the following state-space description represents the given model:

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 0 & 1 & \dots & 0 \\ \vdots & & & & \vdots & \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -a_0(t) & -a_1(t) & -a_2(t) & \dots & -a_{n-2}(t) & -a_{n-1}(t) \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} u$$

$$w = \begin{bmatrix} 1 & 0 & 0 & 0 & \dots & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix}.$$

(The matrix A here is again in one of the companion forms; the two remaining companion forms are the transposes of the one here and the transpose of the one in the previous section.) Suppose now that we want to realize another special case, namely the differential equation

$$r^{(n)} + a_{n-1}r^{(n-1)} + \cdots + a_0 r = \dot{u} \tag{8.30}$$

which is the same equation as (8.29), except that the RHS is $\dot{u}$ rather than u. By linearity, the response of (8.30) will $r = \dot{w}(t)$, and this response can be obtained from the above realization by simply taking the output to be x_2 rather than x_1, since $x_2 = \dot{w} = r$.

Superposing special cases of the preceding form, we see that if we have the differential equation (8.28), with an RHS of

$$c_0 u + c_1 \dot{u} + \cdots + c_{n-1} u^{(n-1)}$$

then the above realization suffices, provided we take the output to be

$$y = c_0 x_1 + c_1 x_2 + \cdots + c_{n-1} x_n. \tag{8.31}$$

i.e., we just change the output equation to have

$$c = \begin{bmatrix} c_0 & c_1 & c_2 & \cdots & c_{n-1} \end{bmatrix}. \tag{8.32}$$

A block diagram of the final realization is shown below in 8.3. This is called the *reachability or controllability canonical form*.

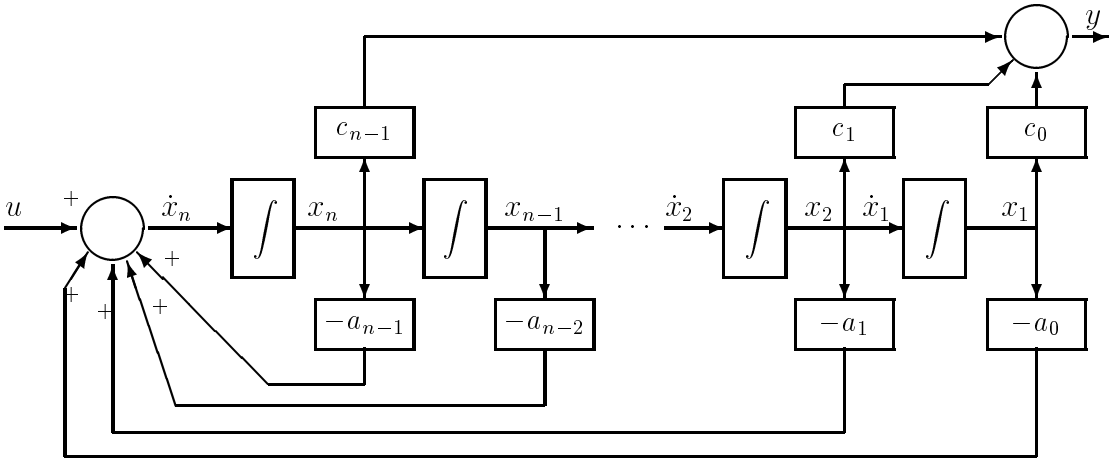


Figure 8.3: Reachability Canonical Form

Finally, for the obvious DT difference equation that is analogous to the CT differential equation that we used in this example, the same scheme will work, with derivatives replaced by differences.

Exercises

Exercise 8.1 Suppose we wish to realize a *two-input* differential equation of the form

$$\begin{aligned} y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_0 y = \quad & b_{01}u_1 + b_{11}\dot{u}_1 + \cdots + b_{n-1,1}u_1^{(n-1)} \\ + \quad & b_{02}u_2 + b_{12}\dot{u}_2 + \cdots + b_{n-1,2}u_2^{(n-1)} \end{aligned}$$

Show how you would modify the observability canonical realization to accomplish this, still using only n integrators.

Exercise 8.2 How would reachability canonical realization be modified if the linear differential equation that we started with was time varying rather than time invariant?

Exercise 8.3 Show how to modify the reachability canonical realization— but still using only n integrators — to obtain a realization of a *two-output* system of the form

$$\begin{aligned} y_1^{(n)} + a_{n-1}y_1^{(n-1)} + \cdots + a_0 y_1 &= c_{10}u + c_{11}\dot{u} + \cdots + c_{1,n-1}u^{(n-1)} \,, \\ y_2^{(n)} + a_{n-1}y_2^{(n-1)} + \cdots + a_0 y_2 &= c_{20}u + c_{21}\dot{u} + \cdots + c_{2,n-1}u^{(n-1)} . \end{aligned}$$

Exercise 8.4 Consider the two-input two-output system:

$$\begin{aligned} \dot{y}_1 &= y_1 + \alpha u_1 + u_2 \,, \\ \dot{y}_2 &= y_2 + u_1 + u_2 \end{aligned}$$

(a) Find a realization with the minimum number of states when $\alpha \neq 1$.

(b) Find a realization with the minimum number of states when $\alpha = 1$.

Chapter 10

Discrete-Time Linear State-Space Models

10.1 Introduction

In the previous chapters we showed how dynamic models arise, and studied some special characteristics that they may possess. We focused on state-space models and their properties, presenting several examples. In this chapter we will continue the study of state-space models, concentrating on solutions and properties of DT *linear* state-space models, both time-varying and time-invariant.

10.2 Time-Varying Linear Models

A general nth-order discrete-time linear state-space description takes the following form:

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \ , \end{aligned} \tag{10.1}$$

where $x(k) \in \mathbb{R}^n$. Given the initial condition $x(0)$ and the input sequence $u(k)$, we would like to find the state sequence or state *trajectory* $x(k)$ as well as the output sequence $y(k)$.

Undriven Response

First let us consider the *undriven response*, that is the response when $u(k) = 0$ for all $k \in \mathbb{Z}$. The state evolution equation then reduces to

$$x(k+1) = A(k)x(k) \ . \tag{10.2}$$

The response can be derived directly from (10.2) by simply iterating forward:

$$\begin{aligned} x(1) &= A(0)x(0) \\ x(2) &= A(1)x(1) \\ &= A(1)A(0)x(0) \\ x(k) &= A(k-1)A(k-2)\dots A(1)A(0)x(0) \end{aligned} \tag{10.3}$$

Motivated by (10.3), we define the **state transition matrix**, which relates the state of the undriven system at time k to the state at an earlier time ℓ:

$$x(k) = \Phi(k,\ell)x(\ell) \;\; k \geq \ell \; . \tag{10.4}$$

The form of the matrix follows directly from (10.3):

$$\Phi(k,\ell) = \begin{cases} A(k-1)A(k-2)\cdots A(\ell) & , \quad k > \ell \geq 0 \\ I & , \quad k = \ell \end{cases} \; . \tag{10.5}$$

If $A(k-1)$, $A(k-2)$,..., $A(\ell)$ are all invertible, then one could use the state transition matrix to obtain $x(k)$ from $x(\ell)$ even when $k < \ell$, but we shall typically assume $k \geq \ell$ when writing $\Phi(k,\ell)$.

The following properties of the discrete-time state transition matrix are worth highlighting:

$$\begin{aligned} \Phi(k,k) &= I \\ x(k) &= \Phi(k,0)x(0) \\ \Phi(k+1,\ell) &= A(k)\Phi(k,\ell). \end{aligned} \tag{10.6}$$

Example 10.1 (A Sufficient Condition for Asymptotic Stability)

The linear system (10.1) is termed *asymptotically stable* if, with $u(k) \equiv 0$, and for all $x(0)$, we have $x(n) \to 0$ (by which we mean $\|x(n)\| \to 0$) as $n \to \infty$. Since $u(k) \equiv 0$, we are in effect dealing with (10.2).

Suppose

$$\|A(k)\| \leq \gamma < 1 \tag{10.7}$$

for all k, where the norm is any submultiplicative norm and γ is a constant (independent of k) that is less than 1. Then

$$\|\Phi(n,0)\| \leq \gamma^n$$

and hence

$$\|x(n)\| \leq \gamma^n \|x(0)\|$$

so $x(n) \to 0$ as $n \to \infty$, no matter what $x(0)$ is. Hence (10.7) constitutes a sufficient condition (though a weak one, as we'll see) for asymptotic stability of (10.1).

Example 10.2 ("Lifting" a Periodic Model to an LTI Model)

Consider an undriven *linear, periodically varying* (LPV) model in state-space form. This is a system of the form (10.2) for which there is a smallest positive integer N such that $A(k+N) = A(k)$ for all k; thus N is the *period* of the system. (If $N = 1$, the system is actually LTI, so the cases of interest here are really those with $N \geq 2$.) Now focus on the state vector $x(mN)$ for integer m, i.e., the state of the LPV system sampled regularly once every period. Evidently

$$\begin{aligned} x(mN+N) &= \Big[A(N-1)A(N-2)\cdots A(0) \Big] x(mN) \\ &= \Phi(N,0)\, x(mN) \end{aligned} \tag{10.8}$$

The sampled state thus admits an LTI state-space model. The process of constructing this sampled model for an LPV system is referred to as *lifting*.

Driven Response

Now let us consider the driven system, *i.e.*, $u(k) \neq 0$ for at least some k. Referring back to (10.1), we have

$$\begin{aligned} x(1) &= A(0)x(0) + B(0)u(0) \\ x(2) &= A(1)x(1) + B(1)u(1) \\ &= A(1)A(0)x(0) + A(1)B(0)u(0) + B(1)u(1) \end{aligned} \tag{10.9}$$

which leads to

$$\begin{aligned} x(k) &= \Phi(k,0)x(0) + \sum_{\ell=0}^{k-1} \Phi(k,\ell+1)B(\ell)u(\ell) \\ &= \Phi(k,0)x(0) + \Gamma(k,0)\mathcal{U}(k,0)\ , \end{aligned} \tag{10.10}$$

where

$$\Gamma(k,0) = \Big[\Phi(k,1)B(0) \mid \Phi(k,2)B(1) \mid \cdots \mid B(k-1) \Big]\ ,\ \ \mathcal{U}(k,0) = \begin{pmatrix} u(0) \\ u(1) \\ \vdots \\ u(k-1) \end{pmatrix} \tag{10.11}$$

What (10.10) shows is that the solution of the system over k steps has the same form as the solution over one step, which is given in the first equation of (10.1). Also note that the system response is divided into two terms: one depends only on the initial state $x(0)$ and the other depends only on the input. These terms are respectively called the *natural* or *unforced* or *zero-input* response, and the *zero-state* response. Note also that the zero-state response has a form that is reminiscent of a convolution sum; this form is sometimes referred to as a *superposition sum*.

If (10.10) had been simply claimed as a solution, without any sort of derivation, then its validity could be verified by substituting it back into the system equations:

$$\begin{aligned} x(k+1) &= \Phi(k+1,0)x(0) + \sum_{\ell=0}^{k} \Phi(k+1,\ell+1)B(\ell)u(\ell) \\ &= \Phi(k+1,0)x(0) + \sum_{\ell=0}^{k-1} \Phi(k+1,\ell+1)B(\ell)u(\ell) + B(k)u(k) \\ &= A(k)\left[\Phi(k,0)x(0) + \sum_{\ell=0}^{k-1} \Phi(k,\ell+1)B(\ell)u(\ell)\right] + B(k)u(k) \\ &= A(k)x(k) + B(k)u(k)\ . \end{aligned} \tag{10.12}$$

Clearly, (10.12) satisfies the system equations (10.1). It remains to be verified that the proposed solution matches the initial state at $k = 0$. We have

$$x(0) = \Phi(0,0)x(0) = x(0)\ , \tag{10.13}$$

which completes the check.

If $\mathcal{Y}(k,0)$ is defined similarly to $\mathcal{U}(k,0)$, then following the sort of derivation that led to (10.10), we can establish that

$$\mathcal{Y}(k,0) = \Theta(k,0)x(0) + \Psi(k,0)\mathcal{U}(k,0) \tag{10.14}$$

for appropriately defined matrices $\Theta(k,0)$ and $\Psi(k,0)$. We leave you to work out the details. Once again, (10.14) for the output over k steps has the same form as the expression for the output at a single step, which is given in the second equation of (10.1).

10.3 Linear Time-Invariant Models

In the case of a *time-invariant* linear discrete-time system, the solutions can be simplified considerably. We first examine a direct time-domain solution, then compare this with a transform-domain solution, and finally return to the time domain, but in modal coordinates.

Direct Time-Domain Solution

For a linear time-invariant system, observe that

$$\left.\begin{aligned} A(k) &= A \\ B(k) &= B \end{aligned}\right\} \quad \text{for all} \quad k \geq 0, \tag{10.15}$$

where A and B are now constant matrices. Thus

$$\Phi(k,\ell) = A(k-1)\ldots A(\ell) = A^{k-\ell}\ , \quad k \geq \ell \tag{10.16}$$

so that, substituting this back into (10.10), we are left with

$$\begin{aligned} x(k) &= A^k x(0) + \sum_{\ell=0}^{k-1} A^{k-\ell-1} B u(\ell) \\ &= A^k x(0) + \left[A^{k-1}B \mid A^{k-2}B \mid \cdots \mid B \right] \begin{pmatrix} u(0) \\ u(1) \\ \vdots \\ u(k-1) \end{pmatrix} \end{aligned} \tag{10.17}$$

Note that the zero-state response in this case exactly corresponds to a convolution sum. Similar expressions can be worked out for the outputs, by simplifying (10.14); we leave the details to you.

Transform-Domain Solution

We know from earlier experience with dynamic linear time-invariant systems that the use of appropriate transform methods can reduce the solution of such a system to the solution of algebraic equations. This expectation does indeed hold up here. First recall the definition of the one-sided $\mathcal{Z}$-transform:

Definition 10.1 *The one-sided $\mathcal{Z}$-transform, $F(z)$, of the sequence $f(k)$ is given by*

$$F(z) = \sum_{k=0}^{\infty} z^{-k} f(k)$$

for all z such that the result of the summation is well defined, denoted by the Region of Convergence (ROC).

The sequence $f(k)$ can be a vector or matrix sequence, in which case $F(z)$ is respectively a vector or matrix as well.

It is easy to show that the transform of a sum of two sequences is the sum of the individual transforms. Also, scaling a sequence by a constant simply scales the transform by the same constant. The following shift property of the one-sided transform is critical, and not hard to establish. Suppose that $f(k) \stackrel{\mathcal{Z}}{\longleftrightarrow} F(z)$. Then

1.

$$g(k) = \begin{cases} f(k-1) & ; \quad k \geq 1 \\ 0 & ; \quad k = 0 \end{cases} \quad \Longrightarrow \quad G(z) = z^{-1} F(z).$$

2.

$$g(k) = f(k+1) \quad \Longrightarrow \quad G(z) = z\left[F(z) - f(0)\right].$$

Convolution is an important operation that can be defined on two sequences $f(k)$, $g(k)$ as

$$f * g(k) = \sum_{m=0}^{k} g(k-m)f(m),$$

whenever the dimensions of f and g are compatible so that the products are defined. The $\mathcal{Z}$ transform of a convolutions of two sequences satisfy

$$\begin{aligned}
\mathcal{Z}(f*g) &= \sum_{k=0}^{\infty} z^{-k} f*g(k) \\
&= \sum_{k=0}^{\infty} z^{-k} \left(\sum_{m=0}^{k} f(k-m)g(m) \right) \\
&= \sum_{m=0}^{\infty} \sum_{k=m}^{\infty} z^{-k} f(k-m)g(m) \\
&= \sum_{m=0}^{\infty} \sum_{k=0}^{\infty} z^{-(k+m)} f(k)g(m) \\
&= \sum_{m=0}^{\infty} z^{-m} \left(\sum_{k=0}^{\infty} z^{-k} f(k) \right) g(m) \\
&= F(z)G(z).
\end{aligned}$$

Now, given the state-space model (10.1), we can take transforms on both sides of the equations there. Using the transform properties just described, we get

$$zX(z) - zx(0) = AX(z) + BU(z) \tag{10.18}$$
$$Y(z) = CX(z) + DU(z). \tag{10.19}$$

This is solved to yield

$$\begin{aligned}
X(z) &= z(zI-A)^{-1}x(0) + (zI-A)^{-1}BU(z) \\
Y(z) &= zC(zI-A)^{-1}x(0) + \underbrace{\left[C(zI-A)^{-1}B + D\right]}_{\text{Transfer Function}} U(z)
\end{aligned} \tag{10.20}$$

To correlate the transform-domain solutions in the above expressions with the time-domain expressions in (10.10) and (10.14), it is helpful to note that

$$(zI-A)^{-1} = z^{-1}I + z^{-2}A + z^{-3}A^2 + \cdots \tag{10.21}$$

as may be verified by multiplying both sides by $(zI-A)$. The region of convergence for the series on the right is all values of z outside of some sufficiently large circle in the complex plane. What this series establishes, on comparison with the definition of the $\mathcal{Z}$-transform, is

that the inverse transform of $z(zI - A)^{-1}$ is the matrix sequence whose value at time k is A^k for $k \geq 0$; the sequence is 0 for time instants $k < 0$. That is we can write

$$\begin{aligned} \Big(I, A, A^2, A^3, A^4, \ldots\Big) &\overset{\mathcal{Z}}{\longleftrightarrow} z(zI - A)^{-1} \\ \Big(0, I, A, A^2, A^3, \ldots\Big) &\overset{\mathcal{Z}}{\longleftrightarrow} (zI - A)^{-1}. \end{aligned}$$

Also since the inverse transform of a product such as $(zI - A)^{-1}BU(z)$ is the convolution of the sequences whose transforms are $(zI - A)^{-1}B$ and $U(z)$ respectively, we get

$$\begin{aligned} \Big(x(0), Ax(0), A^2x(0), A^3x(0), \ldots\Big) &\overset{\mathcal{Z}}{\longleftrightarrow} z(zI - A)^{-1}x(0) \\ \Big(0, B, AB, A^2B, A^3B, \ldots\Big) * (u(0), u(1), u(2), u(3), \ldots) &\overset{\mathcal{Z}}{\longleftrightarrow} (zI - A)^{-1}BU(z). \end{aligned}$$

Putting the above two pieces together, the parallel between the time-domain expressions and the transform-domain expressions in (10.20) should be clear.

Exercises

Exercise 10.1 **(a)** Give an example of a nonzero matrix whose eigenvalues are all 0.

(b) Show that $A^k = 0$ for some *finite* positive power k if and only if all eigenvalues of A equal 0. Such a matrix is termed *nilpotent*. Argue that $A^n = 0$ for a nilpotent matrix of size n.

(c) If the sizes of the Jordan blocks of the nilpotent matrix A are $n_1 \leq n_2 \leq \ldots \leq n_q$, what is the smallest value of k for which $A^k = 0$?

(d) For an *arbitrary* square matrix A, what is the smallest value of k for which the range of A^{k+1} equals that of A^k? (Hint: Your answer can be stated in terms of the sizes of particular Jordan blocks of A.)

Exercise 10.2 Consider the periodically varying system in Problem 7.4. Find the general form of the solution.

Exercise 10.3 Gambler's Ruin

Consider gambling against a bank of capital A_1 in the following way: a coin is fliped, if the outcome is heads, the bank pays one dollar to the player, and if the outcome is tails, the player payes one dollar to the bank. Suppose the probability of a head is equal to p, the capital of the player is A_2, and the game continues until one party looses all of their capital. Calculate the probability of breaking the bank.

Chapter 11

Continuous-Time Linear State-Space Models

11.1 Introduction

In this chapter, we focus on the solution of CT state-space models. The development here follow the previous chapter.

11.2 The Time-Varying Case

Consider the nth-order continuous-time linear state-space description

$$\begin{aligned} \dot{x}(t) &= A(t)x(t) + B(t)u(t) \\ y(t) &= C(t)x(t) + D(t)u(t) \ . \end{aligned} \tag{11.1}$$

We shall always assume that the coefficient matrices in the above model are sufficiently well behaved for there to *exist* a *unique* solution to the state-space model for any specified initial condition $x(t_0)$ and any integrable input $u(t)$. For instance, if these coefficient matrices are piecewise continuous, with a finite number of discontinuities in any finite interval, then the desired existence and uniqueness properties hold.

We can describe the solution of (11.1) in terms of a matrix function $\Phi(t, \tau)$ that has the following two properties:

$$\dot{\Phi}(t, \tau) = A(t)\Phi(t, \tau) \ , \tag{11.2}$$

$$\Phi(\tau, \tau) = I \ . \tag{11.3}$$

This matrix function is referred to as the **state transition matrix**, and under our assumption on the nature of $A(t)$ it turns out that the state transition matrix *exists* and is *unique*.

We will show that, given $x(t_0)$ and $u(t)$,

$$x(t) = \Phi(t, t_0)x(t_0) + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau \ . \tag{11.4}$$

Observe again that, as in the DT case, the terms corresponding to the zero-input and zero-state responses are evident in (11.4). In order to verify (11.4), we differentiate it with respect to t:

$$\dot{x}(t) = \dot{\Phi}(t, t_0)x(t_0) + \int_{t_0}^{t} \dot{\Phi}(t, \tau)B(\tau)u(\tau)d\tau + \Phi(t, t)B(t)u(t) \ . \tag{11.5}$$

Using (11.2) and (11.3),

$$\dot{x}(t) = A(t)\Phi(t, t_0)x(t_0) + \int_{t_0}^{t} A(t)\Phi(t, \tau)B(\tau)u(\tau)d\tau + B(t)u(t) \ . \tag{11.6}$$

Now, since the integral is taken with respect to τ, $A(t)$ can be factored out:

$$\begin{aligned} \dot{x}(t) &= A(t)\left[\Phi(t, t_0)x(t_0) + \int_{t_0}^{t} \Phi(t, \tau)B(\tau)u(\tau)d\tau\right] + B(t)u(t) \ , & (11.7) \\ &= A(t)x(t) + B(t)u(t) \ , & (11.8) \end{aligned}$$

so the expression in (11.4) does indeed satisfy the state evolution equation. To verify that it also matches the specified initial condition, note that

$$x(t_0) = \Phi(t_0, t_0)x(t_0) = x(t_0). \tag{11.9}$$

We have now shown that the matrix function $\Phi(t, \tau)$ satisfying (11.2) and (11.3) yields the solution to the continuous-time system equation (11.1).

Exercise: Show that $\Phi(t, \tau)$ must be nonsingular. (Hint: Invoke our claim about uniqueness of solutions.)

The question that remains is how to find the state transition matrix. For a general linear time-varying system, there is no analytical expression that expresses $\Phi(t, \tau)$ analytically as a function of $A(t)$. Instead, we are essentially limited to numerical solution of the equation (11.2) with the boundary condition (11.3). This equation may be solved one column at a time, as follows. We numerically compute the respective solutions $x^i(t)$ of the homogeneous equation

$$\dot{x}(t) = A(t)x(t) \tag{11.10}$$

for each of the n initial conditions below:

$$x^1(t_0) = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad x^2(t_0) = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \ldots, \quad x^n(t_0) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \ .$$

Then

$$\Phi(t, t_0) = \left[\begin{array}{ccc} x^1(t) & \ldots & x^n(t) \end{array} \right]. \tag{11.11}$$

In summary, knowing n solutions of the homogeneous system for n independent initial conditions, we are able to construct the general solution of this linear time varying system. The underlying reason this construction works is that solutions of a linear system may be superposed, and our system is of order n.

Example 11.1 A Special Case

Consider the following time-varying system

$$\frac{d}{dt} \left[\begin{array}{c} x_1(t) \\ x_2(t) \end{array} \right] = \left[\begin{array}{cc} \alpha(t) & \beta(t) \\ -\beta(t) & \alpha(t) \end{array} \right] \left[\begin{array}{c} x_1(t) \\ x_2(t) \end{array} \right],$$

where $\alpha(t)$ and $\beta(t)$ are continuous functions of t. It turns out that the special structure of the matrix $A(t)$ here permits an analytical solution. Specifically, verify that the state transition matrix of the system is

$$\Phi(t, t_0) = \left[\begin{array}{cc} \exp(\int_{t_0}^t \alpha(\tau)d\tau) \cos(\int_{t_0}^t \beta(\tau)d\tau) & \exp(\int_{t_0}^t \alpha(\tau)d\tau) \sin(\int_{t_0}^t \beta(\tau)d\tau) \\ -\exp(\int_{t_0}^t \alpha(\tau)d\tau) \sin(\int_{t_0}^t \beta(\tau)d\tau) & \exp(\int_{t_0}^t \alpha(\tau)d\tau) \cos(\int_{t_0}^t \beta(\tau)d\tau) \end{array} \right]$$

The secret to solving the above system — or equivalently, to obtaining its state transition matrix — is to transform it to polar co-ordinates via the definitions

$$\begin{aligned} r^2(t) &= (x_1)^2(t) + (x_2)^2(t) \\ \theta(t) &= \tan^{-1}\left(\frac{x_2}{x_1}\right). \end{aligned}$$

We leave you to deduce now that

$$\begin{aligned} \frac{d}{dt} r^2 &= 2\alpha r^2 \\ \frac{d}{dt}\theta &= -\beta. \end{aligned}$$

The solution of this system of equations is then given by

$$r^2(t) = \exp\left(2\int_{t_0}^t \alpha(\tau)d\tau\right) r^2(t_0)$$

and

$$\theta(t) = \theta(t_0) - \int_{t_0}^t \beta(\tau)d\tau$$

Further Properties of the State Transition Matrix

The first property that we present involves the composition of the state transition matrix evaluated over different intervals. Suppose that at an arbitrary time t_0 the state vector is $x(t_0) = x_0$, with x_0 being an arbitrary vector. In the absence of an input the state vector at time t is given by $x(t) = \Phi(t, t_0)x_0$. At any other time t_1, the state vector is given by $x(t_1) = \Phi(t_1, t_0)x_0$. We can also write

$$\begin{aligned} x(t) &= \Phi(t, t_1)x(t_1) = \Phi(t, t_1)\Phi(t_1, t_0)x_0 \\ &= \Phi(t, t_0)x_0. \end{aligned}$$

Since x_0 is arbitrary, it follows that

$$\Phi(t, t_1)\Phi(t_1, t_0) = \Phi(t, t_0)$$

for any t_0 and t_1. (Note that since the state transition matrix in CT is alway invertible, there is no restriction that t_1 lie between t_0 and t — unlike in the DT case, where the state transition matrix may not be invertible).

Another property of interest (but one whose derivation can be safely skipped on a first reading) involves the determinant of the state transition matrix. We will now show that

$$\det(\Phi(t, t_0)) = \exp\Big(\int_{t_0}^{t} \text{trace}[A(\tau)]d\tau\Big) \ , \tag{11.12}$$

a result known as the *Jacobi-Liouville* formula. Before we derive this important formula, we need the following fact from matrix theory. For an $n \times n$ matrix M and a real parameter ϵ, we have

$$\det(I + \epsilon M) = 1 + \epsilon \,\text{trace}\,(M) + O(\epsilon^2) \ ,$$

where $O(\epsilon^2)$ denotes the terms of order greater than or equal to ϵ^2. In order to verify this fact, let U be a similarity transformation that brings M to an upper triangular matrix T, so $M = U^{-1}TU$. Such a U can always be found, in many ways. (One way, for a diagonalizable matrix, is to pick U to be the modal matrix of M, in which case T is actually diagonal; there is a natural extension of this approach in the non-diagonalizable case.) Then the eigenvalues $\{\lambda_i\}$ of M and T are identical, because similarity transformations do not change eigenvalues, and these numbers are precisely the diagonal elements of T. Hence

$$\begin{aligned} \det(I + \epsilon M) &= \det(I + \epsilon T) \\ &= \Pi_{i=1}^{n}\,(1 + \epsilon\lambda_i) \\ &= 1 + \epsilon \,\text{trace}\,(M) + O(\epsilon^2). \end{aligned}$$

Returning to the proof of (11.12), first observe that

$$\begin{aligned} \Phi(t + \epsilon, t_0) &= \Phi(t, t_0) + \epsilon\frac{d}{dt}\Phi(t, t_0) + O(\epsilon^2) \\ &= \Phi(t, t_0) + \epsilon A(t)\Phi(t, t_0) + O(\epsilon^2). \end{aligned}$$

The derivative of the determinant of $\Phi(t, t_0)$ is given by

$$\begin{aligned}
\frac{d}{dt}\det[\Phi(t,t_0)] &= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\det[\Phi(t+\epsilon,t_0)] - \det[\Phi(t,t_0)]\right) \\
&= \lim_{\epsilon\to 0}\frac{1}{\epsilon}\Big(\det[\Phi(t,t_0) + \epsilon A(t)\Phi(t,t_0)] - \det[\Phi(t,t_0)]\Big) \\
&= \det(\Phi(t,t_0))\lim_{\epsilon\to 0}\frac{1}{\epsilon}\left(\det[I+\epsilon A(t)] - 1\right) \\
&= \operatorname{trace}[A(t)]\det[\Phi(t,t_0)].
\end{aligned}$$

Integrating the above equation yields the desired result, (11.12).

11.3 The LTI Case

For linear time-invariant systems in continuous time, it is possible to give an explicit formula for the state transition matrix, $\Phi(t,\tau)$. In this case $A(t) = A$, a constant matrix. Let us *define* the **matrix exponential** of A by an infinite series of the same form that is (or may be) used to define the scalar exponential:

$$\begin{aligned}
e^{(t-t_0)A} &= I + (t-t_0)A + \frac{1}{2!}(t-t_0)^2A^2 + \dots \\
&= \sum_{k=0}^{\infty}\frac{1}{k!}(t-t_0)^kA^k. \qquad (11.13)
\end{aligned}$$

It turns out that this series is as nicely behaved as in the scalar case: it converges absolutely for all $A \in \mathbb{R}^{n\times n}$ and for all $t \in \mathbb{R}$, and it can be differentiated or integrated term by term. There exist methods for computing it, although the task is fraught with numerical difficulties.

With the above definition, it is easy to verify that the matrix exponential satisfies the defining conditions (11.2) and (11.3) for the state transition matrix. The solution of (11.1) in the LTI case is therefore given by

$$x(t) = e^{(t-t_0)A}x(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau. \qquad (11.14)$$

After determining $x(t)$, the system output can be obtained by

$$y(t) = Cx(t) + Du(t). \qquad (11.15)$$

Transform-Domain Solution of LTI Models

We can now parallel our transform-domain treatment of the DT case, except that now we use the one-sided Laplace transform instead of the $\mathcal{Z}$-transform:

Definition 11.1 *The one-sided Laplace transform, $F(s)$, of the signal $f(t)$ is given by*

$$F(s) = \int_{t=0-}^{\infty} e^{-st} f(t)\, dt$$

for all s where the integral is defined, denoted by the region of convergence (R.O.C.).

The various properties of the Laplace transform follow. The shift property of $\mathcal{Z}$ transforms that we used in the DT case is replaced by the following differentiation property: Suppose that $f(t) \stackrel{\mathcal{L}}{\longleftrightarrow} F(s)$. Then

$$g(t) = \frac{df(t)}{dt} \implies G(s) = sF(s) - f(0-)$$

Now, given the state-space model (11.1) in the LTI case, we can take transforms on both sides of the equations there. Using the transform property just described, we get

$$\begin{aligned} sX(s) - x(0-) &= AX(s) + BU(s) && (11.16)\\ Y(s) &= CX(s) + DU(s). && (11.17) \end{aligned}$$

This is solved to yield

$$\begin{aligned} X(s) &= (sI - A)^{-1}x(0-) + (sI - A)^{-1}BU(s)\\ Y(s) &= C(sI - A)^{-1}x(0-) + \underbrace{\left[C(sI - A)^{-1}B + D\right]}_{\text{Transfer Function}} U(s) && (11.18) \end{aligned}$$

which is very similar to the DT case.

An important fact that emerges on comparing (11.18) with its time-domain version (11.14) is that

$$\mathcal{L}\ \ e^{At} = (sI - A)^{-1}.$$

Therefore one way to compute the state transition matrix (a good way for small examples!) is by evaluating the entry-by-entry inverse transform of $(sI - A)^{-1}$.

Example 11.2 Find the state transition matrix associated with the (non-diagonalizable!) matrix

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

Using the above formula,

$$\begin{aligned} \mathcal{L}\ e^{At} &= (sI - A)^{-1} = \begin{bmatrix} s-1 & -2 \\ 0 & s-1 \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \frac{1}{s-1} & \frac{2}{(s-1)^2} \\ 0 & \frac{1}{s-1} \end{bmatrix}. \end{aligned}$$

By taking the inverse Laplace transform of the above matrix we get

$$e^{At} = \begin{bmatrix} e^t & 2te^t \\ 0 & e^t \end{bmatrix}.$$

Exercises

Exercise 11.1 Companion Matrices

(a) The following two matrices and their transposes are said to be *companion matrices* of the polynomial $q(z) = z^n + q_{n-1}z^{n-1} + \ldots + q_0$. Determine the characteristic polynomials of these four matrices, and hence explain the origin of the name. (Hint: First find explanations for why all four matrices must have the same characteristic polynomial, then determine the characteristic polynomial of any one of them.)

$$A_1 = \begin{pmatrix} -q_{n-1} & 1 & 0 & \ldots & 0 \\ -q_{n-2} & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -q_1 & 0 & 0 & \ldots & 1 \\ -q_0 & 0 & 0 & \ldots & 0 \end{pmatrix} \qquad A_2 = \begin{pmatrix} 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & 1 \\ -q_0 & -q_1 & -q_2 & \ldots & -q_{n-1} \end{pmatrix}$$

(b) Show that the matrix A_2 above has only one (right) eigenvector for each distinct eigenvalue λ_i, and that this eigenvector is of the form $[1 \;\; \lambda_i \;\; \lambda_i^2 \;\; \ldots \;\; \lambda_i^{n-1}]^T$.

(c) If

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 6 & 5 & -2 \end{pmatrix}$$

what are A^k and e^{At}? (Your answers may be left as a product of three — or fewer — matrices; do not bother to multiply them out.)

Exercise 11.2 Suppose you are given the state-space equation

$$\dot{x}(t) = Ax(t) + Bu(t)$$

with an input $u(t)$ that is piecewise constant over intervals of length T:

$$u(t) = u[k] \;, \qquad kT < t \leq (k+1)T$$

(a) Show that the sampled state $x[k] = x(kT)$ is governed by a *sampled-data state-space model* of the form

$$x[k+1] = Fx[k] + Gu[k]$$

for constant matrices F and G (i.e. matrices that do not depend on t or k), and determine these matrices in terms of A and B. (Hint: The result will involve the matrix exponential, e^{At}.) How are the eigenvalues and eigenvectors of F related to those of A?

(b) Compute F and G in the above discrete-time sampled-data model when

$$A = \begin{pmatrix} 0 & 1 \\ -\omega_0^2 & 0 \end{pmatrix} , \qquad B = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

(c) Suppose we implement a *state feedback control law* of the form $u[k] = Hx[k]$, where H is a gain matrix. What choice of H will cause the state of the resulting closed-loop system, $x[k+1] = (F + GH)x[k]$, to go to 0 in at most two steps, from any initial condition (H is then said to produce "deadbeat" behavior)? To simplify the notation for your calculations, denote $\cos\omega_0 T$ by c and $\sin\omega_0 T$ by s. Assume now that $\omega_0 T = \pi/6$, and *check your result* by substituting in your computed H and seeing if it does what you intended.

(d) For $\omega_0 T = \pi/6$ and $\omega_0 = 1$, your matrices from (b) should work out to be

$$F = \begin{pmatrix} \sqrt{3}/2 & 1/2 \\ -1/2 & \sqrt{3}/2 \end{pmatrix} , \quad G = \begin{pmatrix} 1 - (\sqrt{3}/2) \\ 1/2 \end{pmatrix}$$

Use Matlab to compute and plot the response of each of the state variables from $k = 0$ to $k = 10$, assuming $x[0] = [4\,,\,0]^T$ and with the following choices for $u[k]$:

- (i) the open-loop system, with $u[k] = 0$;
- (ii) the closed-loop system with $u[k] = Hx[k]$, where H is the feedback gain you computed in (c), with $\omega_0 = 1$; also plot $u[k]$ in this case.

(e) Now suppose the controller is computer-based. The above control law $u[k] = Hx[k]$ is implementable if the time taken to compute $Hx[k]$ is negligible compared to T. Often, however, it takes a considerable fraction of the sampling interval to do this computation, so the control that is applied to the system at time k is forced to use the state measurement at the previous instant. Suppose therefore that $u[k] = Hx[k-1]$. Find a state-space model for the closed-loop system in this case, written in terms of F, G, and H. (Hint: The computer-based controller now has memory!) What are the eigenvalues of the closed-loop system now, with H as in (c)? Again use Matlab to plot the response of the system to the same initial condition as in (d), and compare with the results in (d)(ii). Is there another choice of H that could yield deadbeat behavior? If so, find it; if not, suggest how to modify the control law to obtain deadbeat behavior.

Exercise 11.3 Given the matrix

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

show that

$$\exp\left(t \begin{bmatrix} \sigma & \omega \\ -\omega & \sigma \end{bmatrix}\right) = \begin{bmatrix} e^{\sigma t}\cos(\omega t) & e^{\sigma t}\sin(\omega t) \\ -e^{\sigma t}\sin(\omega t) & e^{\sigma t}\cos(\omega t) \end{bmatrix}$$

Exercise 11.4 Suppose A and B are constant square matrices. Show that

$$\exp\left(t \begin{bmatrix} A & 0 \\ 0 & B \end{bmatrix}\right) = \begin{bmatrix} e^{tA} & 0 \\ 0 & e^{tB} \end{bmatrix}.$$

Exercise 11.5 Suppose A and B are constant square matrices. Show that the solution of the following system of differential equations,

$$\dot{x}(t) = e^{-tA} B e^{tA} x(t) \ ,$$

is given by

$$x(t) = e^{-tA} e^{(t-t_0)(A+B)} e^{t_0 A} x(t_0) \ .$$

Exercise 11.6 Suppose A is a constant square matrix, and $f(t)$ is a continuous scalar function of t. Show that the state transition matrix for the system

$$\dot{x}(t) = f(t) A \, x(t)$$

is given by

$$\Phi(t, t_0) = \exp\left((\int_{t_0}^{t} f(\tau) d\tau) A \right) .$$

Exercise 11.7 (*Floquet Theory*). Consider the system

$$\dot{x}(t) = A(t) x(t)$$

where $A(t)$ is a periodic matrix with period T, so $A(t+T) = A(t)$. We want to study the state transition matrix $\Phi(t, t_0)$ associated with this periodically time-varying system.

1. First let us start with the state transition matrix $\Phi(t, 0)$, which satisfies

$$\begin{aligned} \dot{\Phi} &= A(t)\Phi \\ \Phi(0,0) &= I. \end{aligned}$$

   Define the matrix $\Psi(t, 0) = \Phi(t+T, 0)$ and show that Ψ satisfies

$$\begin{aligned} \dot{\Psi}(t,0) &= A(t)\Psi(t,0) \\ \Psi(0,0) &= \Phi(T,0). \end{aligned}$$

2. Show that this implies that $\Phi(t+T, 0) = \Phi(t, 0)\Phi(T, 0)$.
3. Using Jacobi-Liouville formula, show that $\Phi(T, 0)$ is invertible and therefore can be written as $\Phi(T, 0) = e^{TR}$.
4. Define

$$P(t)^{-1} = \Phi(t, 0) e^{-tR},$$

   and show that $P(t)^{-1}$, and consequently $P(t)$, are periodic with period T. Also show that $P(T) = I$. This means that

$$\Phi(t, 0) = P(t)^{-1} e^{tR}.$$

5. Show that $\Phi(0, t_0) = \Phi^{-1}(t_0, 0)$. Using the fact that $\Phi(t, t_0) = \Phi(t, 0)\Phi(0, t_0)$, show that

$$\Phi(t, t_0) = P(t)^{-1} e^{(t-t_0)R} P(t_0).$$

   What is the significance of this result?

Chapter 12

Modal Decomposition of State-Space Models

12.1 Introduction

The solutions obtained in previous chapters, whether in time domain or transform domain, can be further decomposed to give a geometric understanding of the solution. The modal decomposition expresses the state equation as a linear combination of the various *modes* of the system and shows precisely how the initial conditions as well as the inputs impact these modes.

12.2 The Transfer Function Matrix

It is evident from (10.20) that the *transfer function matrix* for the system, which relates the input transform to the output transform when the initial condition is zero, is given by

$$H(z) = C(zI - A)^{-1}B + D. \tag{12.1}$$

For a multi-input, multi-output (MIMO) system with m inputs and p outputs, this results in a $p \times m$ matrix of rational functions of z. In order to get an idea of the nature of these rational functions, we express the matrix inverse as the adjoint matrix divided by the determinant, as follows:

$$H(z) = \frac{1}{\det(zI - A)}\, C\left[\text{adj}(zI - A)\right] B \quad + \quad D.$$

The determinant $\det(zI - A)$ in the denominator is an n^{th}-order monic (*i.e.* coefficient of z^n is 1) polynomial in z, known as the *characteristic polynomial* of A and denoted by $a(z)$. The

entries of the adjoint matrix (the cofactors) are computed from minors of $(zI - A)$, which are polynomials of degree less than n. Hence the entries of the matrices

$$(zI - A)^{-1} = \frac{1}{\det(zI - A)}\text{adj}(zI - A)$$

and

$$H(z) - D = \frac{1}{\det(zI - A)}C\text{adj}(zI - A)B$$

are strictly proper, *i.e.* have numerator degree strictly less than their denominator degree. With the D term added in, $H(z)$ becomes proper that is all entries have numerator degree less than or equal to the degree of the denominator. For $|z| \nearrow \infty$, $H(z) \to D$.

The polynomial $a(z)$ forms the denominators of all the entries of $(zI - A)^{-1}$ and $H(z)$, except that in some, or even all, of the entries there may be cancellations of common factors that occur between $a(z)$ and the respective numerators. We shall have a lot more to say later about these cancellations and their relation to the concepts of reachability (or controllability) and observability. To compute the inverse transform of $(zI - A)^{-1}$ (which is the sequence A^{k-1}) and the inverse transform of $H(z)$ (which is a matrix sequence whose components are the zero-state unit sample responses from each input to each output), we need to find the inverse transform of rationals whose denominator is $a(z)$ (apart from any cancellations). The roots of $a(z)$ — also termed the *characteristic roots* or *natural frequencies* of the system, thus play a critical role in determining the nature of the solution. A fuller picture will emerge as we proceed.

Multivariable Poles and Zeros

You are familiar with the definitions of poles, zeros, and their multiplicities for the scalar transfer functions associated with single-input, single-output (SISO) LTI systems. For the case of the $p \times m$ transfer function *matrix* $H(z)$ that describes the zero-state input/output behavior of an m-input, p-output LTI system, the definitions of poles and zeros are more subtle. We include some preliminary discussion here, but will leave further elaboration for later in the course.

It is clear what we would want our eventual definitions of MIMO poles and zeros to specialize to in the case where $H(z)$ is nonzero only in its *diagonal* positions, because this corresponds to completely decoupled scalar transfer functions. For this diagonal case, we would evidently like to say that the poles of $H(z)$ are the poles of the individual diagonal entries of $H(z)$, and similarly for the zeros. For example, given

$$H(z) = \text{diagonal}\left(\frac{z+2}{(z+0.5)^2}\ ,\ \frac{z}{(z+2)(z+0.5)}\right)$$

we would say that $H(z)$ has poles of multiplicity 2 and 1 at $z = -0.5$, and a pole of multiplicity 1 at $z = -2$; and that it has zeros of multiplicity 1 at -2, at $z = 0$, and at $z = \infty$. Note that

in the MIMO case we can have poles and zeros at the same frequency (e.g. those at -2 in the above example), without any cancellation! Also note that a pole or zero is not necessarily characterized by a single multiplicity; we may instead have a set of multiplicity indices (e.g. as needed to describe the pole at -0.5 in the above example). The diagonal case makes clear that we do *not* want to define a pole or zero location of $H(z)$ in the general case to be a frequency where *all* entries of $H(z)$ respectively have poles or zeros.

For a variety of reasons, the appropriate definition of a pole location is as follows:

- **Pole Location:** $H(z)$ has a pole at a frequency p_0 if *some* entry of $H(z)$ has a pole at $z = p_0$.

The full definition (which we will present later in the course) also shows us how to determine the set of multiplicities associated with each pole frequency. Similarly, it turns out that the appropriate definition of a zero location is as follows:

- **Zero Location:** $H(z)$ has a zero at a frequency η_0 if the *rank* of $H(z)$ *drops* at $z = \eta_0$.

Again, the full definition also permits us to determine the set of multiplicities associated with each zero frequency. The determination of whether or not the rank of $H(z)$ drops at some value of z is complicated by the fact that $H(z)$ may also have a pole at that value of z; however, all of this can be sorted out very nicely.

12.3 Similarity Transformations

Suppose we have characterized a given dynamic system via a particular state-space representation, say with state variables $x_1, x_2, \cdots, x_n$. The evolution of the system then corresponds to a trajectory of points in the state space, described by the succession of values taken by the state variables. In other words, the state variables may be seen as constituting the *coordinates* in terms of which we have chosen to describe the motion in the state space.

We are free, of course, to choose alternative coordinate bases — i.e., alternative state variables — to describe the evolution of the system. This evolution is not changed by the choice of coordinates; only the *description* of the evolution changes its form. For instance, in the LTI circuit example in the previous chapter, we could have used $i_L - v_C$ and $i_L + v_C$ instead of i_L and v_C. The information in one set is identical with that in the other, and the existence of a state-space description with one set implies the existence of a state-space description with the other, as we now show more concretely and more generally. The flexibility to choose an appropriate coordinate system can be very valuable, and we will find ourselves invoking such coordinate changes very often.

Given that we have a state vector x, suppose we define a constant invertible linear mapping from x to r, as follows:

$$r = T^{-1}x \quad , \quad x = Tr. \tag{12.2}$$

Since T is invertible, this maps each trajectory $x(k)$ to a unique trajectory $r(k)$, and vice versa. We refer to such a transformation as a *similarity transformation*. The matrix T embodies

the details of the transformation from x coordinates to r coordinates — it is easy to see from (12.2) that the columns of T are the representations of the standard unit vectors of r in the coordinate system of x, which is all that is needed to completely define the new coordinate system.

Substituting for $x(k)$ in the standard (LTI version of the) state-space model (10.1), we have

$$T\,r(k+1) = A\left(T\,r(k)\right) + Bu(k) \tag{12.3}$$

$$y(k) = C\left(T\,r(k)\right) + Du(k). \tag{12.4}$$

or

$$r(k+1) = (T^{-1}AT)\,r(k) + (T^{-1}B)\,u(k) \tag{12.5}$$

$$= \widehat{A}\,r(k) + \widehat{B}\,u(k) \tag{12.6}$$

$$y(k) = (CT)\,r(k) + D\,u(k) \tag{12.7}$$

$$= \widehat{C}\,r(k) + D\,u(k) \tag{12.8}$$

We now have a new representation of the system dynamics; it is said to be *similar* to the original representation. It is critical to understand, however, that the dynamic properties of the model are not at all affected by this coordinate change in the state space. In particular, the mapping from $u(k)$ to $y(k)$, *i.e.* the input/output map, is unchanged by a similarity transformation.

12.4 Solution in Modal Coordinates

The proper choice of a similarity transformation may yield a new system model that will be more suitable for analytical purposes. One such transformation brings the system to what are known as *modal coordinates*. We shall describe this transformation now for the case where the matrix A in the state-space model can be *diagonalized*, in a sense to be defined below; we leave the general case for later.

Modal coordinates are built around the *eigenvectors* of A. To get a sense for why the eigenvectors may be involved in obtaining a simple choice of coordinates for studying the dynamics of the system, let us examine the possibility of finding a solution of the form

$$x(k) = \lambda^k v\,, \quad v \neq 0 \tag{12.9}$$

for the undriven LTI system

$$x(k+1) = Ax(k) \tag{12.10}$$

Substituting (12.9) in (12.10), we find the requisite condition to be that

$$(\lambda I - A)\,v = 0 \tag{12.11}$$

i.e., that λ be an *eigenvalue* of A, and v an associated eigenvector. Note from (12.11) that multiplying any eigenvector by a nonzero scalar again yields an eigenvector, so eigenvectors are only defined up to a nonzero scaling; any convenient scaling or normalization can be used. In other words, (12.9) is a solution of the undriven system iff λ is one of the n roots λ_i of the *characteristic polynomial*

$$a(z) = \det(zI - A) = z^n + a_{n-1}z^{n-1} + \cdots + a_0 \tag{12.12}$$

and v is a corresponding eigenvector v_i. A solution of the form $x(k) = \lambda_i^k v_i$ is referred to as a *mode* of the system, in this case the ith mode. The corresponding λ_i is the ith *modal frequency* or *natural frequency*, and v_i is the corresponding *modal shape*. Note that we can excite just the ith mode by ensuring that the initial condition is $x(0) = \lambda_i^0 v_i = v_i$. The ensuing motion is then confined to the direction of v_i, with a scaling by λ_i at each step.

It can be shown fairly easily that eigenvectors associated with *distinct* eigenvalues are (linearly) *independent*, i.e. none of them can be written as a weighted linear combination of the remaining ones. Thus, *if* the n eigenvalues of A are distinct, then the n corresponding eigenvectors v_i are independent, and can actually form a *basis* for the state-space. Distinct eigenvalues are not necessary, however, to ensure that there exists a selection of n independent eigenvectors. In any case, we shall restrict ourselves for now to the case where — because of distinct eigenvalues or otherwise — the matrix A has n independent eigenvectors. Such an A is termed *diagonalizable* (for a reason that will become evident shortly), or *non-defective*. There do exist matrices that are *not* diagonalizable, as we shall see when we examine the Jordan form in detail later in this course.

Because (12.10) is linear, a weighted linear combination of modal solutions will satisfy it too, so

$$x(k) = \sum_{i=1}^{n} \alpha_i v_i \lambda_i^k \tag{12.13}$$

will be a solution of (12.10) for arbitrary weights α_i, with initial condition

$$x(0) = \sum_{i=1}^{n} \alpha_i v_i \tag{12.14}$$

Since the n eigenvectors v_i are independent under our assumption of diagonalizable A, the right side of (12.14) can be made equal to *any* desired $x(0)$ by proper choice of the coefficients α_i, and these coefficients are *unique*. Hence specifying the initial condition of the undriven system (12.10) specifies the α_i via (12.14) and thus, via (12.13), specifies the response of the undriven system. We refer to the expression in (12.13) as the *modal decomposition* of the undriven response. Note that the contribution to the modal decomposition from a conjugate pair of eigenvalues λ and λ^* will be a *real* term of the form $\alpha v \lambda^k + \alpha^* v^* \lambda^{*k}$.

From (12.14), it follows that $\alpha = V^{-1}x(0)$, where α is a vector with components α_i. Let $W = V^{-1}$, and w_i' be the i^{th} row of W, then

$$x(k) = \sum_{i=1}^{n} \lambda_i^k v_i w_i' x(0) \tag{12.15}$$

It easy to see that w_i is a left eigenvector corresponding to the eigenvalue λ_i. The above modal decomposition of the undriven system is the same as obtaining the *diadic* form of A^k. The contribution of $x(0)$ to the i^{th} mode is captured in the term $w_i' x(0)$.

Before proceeding to examine the full response of a linear time-invariant model in modal terms, it is worth noting that the preceding results already allow us to obtain a precise condition for **asymptotic stability** of the system, at least in the case of diagonalizable A (it turns out that the condition below is the right one even for the general case). Recalling the definition in Example 10.1, we see immediately from the modal decomposition that the LTI system (12.10) is asymptotically stable iff $|\lambda_i| < 1$ for all $1 \leq i \leq n$, i.e. iff all the natural frequencies of the system are within the unit circle. Since it is certainly possible to have this condition hold even when $\|A\|$ is arbitrarily greater than 1, we see that the sufficient condition given in Example 1 is indeed rather weak, at least for the time-invariant case.

Let us turn now to the LTI version of the full system in (10.1). Rather than approaching its modal solution in the same style as was done for the undriven case, we shall (for a different point of view) approach it via a similarity transformation to modal coordinates, i.e., to coordinates defined by the eigenvectors $\{v_i\}$ of the system. Consider using the similarity transformation

$$x(k) = V\, r(k) \tag{12.16}$$

where the ith column of the $n \times n$ matrix V is the ith eigenvector, v_i:

$$V = \begin{pmatrix} v_1 & v_2 & & v_n \end{pmatrix} \tag{12.17}$$

We refer to V as the *modal matrix*. Under our assumption of diagonalizable A, the eigenvectors are independent, so V is guaranteed to be invertible, and (12.16) therefore does indeed constitute a similarity transformation. We refer to this similarity transformation as a *modal transformation*, and the variables $r_i(k)$ defined through (12.16) are termed *modal variables* or *modal coordinates*. What makes this transformation interesting and useful is the fact that the state evolution matrix A now transforms to a *diagonal* matrix Λ:

$$V^{-1}AV = \text{diagonal}\,\{\lambda_1, \quad , \lambda_n\} = \begin{bmatrix} \lambda_1 & 0 & & 0 \\ 0 & \lambda_2 & & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_n \end{bmatrix} = \Lambda \tag{12.18}$$

The easiest way to verify this is to establish the equivalent condition that $AV = V\Lambda$, which in turn is simply the equation (12.11), written for $i = 1, \quad , n$ and stacked up in matrix form. The reason for calling A "diagonalizable" when it has a full set of independent eigenvectors is now apparent.

Under this modal transformation, the *undriven* system is transformed into n *decoupled, scalar* equations:

$$r_i(k+1) = \lambda_i r_i(k) \tag{12.19}$$

for $i = 1, 2, \ldots, n$. Each of these is trivial to solve: we have $r_i(k) = \lambda_i^k \, r_i(0)$. Combining this with (12.16) yields (12.13) again, but with the additional insight that

$$\alpha_i = r_i(0) \tag{12.20}$$

Applying the modal transformation (12.16) to the full system, it is easy to see that the transformed system takes the following form, which is once again decoupled into n parallel *scalar* subsystems:

$$\begin{aligned} r_i(k+1) &= \lambda_i r_i(k) + \beta_i u(k) \ , \quad i = 1, 2, \ldots, n & (12.21) \\ y(k) &= \xi_1 r_1(k) + \quad + \xi_n r_n(k) + Du(k) & (12.22) \end{aligned}$$

where the β_i and ξ_i are defined via

$$V^{-1}B = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \quad , \quad CV = \begin{bmatrix} \xi_1 & \xi_2 & & \xi_n \end{bmatrix} \tag{12.23}$$

The scalar equations above can be solved explicitly by elementary methods (compare also with the expression in (22.2):

$$r_i(k) = \underbrace{\lambda_i^k \, r_i(0)}_{\text{ZIR}} + \underbrace{\sum_{0}^{k-1} \lambda_i^{k-\ell-1} \, \beta_i \, u(\ell)}_{\text{ZSR}} \tag{12.24}$$

where "ZIR" denotes the zero-input response, and "ZSR" the zero-state response. From the preceding expression, one can obtain an expression for $y(k)$. Also, substituting (12.24) in (12.16), we can derive a corresponding modal representation for the original state vector $x(k)$. We leave you to write out these details.

Finally, the same concepts hold for CT systems. We leave the details as an exercise.

Example 12.1

Consider the following system:

$$\begin{bmatrix} \dot{x_1} \\ \dot{x_2} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 8 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u \tag{12.25}$$

We will consider the modal decomposition of this system for the zero input response. The eigenvalues of A are -4 and 2 and the associated eigenvectors are $[\ 1 \quad -4\ ]'$ and $[\ 1 \quad 2\ ]'$. The modal matrix is constructed from the eigenvectors above:

$$V = \begin{pmatrix} 1 & 1 \\ -4 & 2 \end{pmatrix} \tag{12.26}$$

Its inverse is given by

$$W = V^{-1} = \frac{1}{6}\begin{bmatrix} 2 & -1 \\ 4 & 1 \end{bmatrix}.$$

It follows that:

$$WAV = \Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} -4 & 0 \\ 0 & 2 \end{bmatrix}.$$

Now let's define r in modal coordinate as

$$x(t) = Tr \rightarrow r(t) = T^{-1}x(t).$$

Then in terms of r, the original system can be transformed into the following:

$$\begin{bmatrix} \dot{r_1} \\ \dot{r_2} \end{bmatrix} = \begin{bmatrix} -4 & 0 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} r_1 \\ r_2 \end{bmatrix}. \tag{12.27}$$

The response of the system for a given initial state and zero input can now be expressed as:

$$\begin{aligned} x(t) &= Vr(t) = Ve^{\Lambda(t-t_0)}Wx(t_0) \\ &= \begin{bmatrix} 1 & 1 \\ -4 & 2 \end{bmatrix}\begin{bmatrix} e^{-4(t-t_0)} & 0 \\ 0 & e^{2(t-t_0)} \end{bmatrix}\frac{1}{6}\begin{bmatrix} 2 & -1 \\ 4 & 1 \end{bmatrix}x(t_0). \end{aligned}$$

For instance, if the initial vector is chosen in the direction of the first eigenvector, i.e., $x(t_0) = v_1 = \begin{bmatrix} 1 & -4 \end{bmatrix}'$ then the response is given by:

$$x(t) = \begin{bmatrix} 1 \\ -4 \end{bmatrix}e^{-4(t-t_0)}.$$

Example 12.2 Inverted Pendulum

Consider the linearized model of the inverted pendulum in Example 7.6 with the parameters given by: $m = 1$, $M = 10$, $l = 1$, and $g = 9.8$. The eigenvalues of the matrix A are 0, 0, 3.1424, and -3.1424. In this case, the eigenvalue at 0 is repeated, and hence the matrix A may not be diagonalizable. However, we can still construct the Jordan form of A by finding the generalized eigenvectors corresponding to 0, and the eigenvectors corresponding to the other eigenvalues. The Jordan form of A, $\Lambda = T^{-1}AT$ and the corresponding transformation T are given by:

$$\Lambda = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 3.1424 & 0 \\ 0 & 0 & 0 & -3.1424 \end{bmatrix}, T = \begin{bmatrix} 0.0909 & 0 & -0.0145 & 0.0145 \\ 0 & 0.0909 & -0.0455 & -0.0455 \\ 0 & 0 & 0.1591 & -0.1591 \\ 0 & 0 & 0.5000 & 0.5000 \end{bmatrix}$$

We can still get quite a bit of insight from this decomposition. Consider the zero input response, and let $x(0) = v_1 = [1 \quad 0 \quad 0 \quad 0 \quad]'$. This is an eigenvector corresponding to the zero eigenvalue, and corresponds to a fixed distance s, zero velocity, zero angular position, and zero angular velocity. In that case, the system remains in the same position and the response is equal to $x(0)$ for all future time. Now, let $x(0) = v_2 = [0 \quad 1 \quad 0 \quad 0 \quad]'$, which corresponds to a non-zero velocity and zero position, angle and angular velocity. This is not an eigenvector but rather a generalized eigenvector, i.e., it satisfies $Av_2 = v_1$. We can easily calculate the response to be $x(t) = [t \quad 1 \quad 0 \quad 0]$ implying that the cart will drift with constant velocity but will remain in the upright position. Notice that the response lies in the linear span of v_1 and v_2.

The case where $x_(0) = v_3$ corresponds to the eigenvalue $\lambda = 3.1424$. In this case, the cart is moving to the left while the pendulum is tilted to the right with clockwise angular velocity. Thus, the pendulum tilts more to the right, which corresponds to unstable behavior. The case where $x(0) = v_4$ corresponds to the eigenvalue $\lambda = -3.1424$. The cart again is moving to the left with clockwise angular velocity, but the pendulum is tilted to the left. With an appropriate combination of these variables (given by the eigenvector v_4) the response of the system converges to the upright equilibrium position at the origin.

Exercises

Exercise 12.1 Use the expression in (12.1) to find the transfer functions of the DT versions of the controller canonical form and the observer canonical form defined in Chapter 8. Verify that the transfer functions are consistent with what you would compute from the input-output difference equation on which the canonical forms are based.

Exercise 12.2 Let v and w' be the right and left eigenvectors associated with some *non-repeated* eigenvalue λ of a matrix A, with the normalization $w'v = 1$. Suppose A is perturbed infinitesimally to $A + dA$, so that λ is perturbed to $\lambda + d\lambda$, v to $v + dv$, and w' to $w' + dw'$. Show that $d\lambda = w'(dA)v$.

Chapter 13

Internal (Lyapunov) Stability

13.1 Introduction

We have already seen some examples of both stable and unstable systems. The objective of this chapter is to formalize the notion of internal stability for general nonlinear state-space models. Apart from defining the various notions of stability, we define an entity known as a *Lyapunov function* and relate it to these various stability notions.

13.2 Notions of Stability

For a general undriven system

$$\dot{x}(t) = f(x(t), 0, t) \qquad (CT) \tag{13.1}$$

$$x(k+1) = f(x(k), 0, k) \qquad (DT), \tag{13.2}$$

we say that a point $\overline{x}$ is an *equilibrium point* from time t_0 for the CT system above if $f(\overline{x}, 0, t) = 0,\ \forall t \geq t_0$, and is an equilibrium point from time k_0 for the DT system above if $f(\overline{x}, 0, k) = \bar{x},\ \forall k \geq k_0$. If the system is started in the state $\bar{x}$ at time t_0 or k_0, it will remain there for all time. Nonlinear systems can have multiple equilibrium points (or equilibria). (Another class of special solutions for nonlinear systems are *periodic* solutions, but we shall just focus on equilibria here.) We would like to characterize the *stability* of the equilibria in some fashion. For example, does the state tend to return to the equilibrium point after a small perturbation away from it? Does it remain close to the equilibrium point in some sense? Does it diverge?

The most fruitful notion of stability for an equilibrium point of a nonlinear system is given by the definition below. We shall assume that the equilibrium point of interest is at the origin, since if $\overline{x} \neq 0$, a simple translation can always be applied to obtain an equivalent system with the equilibrium at 0.

Definition 13.1 A system is called *asymptotically stable* around its equilibrium point at the origin if it satisfies the following two conditions:

1. Given any $\epsilon > 0$, $\exists \delta_1 > 0$ such that if $\|x(t_0)\| < \delta_1$, then $\|x(t)\| < \epsilon$, $\forall\, t > t_0$.
2. $\exists \delta_2 > 0$ such that if $\|x(t_0)\| < \delta_2$, then $x(t) \to 0$ as $t \to \infty$.

The first condition requires that the state trajectory can be confined to an arbitrarily small "ball" centered at the equilibrium point and of radius ϵ, when released from an *arbitrary* initial condition in a ball of sufficiently small (but positive) radius δ_1. This is called *stability in the sense of Lyapunov* (i.s.L.). It is possible to have stability in the sense of Lyapunov without having asymptotic stability, in which case we refer to the equilibrium point as *marginally stable*. Nonlinear systems also exist that satisfy the second requirement without being stable i.s.L., as the following example shows. An equilibrium point that is *not* stable i.s.L. is termed *unstable*.

Example 13.1 (Unstable Equilibrium Point That Attracts All Trajectories)

Consider the second-order system with state variables x_1 and x_2 whose dynamics are most easily described in polar coordinates via the equations

$$\begin{aligned} \dot{r} &= r(1-r) \\ \dot{\theta} &= \sin^2(\theta/2) \end{aligned} \tag{13.3}$$

where the radius r is given by $r = \sqrt{x_1^2 + x_2^2}$ and the angle θ by $0 \leq \theta = \arctan(x_2/x_1) < 2\pi$. (You might try obtaining a state-space description directly involving x_1 and x_2.) It is easy to see that there are precisely two equilibrium points: one at the origin, and the other at $r = 1$, $\theta = 0$. We leave you to verify with rough calculations (or computer simulation from various initial conditions) that the trajectories of the system have the form shown in the figure below.

Evidently all trajectories (except the trivial one that starts and stays at the origin) end up at $r = 1$, $\theta = 0$. However, this equilibrium point is not stable i.s.L., because these trajectories cannot be confined to an arbitrarily small ball around the equilibrium point when they are released from arbitrary points with any ball (no matter how small) around this equilibrium.

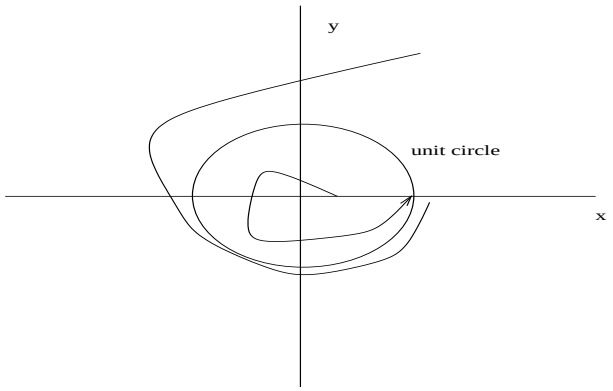


Figure 13.1: System Trajectories

13.3 Stability of Linear Systems

We may apply the preceding definitions to the LTI case by considering a system with a diagonalizable A matrix (in our standard notation) and $\mathbf{u} \equiv 0$. The unique equilibrium point is at $x = 0$, provided A has no eigenvalue at 0 (respectively 1) in the CT (respectively DT) case. (Otherwise every point in the entire eigenspace corresponding to this eigenvalue is an equilibrium.) Now

$$\begin{aligned} \dot{x}(t) &= e^{At}x(0) \\ &= V \begin{bmatrix} e^{\lambda_1 t} & & \\ & \ddots & \\ & & e^{\lambda_n t} \end{bmatrix} Wx(0) \qquad (CT) \end{aligned} \tag{13.4}$$

$$\begin{aligned} x(k) &= A^k x(0) \\ &= V \begin{bmatrix} \lambda_1^k & & \\ & \ddots & \\ & & \lambda_n^k \end{bmatrix} Wx(0) \qquad (DT) \end{aligned} \tag{13.5}$$

Hence, it is clear that in continuous time a system with a diagonalizable A is asymptotically stable iff

$$\mathcal{R}e(\lambda_i) < 0, \quad i \in \{1, \ldots, n\}, \tag{13.6}$$

while in discrete time the requirement is that

$$|\lambda_i| < 1 \quad i \in \{1, \ldots, n\}, \tag{13.7}$$

Note that if $\mathcal{R}e(\lambda_i) = 0$ (CT) or $|\lambda_i| = 1$ (DT), the system is not asymptotically stable, but is marginally stable.

Exercise: For the nondiagonalizable case, use your understanding of the Jordan form to show that the conditions for *asymptotic* stability are the *same* as in the diagonalizable case. For *marginal* stability, we require in the CT case that $\mathcal{R}e(\lambda_i) \leq 0$, with equality holding for at least one eigenvalue; furthermore, every eigenvalue whose real part equals 0 should have its geometric multiplicity equal to its algebraic multiplicity, i.e., all its associated Jordan blocks should be of size 1. (Verify that the presence of Jordan blocks of size greater than one for these imaginary-axis eigenvalues would lead to the state variables *growing polynomially* with time.) A similar condition holds for marginal stability in the DT case.

Stability of Linear Time-Varying Systems

Recall that the general unforced solution to a linear time-varying system is

$$x(t) = \Phi(t, t_0)x(t_0),$$

where $\Phi(t, \tau)$ is the state transition matrix. It follows that the system is

1. stable i.s.L. at $\overline{x} = 0$ if $\sup_t \|\Phi(t, t_0)\| = m(t_0) < \infty$.
2. asymptotically stable at $\overline{x} = 0$ if $\lim_{t\to\infty} \|\Phi(t, t_0)\| \to 0, \ \forall t_0$.

These conditions follow directly from Definition 13.1.

13.4 Lyapunov's Direct Method

General Idea

Consider the continuous-time system

$$\dot{x}(t) = f(x(t)) \tag{13.8}$$

with an equilibrium point at $x = 0$. This is a time-invariant (or "autonomous") system, since f does not depend explicitly on t. The stability analysis of the equilibrium point in such a system is a difficult task in general. This is due to the fact that we cannot write a simple formula relating the trajectory to the initial state. The idea behind Lyapunov's "direct" method is to establish properties of the equilibrium point (or, more generally, of the nonlinear system) by studying how certain carefully selected scalar functions of the state evolve as the system state evolves. (The term "direct" is to contrast this approach with Lyapunov's "indirect" method, which attempts to establish properties of the equilibrium point by studying the behavior of the *linearized* system at that point. We shall study this next Chapter.)

Consider, for instance, a continuous scalar function $V(x)$ that is 0 at the origin and positive elsewhere in some ball enclosing the origin, i.e. $V(0) = 0$ and $V(x) > 0$ for $x \neq 0$ in this ball. Such a $V(x)$ may be thought of as an "energy" function. Let $\dot{V}(x)$ denote the time derivative of $V(x)$ along any trajectory of the system, i.e. its rate of change as $x(t)$ varies

according to (13.8). If this derivative is negative throughout the region (except at the origin), then this implies that the energy is strictly decreasing over time. In this case, because the energy is lower bounded by 0, the energy must go to 0, which implies that all trajectories converge to the zero state. We will formalize this idea in the following sections.

Lyapunov Functions

Definition 13.2 Let V be a continuous map from $\mathbb{R}^n$ to $\mathbb{R}$. We call $V(x)$ a *locally positive definite* (lpd) function around $x = 0$ if

1. $V(0) = 0$.
2. $V(x) > 0,\ 0 < \|x\| < r$ for some r.

Similarly, the function is called locally *positive semidefinite* (lpsd) if the strict inequality on the function in the second condition is replaced by $V(x) \geq 0$. The function $V(x)$ is locally *negative definite* (lnd) if $-V(x)$ is lpd, and locally *negative semidefinite* (lnsd) if $-V(x)$ is lpsd. What may be useful in forming a mental picture of an lpd function $V(x)$ is to think of it as having "contours" of constant V that form (at least in a small region around the origin) a nested set of closed surfaces surrounding the origin. The situation for $n = 2$ is illustrated in Figure 13.2.

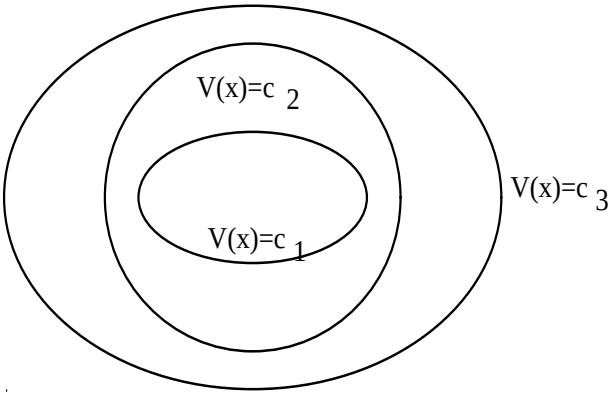


Figure 13.2: Level lines for a Lyapunov function, where $c_1 < c_2 < c_3$.

Throughout our treatment of the CT case, we shall restrict ourselves to $V(x)$ that have continuous first partial derivatives. (Differentiability will not be needed in the DT case — continuity will suffice there.) We shall denote the derivative of such a V with respect to time *along a trajectory of the system* (13.8) by $\dot{V}(x(t))$. This derivative is given by

$$\dot{V}(x(t)) = \frac{dV(x)}{dx}\,\dot{x} = \frac{dV(x)}{dx} f(x)$$

where $\frac{dV(x)}{dx}$ is a row vector — the *gradient* vector or *Jacobian* of V with respect to x — containing the component-wise partial derivatives $\frac{\partial V}{\partial x_i}$.

Definition 13.3 Let V be an lpd function (a "candidate Lyapunov function"), and let $\dot{V}$ be its derivative along trajectories of system (13.8). If $\dot{V}$ is lnsd, then V is called a *Lyapunov function* of the system (13.8).

Lyapunov Theorem for Local Stability

Theorem 13.1 If there exists a Lyapunov function of system (13.8), then $x = 0$ is a stable equilibrium point in the sense of Lyapunov. If in addition $\dot{V}(x) < 0$, $0 < \|x\| < r_1$ for some r_1, i.e. if $\dot{V}$ is lnd, then $x = 0$ is an asymptotically stable equilibrium point.

Proof: First, we prove stability in the sense of Lyapunov. Suppose $\epsilon > 0$ is given. We need to find a $\delta > 0$ such that for all $\|x(0)\| < \delta$, it follows that $\|x(t)\| < \epsilon$, $\forall t > 0$. The Figure 19.6 illustrates the constructions of the proof for the case $n = 2$. Let $\epsilon_1 = \min(\epsilon, r)$. Define

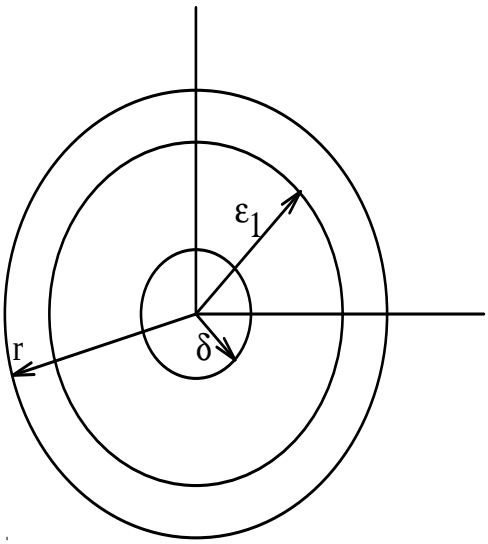


Figure 13.3: Illustration of the neighborhoods used in the proof

$$m = \min_{\|x\|=\epsilon_1} V(x).$$

Since $V(x)$ is continuous, the above m is well defined and positive. Choose δ satisfying $0 < \delta < \epsilon_1$ such that for all $\|x\| < \delta$, $V(x) < m$. Such a choice is always possible, again because of the continuity of $V(x)$. Now, consider any $x(0)$ such that $\|x(0)\| < \delta$, $V(x(0)) < m$, and let $x(t)$ be the resulting trajectory. $V(x(t))$ is non-increasing (i.e. $\dot{V}(x(t)) \leq 0$) which results in $V(x(t)) < m$. We will show that this implies that $\|x(t)\| < \epsilon_1$. Suppose there exists t_1 such that $\|x(t_1)\| > \epsilon_1$, then by continuity we must have that at an earlier time t_2, $\|x(t_2)\| = \epsilon_1$, and $\min_{\|x\|=\epsilon_1} \|V(x)\| = m > V(x(t_2))$, which is a contradiction. Thus stability in the sense of Lyapunov holds.

To prove asymptotic stability when $\dot{V}$ is lnd, we need to show that as $t \to \infty$, $V(x(t)) \to 0$; then, by continuity of V, $\|x(t)\| \to 0$. Since $V(x(t))$ is strictly decreasing, and $V(x(t)) \geq 0$ we know that $V(x(t)) \to c$, with $c \geq 0$. We want to show that c is in fact zero. We can argue by contradiction and suppose that $c > 0$. Let the set S be defined as

$$S = \{x \in \mathbb{R}^n | V(x) \leq c\},$$

and let B_α be a ball inside S of radius α,

$$B_\alpha = \{x \in S | \|x\| < \alpha\}.$$

Suppose $x(t)$ is a trajectory of the system that starts at $x(0)$, we know that $V(x(t))$ is decreasing monotonically to c and $V(x(t)) > c$ for all t. Therefore, $x(t) \notin B_\alpha$; recall that $B_\alpha \subset S$ which is defined as all the elements in $\mathbb{R}^n$ for which $V(x) \leq c$. In the first part of the proof, we have established that if $\|x(0)\| < \delta$ then $\|x(t)\| < \epsilon$. We can define the largest derivative of $V(x)$ as

$$-\gamma = \max_{\alpha \leq \|x\| \leq \epsilon} \dot{V}(x).$$

Clearly $-\gamma < 0$ since $\dot{V}(x)$ is lnd. Observe that,

$$\begin{aligned} V(x(t) &= V(x(0)) + \int_0^t \dot{V}(x(\tau)) d\tau \\ &\leq V(x(0)) - \gamma t, \end{aligned}$$

which implies that $V(x(t))$ will be negative which will result in a contradiction establishing the fact that c must be zero.

Example 13.2 Consider the dynamical system which is governed by the differential equation

$$\dot{x} = -g(x)$$

where $g(x)$ has the form given in Figure 13.4. Clearly the origin is an equilibrium point. If we define a function

$$V(x) = \int_0^x g(y) dy$$

then it is clear that $V(x)$ is locally positive definite (lpd) and

$$\dot{V}(x) = -g(x)^2$$

which is locally negative definite (lnd). This implies that $x = 0$ is an asymptotically stable equilibrium point.

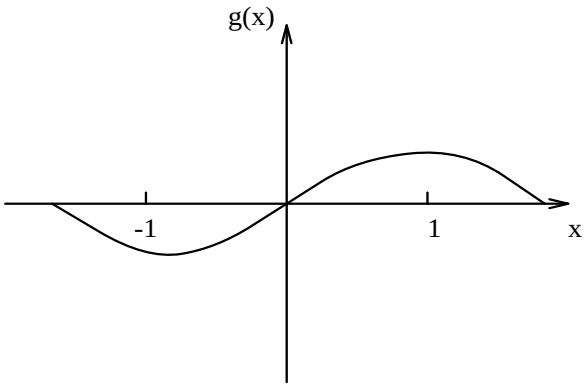


Figure 13.4: Graphical Description of $g(x)$

Lyapunov Theorem for Global Asymptotic Stability

The region in the state space for which our earlier results hold is determined by the region over which $V(x)$ serves as a Lyapunov function. It is of special interest to determine the "basin of attraction" of an asymptotically stable equilibrium point, i.e. the set of initial conditions whose subsequent trajectories end up at this equilibrium point. An equilibrium point is *globally asymptotically stable* (or asymptotically stable "in the large") if its basin of attraction is the entire state space.

If a function $V(x)$ is positive definite on the entire state space, *and* has the additional property that $|V(x)| \nearrow \infty$ as $\|x\| \nearrow \infty$, *and* if its derivative $\dot{V}$ is negative definite on the entire state space, then the equilibrium point at the origin is globally asymptotically stable. We omit the proof of this result. Other versions of such results can be stated, but are also omitted.

Example 13.3

Consider the nth-order system

$$\dot{x} = -C(x)$$

with the property that $C(0) = 0$ and $x'C(x) > 0$ if $x \neq 0$. Convince yourself that the unique equilibrium point of the system is at 0. Now consider the candidate Lyapunov function

$$V(x) = x'x$$

which satisfies all the desired properties, including $|V(x)| \nearrow \infty$ as $\|x\| \nearrow \infty$. Evaluating its derivative along trajectories, we get

$$\dot{V}(x) = 2x'\dot{x} = -2x'C(x) < 0 \quad \text{for } x \neq 0$$

Hence, the system is globally asymptotically stable.

Example 13.4 Consider the following dynamical system

$$\begin{aligned} \dot{x}_1 &= -x_1 + 4x_2 \\ \dot{x}_2 &= -x_1 - x_2^3. \end{aligned}$$

The only equlibrium point for this system is the origin $x = 0$. To investigate the stability of the origin let us propose a quadratic Lyapunov function $V = x_1^2 + ax_2^2$, where a is a positive constant to be determined. It is clear that V is positive definite on the entire state space $\mathbb{R}^2$. In addition, V is radially unbounded, that is it satisfies $|V(x)| \nearrow \infty$ as $\|x\| \nearrow \infty$. The derivative of V along the trajectories of the system is given by

$$\begin{aligned} \dot{V} &= \begin{bmatrix} 2x_1 & 2ax_2 \end{bmatrix} \begin{bmatrix} -x_1 + 4x_2 \\ -x_1 - x_2^3 \end{bmatrix} \\ &= -2x_1^2 + (8 - 2a)x_1x_2 - 2ax_2^4. \end{aligned}$$

If we choose $a = 4$ then we can eliminate the cross term x_1x_2, and the derivative of V becomes

$$\dot{V} = -2x_1^2 - 8x_2^4,$$

which is clearly a negative definite function on the entire state space. Therefore we conclude that $x = 0$ is a globally asymptotically stable equilibrium point.

Example 13.5 A highly studied example in the area of dynamical systems and chaos is the famous Lorenz system, which is a nonlinear system that evolves in $\mathbb{R}^3$ whose equations are given by

$$\begin{aligned} \dot{x} &= \sigma(y - x) \\ \dot{y} &= rx - y - xz \\ \dot{z} &= xy - bz, \end{aligned}$$

where σ, r and b are positive constants. This system of equations provides an approximate model of a horizontal fluid layer that is heated from below. The warmer fluid from the bottom rises and thus causes convection currents. This approximates what happens in the atmosphere. Under intense heating this model exhibits complex dynamical behaviour. However, in this example we would like to analyze the stability of the origin under the condition $r < 1$, which is known not to lead to complex behaviour. Le us define $V = \alpha_1 x^2 + \alpha_2 y^2 + \alpha_3 z^2$, where α_1, α_2, and α_3 are positive constants to be determined. It is clear that V is positive definite on $\mathbb{R}^3$ and is radially unbounded. The derivative of V along the trajectories of the system is given by

$$\dot{V} = \begin{bmatrix} 2\alpha_1 x & 2\alpha_2 y & 2\alpha_3 z \end{bmatrix} \begin{bmatrix} \sigma(y - x) \\ rx - y - xz \\ xy - bz \end{bmatrix}$$

$$
\begin{aligned}
&= -2\alpha_1\sigma x^2 - 2\alpha_2 y^2 - 2\alpha_3 b z^2 \\
&\quad + xy(2\alpha_1\sigma + 2r\alpha_2) + (2\alpha_3 - 2\alpha_2)xyz.
\end{aligned}
$$

If we choose $\alpha_2 = \alpha_3 = 1$ and $\alpha_1 = \frac{1}{\sigma}$ then the $\dot{V}$ becomes

$$
\begin{aligned}
\dot{V} &= -2\left(x^2 + y^2 + 2bz^2 - (1+r)xy\right) \\
&= -2\left[\left(x - \frac{1}{2}(1+r)y\right)^2 + \left(1 - (\frac{1+r}{2})^2\right)y^2 + bz^2\right].
\end{aligned}
$$

Since $0 < r < 1$ it follows that $0 < \frac{1+r}{2} < 1$ and therefore $\dot{V}$ is negative definite on the entire state space $\mathbb{R}^3$. This implies that the origin is globally asymptotically stable.

Example 13.6 (Pendulum)

The dynamic equation of a pendulum comprising a mass M at the end of a rigid but massless rod of length R is

$$
MR\ddot{\theta} + Mg\sin\theta = 0
$$

where θ is the angle made with the downward direction, and g is the acceleration due to gravity. To put the system in state-space form, let $x_1 = \theta$, and $x_2 = \dot{\theta}$; then

$$
\begin{aligned}
\dot{x_1} &= x_2 \\
\dot{x_2} &= -\frac{g}{R}\sin x_1
\end{aligned}
$$

Take as a candidate Lyapunov function the total energy in the system. Then

$$
\begin{aligned}
V(x) &= \frac{1}{2}MR^2x_2^2 + MgR(1 - \cos x_1) = \text{kinetic + potential} \\
\dot{V} = \frac{dV}{dx}f(x) &= \begin{bmatrix} MgR\sin x_1 & MR^2x_2 \end{bmatrix}\begin{bmatrix} x_2 \\ -\frac{g}{R}\sin x_1 \end{bmatrix} \\
&= 0
\end{aligned}
$$

Hence, V is a Lyapunov function and the system is stable i.s.L. We cannot conclude asymptotic stability with this analysis.

Consider now adding a damping torque proportional to the velocity, so that the state-space description becomes

$$
\begin{aligned}
\dot{x_1} &= x_2 \\
\dot{x_2} &= -Dx_2 - \frac{g}{R}\sin x_1
\end{aligned}
$$

With this change, but the same V as before, we find

$$\dot{V} = -DMR^2 x_2^2 \leq 0.$$

From this we can conclude stability i.s.L. We still cannot directly conclude asymptotic stability. Notice however that $\dot{V} = 0 \Rightarrow \dot{\theta} = 0$. Under this condition, $\ddot{\theta} = -(g/R)\sin\theta$. Hence, $\ddot{\theta} \neq 0$ if $\theta \neq k\pi$ for integer k, i.e. if the pendulum is not vertically down or vertically up. This implies that, unless we are at the bottom or top with zero velocity, we shall have $\ddot{\theta} \neq 0$ when $\dot{V} = 0$, so $\dot{\theta}$ will not remain at 0, and hence the Lyapunov function will begin to decrease again. The only place the system can end up, therefore, is with zero velocity, hanging vertically down or standing vertically up, i.e. at one of the two equilibria. The formal proof of this result in the general case ("LaSalle's invariant set theorem") is beyond the scope of this course.

The conclusion of local asymptotic stability can also be obtained directly through an alternative choice of Lyapunov function. Consider the Lyapunov function candidate

$$V(x) = \frac{1}{2}x_2^2 + \frac{1}{2}(x_1 + x_2)^2 + 2(1 - \cos x_1).$$

It follows that

$$\dot{V} = -(x_2^2 + x_1 \sin x_1) = - - (\dot{\theta}^2 + \theta \sin\theta) \leq 0.$$

Also, $\dot{\theta}^2 + \theta\sin\theta = 0 \Rightarrow \dot{\theta}^2 = 0,\ \theta\sin\theta = 0 \Rightarrow \theta = 0,\ \dot{\theta} = 0$. Hence, $\dot{V}$ is strictly negative in a small neighborhood around 0. This proves asymptotic stability.

Discrete-Time Systems

Essentially identical results hold for the system

$$x(k+1) = f(x(k)) \tag{13.9}$$

provided we interpret $\dot{V}$ as

$$\dot{V}(x) \stackrel{\triangle}{=} V(f(x)) - V(x)\ ,$$

i.e. as

$$V(\text{next state}) - V(\text{present state})$$

Example 13.7 (DT System)

Consider the system

$$\begin{aligned} x_1(k+1) &= \frac{x_2(k)}{1 + x_2^2(k)} \\ x_2(k+1) &= \frac{x_1(k)}{1 + x_2^2(k)} \end{aligned}$$

which has its only equilibrium at the origin. If we choose the quadratic Lyapunov function

$$V(x) = x_1^2 + x_2^2$$

we find

$$\dot{V}(x(k)) = V(x(k)) \left(\frac{1}{[1 + x_2^2(k)]^2} - 1 \right) \leq 0$$

from which we can conclude that the equilibrium point is stable i.s.L. In fact, examining the above relations more carefully (in the same style as we did for the pendulum with damping), it is possible to conclude that the equilibrium point is actually *globally asymptotically stable*.

Notes

The system in Example 2 is taken from the eminently readable text by F. Verhulst, *Nonlinear Differential Equations and Dynamical Systems*, Springer-Verlag, 1990.

Exercises

Exercise 13.1 Consider the horizontal motion of a particle of unit mass sliding under the influence of gravity on a frictionless wire. It can be shown that, if the wire is bent so that its height h is given by $h(x) = V_\alpha(x)$, then a state-space model for the motion is given by

$$\begin{aligned} \dot{x} &= z \\ \dot{z} &= -\frac{d}{dx} V_\alpha(x), \end{aligned}$$

Suppose $V_\alpha(x) = x^4 - \alpha x^2$.

(a) Verify that the above model has $(z, x) = (0, 0)$ as equilibrium point for any α in the interval $-1 \leq \alpha \leq 1$, and it also has $(z, x) = \left(0, \pm\sqrt{\frac{\alpha}{2}}\right)$ as equilibrium points when α is in the interval $0 < \alpha \leq 1$.

(b) Verify that the linearized model about any of the equilibrium points is neither asymptotically stable nor unstable for any α in the interval $-1 \leq \alpha \leq 1$.

Exercise 13.2 Consider the dynamic system described below:

$$\ddot{y} + a_1 \dot{y} + a_2 y + c y^2 = u + \dot{u},$$

where y is the output and u is the input.

(a) Obtain a state-space realization of dimension 2 that describes the above system.

(b) If $a_1 = 3$, $a_2 = 2$, $c = 2$, show that the system is asymptotically stable at the origin.

(c) Find a region (a disc of non-zero radius) around the origin such that every trajectory, with an initial state starting in this region, converges to zero as t approaches infinity. This is known as a region of attraction.

Exercise 13.3 Consider the system

$$\dot{x}(t) = -\frac{dP(x)}{dx}$$

where $P(x)$ has continuous first partial derivatives. The function $P(x)$ is referred to as the *potential function* of the system, and the system is said to be a *gradient* system. Let $\bar{x}$ be an isolated local minimum of $P(x)$, i.e. $P(\bar{x}) < P(x)$ for $0 < \|x - \bar{x}\| < r$, some r.

(a) Show that $\bar{x}$ is an equilibrium point of the gradient system.

(b) Use the candidate Lyapunov function

$$V(x) = P(x) - P(\bar{x})$$

to try and establish that $\bar{x}$ is an asymptotically stable equilibrium point.

Exercise 13.4 The objective of this problem is to analyze the convergence of the gradient algorithm for finding a local minimum of a function. Let $f : \mathbb{R}^n \to \mathbb{R}$ and assume that x^* is a local minimum; i.e., $f(x^*) < f(x)$ for all x close enough but not equal to x^*. Assume that f is continuously differentiable. Let $g^T : \mathbb{R} \to \mathbb{R}^n$ be the gradient of f:

$$g^T = \left(\frac{\partial g}{\partial x_1} \quad \cdots \quad \frac{\partial g}{\partial x_n} \right).$$

It follows from elementary Calculus that $g(x^*) = 0$.

If one has a good estimate of x^*, then it is argued that the solution to the dynamic system:

$$\dot{x} = -g(x) \tag{13.10}$$

with $x(0)$ close to x^* will give $x(t)$ such that

$$\lim_{t \to \infty} x(t) = x^*.$$

(a) Use Lyapunov stability analysis methods to give a precise statement and a proof of the above argument.

(b) System 13.10 is usually solved numerically by the discrete-time system

$$x(k+1) = x(k) - \alpha(x_k) g(x_k), \tag{13.11}$$

where $\alpha(x_k)$ is some function from $\mathbb{R}^n \to \mathbb{R}$. In certain situations, α can be chosen as a constant function, but this choice is not always good. Use Lyapunov stability analysis methods for discrete-time systems to give a possible choice for $\alpha(x_k)$ so that

$$\lim_{k \to \infty} x(k+1) = x^*.$$

(c) Analyze directly the gradient algorithm for the function

$$f(x) = \frac{1}{2} x^T Q x, \quad Q \text{ Symmetric, Positive Definite.}$$

Show directly that system 13.10 converges to zero ($= x^*$). Also, show that α in system 13.11 can be chosen as a real constant, and give tight bounds on this choice.

Exercise 13.5 **(a)** Show that any (possibly complex) square matrix M can be written uniquely as the sum of a Hermitian matrix H and a skew-Hermitian matrix S, i.e. $H' = H$ and $S' = -S$. (Hint: Work with combinations of M and M'.) Note that if M is real, then this decomposition expresses the matrix as the sum of a symmetric and skew-symmetric matrix.

(b) With M, H, and S as above, show that the real part of the quadratic form $x'Mx$ equals $x'Hx$, and the imaginary part of $x'Mx$ equals $x'Sx$. (It follows that if M and x are real, then $x'Mx = x'Hx$.)

(c) Let $V(x) = x'Mx$ for real M and x. Using the standard definition of $dV(x)/dx$ as a Jacobian matrix — actually just a row vector in this case — whose jth entry is $\partial V(x)/\partial x_j$, show that

$$\frac{dV(x)}{dx} = 2x'H$$

where H is the symmetric part of M, as defined in part (a).

(d) Show that a Hermitian matrix always has real eigenvalues, and that the eigenvectors associated with distinct eigenvalues are *orthogonal* to each other.

Exercise 13.6 Consider the (real) continuous-time LTI system $\dot{x}(t) = Ax(t)$.

(a) Suppose the (continuous-time) *Lyapunov equation*

$$PA + A'P = -I \tag{3.1}$$

has a symmetric, positive definite solution P. Note that (3.1) can be written as a *linear* system of equations in the entries of P, so solving it is in principle straightforward; good numerical algorithms exist.

Show that the function $V(x) = x'Px$ serves as a Lyapunov function, and use it to deduce the global asymptotic stability of the equilibrium point of the LTI system above, i.e. to deduce that the eigenvalues of A are in the open left-half plane. (The result of Exercise 13.5 will be helpful in computing $\dot{V}(x)$.)

What part (a) shows is that the existence of a symmetric, positive definite solution of (3.1) is *sufficient* to conclude that the given LTI system is asymptotically stable. The existence of such a solution turns out to also be *necessary*, as we show in what follows. [Instead of $-I$ on the right side of (3.1), we could have had $-Q$ for any positive definite matrix Q. It would still be true that the system is asymptotically stable if and only if the solution P is symmetric, positive definite. We leave you to modify the arguments here to handle this case.]

(b) Suppose the LTI system above is asymptotically stable. Now define

$$P = \int_0^\infty R(t)dt \ , \qquad R(t) = e^{A't}e^{At} \tag{3.2}$$

The reason the integral exists is that the system is asymptotically stable — explain this in more detail! Show that P is symmetric and positive definite, and that it is the *unique* solution of the Lyapunov equation (3.1). You will find it helpful to note that

$$R(\infty) - R(0) = \int_0^\infty \frac{dR(t)}{dt}\, dt$$

The results of this problem show that one can decide whether a matrix A has all its eigenvalues in the open left-half plane without solving for all its eigenvalues. We only need to test for the positive definiteness of the solution of the linear system of equations (3.1). This can be simpler.

Exercise 13.7 This problem uses Lyapunov's direct method to justify a key claim of his *indirect method*: if the *linearized* model at an equilibrium point is asymptotically stable, then this equilibrium point of the nonlinear system is asymptotically stable. (We shall actually only consider an equilibrium point at the origin, but the approach can be applied to any equilibrium point, after an appropriate change of variables.)
Consider the time-invariant continuous-time *nonlinear* system given by

$$\dot{x}(t) = Ax(t) + h(x(t)) \tag{4.1}$$

where A has all its eigenvalues in the open left-half plane, and $h(.)$ represents "higher-order terms", in the sense that $\|h(x)\|/\|x\| \to 0$ as $\|x\| \to 0$.

(a) Show that the origin is an equilibrium point of the system (4.1), and that the linearized model at the origin is just $\dot{x}(t) = Ax(t)$.

(b) Let P be the positive definite solution of the Lyapunov equation in (3.1). Show that $V(x) = x'Px$ qualifies as a *candidate* Lyapunov function for testing the stability of the equilibrium point at the origin in the system (4.1). Determine an expression for $\dot{V}(x)$, the rate of change of $V(x)$ along trajectories of (4.1)

(c) Using the fact that $x'x = \|x\|^2$, and that $\|Ph(x)\| \leq \|P\|\|h(x)\|$, how small a value (in terms of $\|P\|$) of the ratio $\|h(x)\|/\|x\|$ will allow you to conclude that $\dot{V}(x(t)) < 0$ for $x(t) \neq 0$? Now argue that you can indeed limit $\|h(x)\|/\|x\|$ to this small a value by choosing a small enough neighborhood of the equilibrium. In this neighborhood, therefore, $\dot{V}(x(t)) < 0$ for $x(t) \neq 0$. By Lyapunov's direct method, this implies asymptotic stability of the equilibrium point.

Exercise 13.8 For the discrete-time LTI system $x(k+1) = Ax(k)$, let $V(x) = x'Px$, where P is a symmetric, positive definite matrix. What condition will guarantee that $V(x)$ is a Lyapunov function for this system? What condition involving A and P will guarantee asymptotic stability of the system? (Express your answers in terms of the positive semidefiniteness and definiteness of a matrix.)

Chapter 14

Internal Stability for LTI Systems

14.1 Introduction

Constructing a Lyapunov function for an arbitrary nonlinear system is not a trivial exercise. The complication arises from the fact that we cannot restrict the class of functions to search from in order to prove stability. The situation is different for LTI systems. In this chapter, we address the question of constructing Lyapunov functions for linear systems and then we present and verify Lyapunov indirect method for proving stability of a nonlinear system.

14.2 Quadratic Lyapunov Functions for LTI Systems

Consider the continuous-time system

$$\dot{x}(t) = Ax(t) \; . \tag{14.1}$$

We have already established that the system (14.1) is asymptotically stable if and only if all the eigenvalues of A are in the open left half plane. In this section we will show that this result can be inferred from Lyapunov theory. Moreover, it will be shown that *quadratic* Lyapunov functions suffice. A consequence of this is that stability can be assessed by methods that may be computationally simpler than eigenanalysis. More importantly, quadratic Lyapunov functions and the associated mathematics turn up in a variety of other problems, so they are worth mastering in the context of stability evaluation.

Quadratic Positive-Definite Functions

Consider the function

$$V(x) = x^T P x, \quad x \in \mathbb{R}^n$$

where P is a symmetric matrix. This is the general form of a quadratic function in $\mathbb{R}^n$. It is sufficient to consider symmetric matrices; if P is not symmetric, we can define $P_1 = \frac{1}{2}(P+P^T)$. It follows immediately that $x^TPx = x^TP_1x$ (verify, using the fact that x^TPx is a scalar).

Proposition 14.1 $V(x)$ is a positive definite function if and only if all the eigenvalues of P are positive.

Proof: Since P is symmetric, it can be diagonalized by an orthogonal matrix, *i.e.*,

$$P = U^TDU \text{ with } U^TU = I \text{ and } D \text{ diagonal.}$$

Then, if $y = Ux$

$$V(x) = x^TU^TDUx = y^TDy = \sum_i \lambda_i |y_i|^2.$$

Thus,

$$V(x) > 0 \ \ \forall x \neq 0 \Leftrightarrow \lambda_i > 0, \ \forall i.$$

Definition 14.1 A matrix P that satisfies

$$x^TPx > 0 \quad \forall x \neq 0 \tag{14.2}$$

is called *positive definite*. When P is symmetric (which is usually the case of interest, for the reason mentioned above), we will denote its positive definiteness by $P > 0$. If $x^TPx \geq 0 \quad \forall x \neq 0$, then P is positive semi-definite, which we denote in the symmetric case by $P \geq 0$.

For a symmetric positive definite matrix, it follows that

$$\lambda_{\min}(P)\|x\|^2 \leq V(x) \leq \lambda_{\max}(P)\|x\|^2.$$

This inequality follows directly from the proof of Proposition 14.1.

It is also evident from the above discussion that the singular values and eigenvalues of any positive definite matrix coincide.

Exercise: Show that $P > 0$ if and only if $P = G^TG$ where G is nonsingular. The matrix G is called a square root of P and is denoted by $P^{\frac{1}{2}}$. Show that H is another square root of P if and only if $G = WH$ for some orthogonal matrix W. Can you see how to construct a *symmetric* square root? (You may find it helpful to begin with the eigen-decomposition $P = U^TDU$, where U is orthogonal and D is diagonal.)

Quadratic Lyapunov Functions for CT LTI Systems

Consider defining a Lyapunov function candidate of the form

$$V(x) = x^T P x, \quad P > 0, \tag{14.3}$$

for the system (14.1). Then

$$\begin{aligned} \dot{V}(x) &= \dot{x}^T P x + x^T P \dot{x} \\ &= x^T A^T P x + x^T P A x \\ &= x^T (A^T P + P A) x \\ &= -x^T Q x \ , \end{aligned}$$

where we have introduced the notation $Q = -(A^T P + PA)$; note that Q is symmetric. Now invoking the Lyapunov stability results from Lecture 5, we see that V is a Lyapunov function if $Q \geq 0$, in which case the equilibrium point at the origin of the system (14.1) is stable i.s.L. If $Q > 0$, then the equilibrium point at the origin is *globally asymptotically stable*. In this latter case, the origin must be the only equilibrium point of the system, so we typically say the *system* (rather than just the equilibrium point) is asymptotically stable.

The preceding relationships show that in order to find a quadratic Lyapunov function for the system (14.1), we can pick $Q > 0$ and then try to solve the equation

$$A^T P + PA = -Q \tag{14.4}$$

for P. This equation is referred to as a *Lyapunov equation*, and is a *linear* system of equations in the entries of P. If it has a solution, then it has a symmetric solution (show this!), so we only consider symmetric solutions. If it has a positive definite solution $P > 0$, then we evidently have a Lyapunov function $x^T P x$ that will allow us to prove the asymptotic stability of the system (14.1). The interesting thing about LTI systems is that the converse also holds: If the system is asymptotically stable, then the Lyapunov equation (14.4) has positive definite solution $P > 0$ (which, as we shall show, is unique). This result is stated and proved in the following theorem.

Theorem 14.1 Given the dynamic system (14.1) and any $Q > 0$, there exists a positive definite solution P of the Lyapunov equation

$$A^T P + PA = -Q$$

if and only if all the eigenvalues of A are in the open left half plane (OLHP). The solution P in this case is unique.

Proof: If $P > 0$ is a solution of (14.4), then $V(x) = x^T P x$ is a Lyapunov function of system (14.1) with $\dot{V}(x) < 0$ for any $x \neq 0$. Hence, system (14.1) is (globally) asymptotically stable and thus the eigenvalues of A are in the OLHP.

To prove the converse, suppose A has all eigenvalues in the OLHP, and $Q > 0$ is given. Define the symmetric matrix P by

$$P = \int_0^\infty e^{tA^T} Q e^{tA}\, dt\ . \tag{14.5}$$

This integral is well defined because the integrand decays exponentially to the origin, since the eigenvalues of A are in the OLHP. Now

$$\begin{aligned} A^T P + PA &= \int_0^\infty A^T e^{tA^T} Q e^{tA} dt + \int_0^\infty e^{tA^T} Q e^{tA} A dt \\ &= \int_0^\infty \frac{d}{dt}\left[e^{tA^T} Q e^{tA}\right] dt \\ &= -Q \end{aligned}$$

so P satisfies the Lyapunov equation.

To prove that P is positive definite, note that

$$\begin{aligned} x^T P x &= \int_0^\infty x^T e^{tA^T} Q e^{tA} x dt \\ &= \int_0^\infty \|Q^{\frac{1}{2}} e^{tA} x\|^2 dt \geq 0 \end{aligned}$$

and

$$x^T P x = 0 \Rightarrow Q^{\frac{1}{2}} e^{tA} x = 0 \Rightarrow x = 0\ ,$$

where $Q^{\frac{1}{2}}$ denotes a square root of Q. Hence P is positive definite.

To prove that the P defined in (14.5) is the unique solution to (14.4) when A has all eigenvalues in the OLHP, suppose that P_2 is another solution. Then

$$\begin{aligned} P_2 &= -\int_0^\infty \frac{d}{dt}\left[e^{tA^T} P_2 e^{tA}\right] dt \qquad \text{(verify this identity)} \\ &= -\int_0^\infty e^{tA^T}\left(A^T P_2 + P_2 A\right) e^{tA} dt \\ &= \int_0^\infty e^{tA^T} Q e^{tA} dt = P \end{aligned}$$

This completes the proof of the theorem.

A variety of generalizations of this theorem are known.

Quadratic Lyapunov Functions for DT LTI Systems

Consider the system

$$x(t+1) = Ax(t) = f\Big(x(t)\Big) \tag{14.6}$$

If

$$V(x) = x^T P x,$$

then

$$\dot{V}(x) \stackrel{\triangle}{=} V(f(x)) - V(x) = x^T A^T P A x - x^T P x.$$

Thus the resulting Lyapunov equation to study is

$$A^T P A - P = -Q\,. \tag{14.7}$$

The following theorem is analogous to what we proved in the CT case, and we leave its proof as an exercise.

Theorem 14.2 Given the dynamic system (14.6) and any $Q > 0$, there exists a positive definite solution P of the Lyapunov equation

$$A^T P A + P = -Q$$

if and only if all the eigenvalues of A have magnitude less than 1 (i.e. are in the open unit disc). The solution P in this case is unique.

Example 14.1 Differential Inclusion

In many situations, the evolution of a dynamic system can be uncertain. One way of modeling this uncertainty is by differential (difference) inclusion which can be described as follows:

$$\dot{x}(t) \subset \{Ax(t) \,|\, A \subset \mathcal{A}\}$$

where $\mathcal{A}$ is a set of matrices. Consider the case where $\mathcal{A}$ is a finite set of matrices and their convex combinations:

$$\mathcal{A} = \{A = \sum_{i=1}^{m} \alpha_i A_i \,|\, \sum_{i=1}^{m} \alpha_i = 1\}$$

One way to guarantee the stability of this system is to find *one* Lyapunov function for all systems defined by $\mathcal{A}$. If we look for a quadratic Lyapunov function, then it suffices to find a P that satisfies:

$$A_i^T P + P A_i < -Q, \quad i = 1\ \ 2\ \ \ldots m$$

for some positive definite Q. Then $V(x) = x^T P x$ satisfies $\dot{V}(x) < -x^T Q x$ (verify) showing that the system is asymptotically stable.

Example 14.2 Set of Bounded Norm

In this problem, we are interested in studying the stability of linear time-invariant systems of the form $\dot{x}(t) = (A + \Delta)x(t)$ where Δ is a real matrix perturbation with bounded norm. In particular, we are interested in calculating a good bound on the size of the smallest perturbation that will destabilize a stable matrix A.

This problem can be cast as a differntial inclusion problem as in the previous example with

$$\mathcal{A} = \{A + \Delta \,\|\, \|\Delta\| \leq \gamma \;\; \Delta \text{ is a real matrix}\}$$

Since A is stable, we can calculate a quadratic Lyapunov function with a matrix P satisfying $A^T P + PA < -Q$ and Q is positive definite. Applying the same Lyapunov function to the perturbed system we get:

$$\dot{V}(x) = x^T \quad A^T P + PA + \Delta^T P + P\Delta \quad x$$

It is evident that all perturbations satisfying

$$\Delta^T P + P\Delta < Q$$

will result in a stable system. This can be guaranteed if

$$2\sigma_{max}(P)\sigma_{max}(\Delta) < \sigma_{min}(Q)$$

This provides a bound on the perturbation although it is potentially conservative.

Example 14.3 Bounded Perturbation

Casting the perturbation in the previous example in terms of differential inclusion introduces a degree of conservatism in that the value Δ takes can change as a function of time. Consider the system:

$$\dot{x}(t) = (A + \Delta)x(t)$$

where A is a known fixed stable matrix and Δ is an unknown fixed real perturbation matrix. The *stability margin* of this system is defined as

$$\gamma(A) = \min_{\Delta \in \mathbb{R}^{n \times n}} \{ \|\Delta\| \,|\, A + \Delta \text{ is unstable } \}.$$

We desire to compute a good lower bound on $\gamma(A)$. The previous example gave one such bound.

First, it is easy to argue that the minimizing solution Δ_o of the above problem results in $A + \Delta_0$ having eigenvalues at the imaginary axis (either at the origin, or in two complex conjugate locations). This is a consequence of the fact that the eigenvalues of $A + p\Delta_0$ will move continuously in the complex plane as the parameter p varies from 0 to 1. The intersection with the imaginary axis will happen at $p = 1$; if not, a perturbation of smaller size can be found.

We can get a lower bound on γ by dropping the condition that Δ is a real matrix, and allowing complex matrices (is it clear why this gives a lower bound?). We can show:

$$\min_{\Delta \in \mathbf{C}^{n \times n}} \{ \|\Delta\| \,|\, A + \Delta \text{ is unstable } \} = \min_{\omega \in \mathbb{R}} \sigma_{min}(A - j\omega I).$$

To verify this, notice that if the minimizing solution has an eigenvalue at the imaginary axis, then $j\omega_0 I - A - \Delta_0$ should be singular while we know that $j\omega_0 - A$ is not. The smallest possible perturbation that achieves this has size $\sigma_{min}(A - j\omega_0 I)$. We can then choose ω_0 that gives the smallest possible size. In the exercises, we further improve this bound.

14.3 Lyapunov's Indirect Method: Analyzing the Linearization

Suppose the system

$$\dot{x} = f(x) \tag{14.8}$$

has an equilibrium point at $\overline{x} = 0$ (an equilibrium at any other location can be dealt with by a preliminary change of variables to move that equilibrium to the origin). Assume we can write

$$f(x) = Ax + h(x)$$

where

$$\lim_{\|x\| \to 0} \frac{\|h(x)\|}{\|x\|} = 0$$

i.e. $h(x)$ denotes terms that are higher order than linear, and A is the Jacobian matrix associated with the linearization of (14.8) about the equilibrium point. The linearized system is thus given by

$$\dot{x} = Ax\ . \tag{14.9}$$

We might expect that if (14.9) is asymptotically stable, then in a small neighborhood around the equilibrium point, the system (14.8) behaves like (14.9) and will be stable. This is made precise in the following theorem.

Theorem 14.3 If the system (14.9) is asymptotically stable, then the equilibrium point of system (14.8) at the origin is (locally) asymptotically stable.

Proof: If system (14.9) is asymptotically stable, then for any $Q > 0$, there exists $P > 0$ such that

$$A^T P + PA = -Q$$

and $V(x) = x^T P x$ is a Lyapunov function for system (14.9). Consider $V(x)$ as a Lyapunov function candidate for system (14.8). Then

$$\begin{aligned}
\dot{V}(x) &= x^T(A^T P + PA)x + 2x^T P h(x) \\
&\leq -\lambda_{\min}(Q)\|x\|^2 + 2\|x\| \cdot \|h(x)\| \cdot \lambda_{\max}(P) \\
&\leq -\left[\lambda_{\min}(Q) - 2\lambda_{\max}(P)\frac{\|h(x)\|}{\|x\|}\right] \cdot \|x\|^2
\end{aligned}$$

From the assumption on h, for every $\epsilon > 0$, there exists $r > 0$ such that

$$\|h(x)\| < \epsilon \|x\| \qquad \forall \ \|x\| < r.$$

This implies that $\dot{V}$ is strictly negative for all $\|x\| < r$, where r is chosen for

$$\epsilon < \frac{\lambda_{\min}(Q)}{2\lambda_{\max}(P)}.$$

This concludes the proof.

Notice that asymptotic stability of the equilibrium point of the system (14.8) can be concluded from the asymptotic stability of the linearized system (14.9) only when the eigenvalues of A have negative real parts. It can also be shown that if there is any eigenvalue of A in the right half plane, i.e. if the linearization is exponentially unstable, then the equilibrium point of the nonlinear system is unstable. The above theorem is inconclusive if there are eigenvalues on the imaginary axis, but none in the right half plane. The higher-order terms of the nonlinear model can in this case play a decisive role in determining stability; for instance, if the linearization is polynomially (rather than exponentially) unstable, due to the presence of one or more Jordan blocks of size greater than 1 for eigenvalues on the imaginary axis (and the absence of eigenvalues in the right half plane), then the higher-order terms can still cause the equilibrium point to be stable.

It turns out that stronger versions of the preceding theorem hold if A has no eigenvalues on the imaginary axis: not only the stability properties of the equilibrium point, but also the local behavior of (14.8) can be related to the behavior of (14.9). We will not discuss these results further here.

Similar results hold for discrete-time systems.

Example 14.4

The equations of motion for a pendulum with friction are

$$\begin{aligned} \dot{x_1} &= x_2 \\ \dot{x_2} &= -x_2 - \sin x_1 \end{aligned}$$

The two equilibrium points of the system are at (0 0) and (π 0). The linearized system at the origin is given by

$$\begin{aligned} \dot{x_1} &= x_2 \\ \dot{x_2} &= -x_1 - x_2 \end{aligned}$$

or

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & -1 \end{bmatrix} x = Ax \ .$$

This A has all its eigenvalues in the OLHP. Hence the equilibrium point at the origin is asymptotically stable. Note, however, that if there were no damping, then the linearized system would be

$$\dot{x} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} x$$

and the resulting matrix A has eigenvalues on the imaginary axis. No conclusions can be drawn from this situation using Lyapunov linearization methods. Lyapunov's direct method, by contrast, allowed us to conclude stability even in the case of zero damping, and also permitted some detailed global conclusions in the case with damping.

The linearization around the equilibrium point at $(\pi \ 0)$ is

$$\begin{aligned} \dot{z_1} &= z_2 \\ \dot{z_2} &= +z_1 - z_2 \end{aligned}$$

where $z_1 = x_1 - \pi$ and $z_2 = x_2$, so these variables denote the (small) deviations of x_1 and x_2 from their respective equilibrium values. Hence

$$A = \begin{bmatrix} 0 & 1 \\ 1 & -1 \end{bmatrix} x = Ax$$

which has one eigenvalues in the RHP, indicating that this equilibrium point is unstable.

Exercises

Exercise 14.1 Bounded Perturbation Recall Example 14.3. In this problem we want to improve the lower bound on $\gamma(A)$.

(a) To improve the lower bound, we use the information that if Δ is real, then poles appear in complex conjugate pair. Define

$$A_w = \begin{pmatrix} A & wI \\ -wI & A \end{pmatrix}.$$

Show that

$$\gamma(A) \geq \min_{w \in \mathbb{R}} \sigma_{min}[A_w].$$

(b) If you think harder about your proof above, you will be able to further improve the lower bound. In fact, it follows that

$$\gamma(A) \geq \min_{w \in \mathbb{R}} \sigma_{2n-1}[A_w]$$

where σ_{2n-1} is the next to last singular value. Show this result.

Exercise 14.2 Consider the LTI unforced system given below:

$$\dot{x} = Ax = \begin{pmatrix} 0 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_{N-1} & -a_{N-2} & \dots & \dots & \dots & -a_0 \end{pmatrix} x$$

(a) Under what conditions is this system asymptotically stable?

Assume the system above is asymptotically stable. Now, consider the perturbed system

$$\dot{x} = Ax + \Delta x,$$

where Δ is given by

$$\Delta = \begin{pmatrix} 0 & 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -\delta_{N-1} & -\delta_{N-2} & \dots & \dots & \dots & -\delta_0 \end{pmatrix}, \qquad \delta_i \in \mathbb{R}.$$

(b) Argue that the perturbation with the smallest Frobenius norm that destabilizes the system (makes the system not asymptotically stable) will result in $A + \Delta$ having an eigenvalue at the imaginary axis.

(c) Derive an exact expression for the smallest Frobenius norm of Δ necessary to destabilize the above system (i.e., $\dot{x} = (A+\Delta)x$ is not asymptotically stable). Give an expression for the perturbation Δ that attains the minimum.

(d) Evaluate your answer in part 3 for the case $N = 2$, and $a_0 = a_1$.

Exercise 14.3 Periodic Controllers

(a) Show that the periodically varying system in Exercise 7.4 is asymptotically stable if and only if all the eigenvalues of the matrix $[A_{N-1}...A_0]$ have magnitude less than 1.

(b) (i) Given the system

$$x(k+1) = \begin{pmatrix} 0 & 1 \\ 1 & -1 \end{pmatrix} x(k) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u(k) \ , \qquad y(k) = \begin{pmatrix} 1 & 1 \end{pmatrix} x(k)$$

write down a linear state-space representation of the closed-loop system obtained by implementing the linear output feedback control $u(k) = g(k)y(k)$.

(ii) It turns out that there is *no* constant gain $g(k) = g$ for which the above system is asymptotically stable. (**Optional:** Show this.) However, consider the periodically varying system obtained by making the gain take the value -1 for even k and the value 3 for odd k. Show that any nonzero initial condition in the resulting system will be brought to the origin in at most 4 steps. (The moral of this is that periodically varying output feedback can do more than constant output feedback.)

Exercise 14.4 Delay Systems

The material we covered in class has focused on finite-dimensional systems, i.e., systems that have state-space descriptions with a finite number of state variables. One class of systems that does not belong to the class of finite-dimensional systems is continuous-time systems with delays.

Consider the following forced continuous-time system:

$$y(t) + a_1 y(t-1) + a_2 y(t-2) + \ldots + a_N y(t-N) = u(t) \qquad t \geq N, \ t \in \mathbb{R}.$$

This is known as a delay system with commensurate delays (multiple of the same delay unit). We assume that $u(t) = 0$ for all $t < N$.

(a) Show that we can compute the solution $y(t)$, $t \geq N$, if $y(t)$ is completely known in the interval [0,N). Explain why this system cannot have a finite-dimensional state space description.

(b) To compute the solution $y(t)$ given the initial values (denote those by the function $f(t)$, $t \in [0, N)$, which we will call the initial function) and the input u, it is useful to think of every non-negative real number as $t = \tau + k$ with $\tau \in [0, 1)$ and k being a non-negative integer. Show that for every fixed τ, the solution evaluated at $\tau + k$ ($y(\tau + k)$) can be computed using discrete-time methods and can be expressed in terms of the matrix

$$A = \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 \\ 0 & 0 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ -a_N & -a_{N-1} & \ldots & \ldots & \ldots & -a_1 \end{pmatrix}$$

and the initial vector

$$\begin{pmatrix} f(\tau) & f(\tau+1) & \ldots & f(\tau+N-1) \end{pmatrix}^T .$$

Write down the general solution for $y(t)$.

(c) Compute the solution for $N = 2$, $f(t) = 1$ for $t \in [0, 2)$, and $u(t) = e^{-(t-2)}$ for $t \geq 2$.

(d) This system is asymptotically stable if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all initial functions with $|f(t)| < \delta$, $t \in [0, N)$, and $u = 0$, it follows that $|y(t)| < \epsilon$, and $\lim_{t\to\infty} y(t) = 0$. Give a necessary and sufficient condition for the asymptotic stability of this system. Explain your answer.

(e) Give a necessary and sufficient condition for the above system to be BIBO stable (∞-stable). Verify your answer.

Exercise 14.5 Local Stabilization

(a) One method for stabilizing a nonlinear system is to linearize it around an equilibrium point and then stabilize the resulting linear system. More formally, consider a nonlinear time-invariant system

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

and its linearization around an equilibrium point $(\tilde{x}, \tilde{u})$

$$\dot{\delta x} = A\delta x + B\delta u.$$

As usual, $\delta x = x - \tilde{x}$ and $\delta u = u - \tilde{u}$. Suppose that the feedback $\delta u = K\delta x$ asymptotically stabilizes the linearized system.

1. What can you say about the eigenvalues of the matrix $A + BK$.
2. Show that $\dot{x} = f(x, Kx)$ is (locally) asymptotically stable around $\tilde{x}$.

(b) Consider the dynamic system S_1 governed by the following differential equation:

$$\ddot{y} + \dot{y}^4 + \dot{y}^2 u + y^3 = 0$$

where u is the input.

1. Write down a state space representation for the system S_1 and find its unique equilibrium point x^*.
2. Now try to apply the above method to the system S_1 at the equilibrium point x^* and $u^* = 0$. Does the linearized system provide information about the stability of S_1. Explain why the method fails.

(c) To find a stabilizing controller for S_1, we need to follow approaches that are not based on local linearization. One approach is to pick a positive definite function of the states and then construct the control such that this function becomes a Lyapunov function. This can be a very frustrating exercise. A trick that is commonly used is to find an input as a function of the states so that the resulting system belongs to a class of systems that are known to be stable (e.g. a nonlinear circuit or a mechanical system that are known to be stable). Use this idea to find an input u as function of the states such that S_1 is stable.

Exercise 14.6 For the system

$$\dot{x}(t) = sin[x(t) + y(t)]$$
$$\dot{y}(t) = e^{x(t)} - 1$$

determine *all* equilibrium points, and using Lyapunov's indirect method (i.e. linearization), classify each equilibrium point as asymptotically stable or unstable.

Exercise 14.7 For each of the following parts, all of them **optional**, use Lyapunov's indirect method to determine, if possible, whether the origin is an asymptotically stable or unstable equilibrium point.

(a)

$$\dot{x}_1 = -x_1 + x_2^2$$
$$\dot{x}_2 = -x_2(x_1 + 1)$$

(b)

$$\dot{x}_1 = x_1^3 + x_2$$
$$\dot{x}_2 = x_1 - x_2$$

(c)

$$\dot{x}_1 = -x_1 + x_2$$
$$\dot{x}_2 = -x_2 + x_1^2$$

(d)

$$x_1(k+1) = 2x_1(k) + x_2(k)^2$$
$$x_2(k+1) = x_1(k) + x_2(k)$$

(e)

$$x_1(k+1) = 1 - e^{x_1(k)x_2(k)}$$
$$x_2(k+1) = x_1(k) + 2x_2(k)$$

Exercise 14.8 For each of the nonlinear systems below, construct a linearization for the equilibrium point at the origin, assess the stability of the linearization, and decide (using the results of Lyapunov's *indirect* method) whether you can infer something about the stability of the equilibrium of the nonlinear system at the origin. Then use Lyapunov's *direct* method prove that the origin is actually stable in each case; if you can make further arguments to actually deduce *asymptotic* stability or even global asymptotic stability, do so. [Hints: In part (a), find a suitable Lyapunov (energy) function by interpreting the model as the dynamic equation for a mass attached to a nonlinear (cubic) spring. In parts (b) and (c), try a simple quadratic Lyapunov function of the form $px^2 + qy^2$, then choose p and q appropriately. In part (d), use the indicated Lyapunov function.]

(a)

$$\dot{x} = y$$
$$\dot{y} = -x^3$$

(b)

$$\dot{x} = -x^3 - y^2$$
$$\dot{y} = xy - y^3$$

(c)

$$x_1(k+1) = \frac{x_2(k)}{1 + x_2^2(k)}$$
$$x_2(k+1) = \frac{x_1(k)}{1 + x_2^2(k)}$$

(d)

$$\dot{x} = y(1-x)$$
$$\dot{y} = -x(1-y)$$
$$V(x,y) = -x - \ln(1-x) - y - \ln(1-y)$$

Chapter 15

External Input-Output Stability

15.1 Introduction

In this lecture, we introduce the notion of external, or input-output, stability for systems. There are many connections between this notion of stability and that of Lyapunov stability which we discussed in the previous two chapters. We will only make the connection in the LTI case. In addition, we will point out the fact that the notion of input-output stability depends in a non-trivial fashion on the way we measure the inputs and the outputs.

15.2 Signal Measures

The signals of interest to us are defined as maps from a time set into $\mathbb{R}^n$. A continuous-time signal is a map from $\mathbb{R} \to \mathbb{R}^n$, and a discrete-time signal is a map from $\mathbb{Z} \to \mathbb{R}^n$. If $n = 1$ we have a scalar signal, otherwise we have a vector-valued signal. It is helpful, in understanding the various signal measures defined below, to visualize a discrete-time signal $w(k)$ as just a *vector* of infinite (or, if our signal is defined only for non-negative time, then a vector of semi-infinite) length or dimension, concretely representing it as the array

$$\begin{pmatrix} \vdots \\ w(0) \\ w(1) \\ \vdots \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} w(0) \\ w(1) \\ \vdots \end{pmatrix} . \tag{15.1}$$

Three of the most commonly used DT signal measures are then natural generalizations of the finite-dimensional vector norms (∞-, 2- and 1-norms) that we have already encountered in earlier chapters, generalized to such infinite-dimensional vectors. We shall examine these three measures, and a fourth that is related to the 2-norm, but is not quite a norm. We shall also define CT signal measures that are natural counterparts of the DT measures.

The signal measures that we study below are:

1. peak magnitude (or ∞-norm);
2. energy (whose square root is the 2-norm);
3. power (or mean energy, whose square root is the "rms" or root-mean-square value);
4. "action" (or 1-norm).

Peak Magnitude: The ∞-Norm

The ∞-norm $\|w\|_\infty$ of a signal is its peak magnitude, evaluated over all signal components and all times:

$$\begin{aligned} \|w\|_\infty &\triangleq \text{max magnitude of } w \\ &\triangleq \sup_k \max_i |w_i(k)| = \sup_k \|w(k)\|_\infty \qquad \text{(for DT systems)} \qquad (15.2) \\ &\triangleq \sup_t \max_i |w_i(t)| = \sup_t \|w(t)\|_\infty \qquad \text{(for CT systems)} \quad , \qquad (15.3) \end{aligned}$$

where $w_i(k)$ indicates the i-th component of the signal vector $w(k)$. Note that $\|w(k)\|_\infty$ denotes the ∞-norm of the signal value *at time* k, i.e. the familiar ∞ norm of an n-vector, namely the maximum magnitude among its components. On the other hand, the notation $\|w\|_\infty$ denotes the ∞-norm of the *entire signal*. The "sup" denotes the *supremum* or *least upper bound*, the value that is approached arbitrarily closely but never (i.e., at any finite time) exceeded. We use "sup" instead of "max" because over an infinite time set the signal magnitude may not have a maximum, i.e. a peak value that is actually attained — consider, for instance, the simple case of the signal

$$1 - \frac{1}{1+|k|} \, ,$$

which does not attain its supremum value of 1 for any finite k.

Note that the DT definition is the natural generalization of the standard ∞-norm for finite-dimensional vectors to the case of our infinite vector in (15.1), while the CT definition is the natural counterpart of the DT definition. This pattern is typical for all the signal norms we deal with, and we shall not comment on it explicitly again.

Example 15.1 Some bounded signals:

(a) For $w(t) = 1$, $t \in \mathbb{R}$, $t \geq 0$:
$\|w\|_\infty = 1$.

(b) For $w(t) = a^t$, $t \in \mathbb{Z}$:
$\|w\|_\infty = \infty$ if $|a| \neq 1$ and $\|w\|_\infty = 1$ otherwise.

The space of all signals with finite ∞-norm are generally denoted by ℓ_∞ and $\mathcal{L}_\infty$ for DT and CT signals respectively. For vector-valued signals, the size of the vector may be explicitly added to the symbol, e.g., ℓ_∞^n. These form normed-vector spaces.

Energy and the 2-Norm

The 2-norm of a signal is the square root of its "energy", which is in turn defined as the sum (in DT) or integral (in CT) of the squares of all components over the entire time set:

$$\begin{aligned} \|w\|_2 &\triangleq \text{square-root of } \textit{energy} \text{ in } w \\ &\triangleq \left[\sum_k w^T(k)w(k)\right]^{\frac{1}{2}} = \left[\sum_k \|w(k)\|_2^2\right]^{\frac{1}{2}} \qquad \text{(for DT systems)} \qquad (15.4) \\ &\triangleq \left[\int w^T(t)w(t)\,dt\right]^{\frac{1}{2}} = \left[\int_t \|w(t)\|_2^2\,dt\right]^{\frac{1}{2}} \qquad \text{(for CT systems)} \quad . \quad (15.5) \end{aligned}$$

Example 15.2 Some examples:

(a) For $w(t) = e^{-at}$ and time set $t \geq 0$, with $a > 0$:
$\|w\|_2 = \frac{1}{\sqrt{2a}} < \infty$

(b) For $w(t) = 1$ and time set $t \geq 0$:
$\|w\|_2 = \infty$

(c) For $w(t) = \cos\omega_o t$ and time set $t \geq 0$:
$\|w\|_2 = \infty$.

These examples suggest that bounded-energy signals go to zero as time progresses. For discrete-time signals, this expectation holds up: if $\|w\|_2 < \infty$, then $\|w(k)\| \longrightarrow 0$ as $k \longrightarrow \infty$. However, for continuous-time signals, the property of having bounded energy does not imply that $\|w(t)\| \longrightarrow 0$ as $t \longrightarrow \infty$, unless additional assumptions are made. This is because continuous-time bounded energy signals can still have arbitrarily large excursions in amplitude, provided these excursions occur over sufficiently narrow intervals of time that the integral of the square remains finite — consider, for instance, a CT signal that is zero everywhere, except for a triangular pulse of height k and base $1/k^4$ centered at every nonzero integer value k. If the continuous-time signal $w(t)$ is differentiable and both w and its derivative $\dot{w}$ have bounded energy (which is *not* the case for the preceding triangular-pulse example), then it *is* true that $\|w(t)\| \longrightarrow 0$ as $t \longrightarrow \infty$. The reader may wish to verify this fact.

It is not hard to show that DT or CT signals with finite 2-norms form a vector space. On the vector space ℓ_2 (respectively $\mathcal{L}_2$) of DT (respectively CT) signals with finite 2-norm, one can define a natural inner product as follows, between signals x and y:

$$\langle x\,,\,y\rangle \triangleq \left[\sum_k x^T(k)y(k)\right] \qquad \text{(for DT systems)} \qquad (15.6)$$

$$\stackrel{\triangle}{=} \quad \int x^T(t)y(t)\,dt \qquad \text{(for CT systems)} \quad . \tag{15.7}$$

(The 2-norm is then just the square root of the inner product of a signal with itself.) These particular infinite-dimensional inner-product vector spaces are of great importance in applications, and are the prime examples of what are known as Hilbert spaces.

Power and RMS Value

Another signal measure of interest is the "power" or mean energy of the signal. One also often deals with the square root of the power, which is commonly termed the "root-mean-square" (or "rms") value. For a signal w for which the following limits exist, we define the power by

$$P_w \stackrel{\triangle}{=} \lim_{N\to\infty}\left[\frac{1}{2N}\sum_{k=-(N-1)}^{N-1} w^T(k)w(k)\right] \qquad \text{(for discrete}-\text{time systems)} \tag{15.8}$$

$$\stackrel{\triangle}{=} \lim_{L\to\infty}\left[\frac{1}{2L}\int_{-L}^{L} w^T(t)w(t)dt\right] \qquad \text{(for continuous}-\text{time systems)} \quad . \tag{15.9}$$

(The above definitions assume that the time set is the entire time axis, but the necessary modifications for other choices of time set should be obvious.) We shall use the symbol ρ_w to denote the rms value, namely $\sqrt{P_w}$. The reason that ρ_w is *not* a norm, according to the technical definition of a norm, is that $\rho_w = 0$ does *not* imply that $w = 0$.

Example 15.3 Some finite-power signals:

(a) For $w(t) = 1$:
$\rho_w = 1$

(b) For $w(t)$ such that $\|w\|_2 < \infty$:
$\rho_w = 0$

(c) For $w(t) = \cos\omega_0 t$ (with $t \in \mathbb{R}$ or $t \in \mathbb{Z}$):
$\rho_w = \frac{1}{\sqrt{2}}$.

Example c) points out an important difference between bounded power and bounded energy signals: unlike bounded energy signals, if $\rho_w < \infty$, the signal doesn't necessarily decay to zero.

As a final comment on the definition of the power of a signal, we elaborate on the hint in the preamble to our definition that the limit required by the definition may not exist for certain signals. The limit of a sequence or function (in our case, the sequence or function is the set of finite-interval rms values, considered over intervals of increasing length) may not exist even if the sequence or function stays bounded, as when it oscillates between two different finite values. The following signal is an example of a CT signal that is bounded but does not have a well-defined power, because the required limit does not exist:

$$w(t) = \begin{cases} 1 & \text{if } t \in [2^{2k}\ \ 2^{2k+1}], \text{ for } k = 0\ 1\ 2\ \ldots \\ 0 & \text{otherwise} \end{cases}$$

Also note that the desired limit may exist, but not be finite. For instance, the limit of a sequence is $+\infty$ if the values of the sequence remain above any chosen finite positive number for sufficiently large values of the index.

Action: The 1-Norm

The 1-norm of a signal is also sometimes termed the "action" of the signal, which is in turn defined as the sum (in DT) or integral (in CT) of the 1-norm of the signal value at each time, taken over the entire time set:

$$\begin{aligned} \|w\|_1 &\triangleq \textit{action of } w \\ &\triangleq \left[\sum_k \|w(k)\|_1\right] \qquad (\text{for discrete} - \text{time systems}) \qquad (15.10) \\ &\triangleq \int \|w(t)\|_1 \, dt \qquad (\text{for continuous} - \text{time systems}) \quad . \qquad (15.11) \end{aligned}$$

Recall that $\|w(k)\|$ for the n-vector $w(k)$ denotes the sum of magnitudes of its components.

The space of all signals with finite 1-norm are generally denoted by ℓ_1 and $\mathcal{L}_1$ for DT and CT signals respectively. These form normed-vector spaces.

We leave you to construct examples that show familiar signals of finite and infinite 1-norm.

Relationships Among Signal Measures

a) If w is a discrete-time sequence, then

$$\|w\|_2 < \infty \Longrightarrow \|w\|_\infty < \infty \qquad (15.12)$$

but

$$\|w\|_2 < \infty \not\Longleftarrow \|w\|_\infty < \infty \qquad (15.13)$$

b) If w is a continuous-time signal, then

$$\|w\|_2 < \infty \not\Longrightarrow \|w\|_\infty < \infty. \qquad (15.14)$$

and

$$\|w\|_2 < \infty \not\Longleftarrow \|w\|_\infty < \infty. \qquad (15.15)$$

c) If $\|w\|_\infty < \infty$, then (when ρ_w exists)

$$\rho_w \leq \|w\|_\infty$$

Item a) is true because of the relationship between energy and magnitude for discrete-time signals. Since the energy of a DT signal is the sum of squared magnitudes, if the energy is bounded, then the magnitude must be bounded. However, the converse is not true —take for example, the signal $w(k) = 1$. As item b) indicates, though, bounded energy implies nothing about the boundedness of magnitude for continuous time signals.

(Many more relationships of the above form can be stated.)

15.3 Input-Output Stability

At this point, it is important to make a connection between the stability of a system and its input-output behavior. The most important notion is that of ℓ_p-stability (p-stability).

Definition 15.1 A system with input signal u and output signal y that is obtained from u through the action of an arbitrary operator H, so $y = H(u)$, is ℓ_p-stable or p-stable ($p = 1 \;\; 2 \;\; \infty$) if there exists a finite $C \in \mathbb{R}$ such that

$$\|y\|_p \leq C\|u\|_p \tag{15.16}$$

for every input u.

A p-stable system is therefore characterized by the requirement that every input of finite p-norm gives rise to an output of finite p-norm. For the case $p = \infty$, this notion is known as Bounded-Input Bounded-Output (BIBO) stability. We will see that BIBO stability is equivalent to p-stability for finite-dimensional LTI state-space systems, but not necessarily in other cases.

Example 15.4 The system described by one integrator:

$$\dot{y} = u$$

is not BIBO stable. A step input is mapped to a ramp which is unbounded. It is not hard to see that this system is not p-stable for any p.

15.3.1 BIBO Stability of LTI Systems

A continuous-time LTI system may be characterized by its impulse response *matrix*, $\mathcal{H}(\,\cdot\,)$, whose (i,j)th entry $h_{ij}(\,\cdot\,)$ is the impulse response from the jth input to the ith output. In other words the input-output relation is given by

$$y(t) = \int \mathcal{H}(t-\tau)u(\tau)d\tau \; .$$

Theorem 15.1 A CT LTI system with m inputs, p outputs, and impulse response matrix $\mathcal{H}(t)$ is BIBO stable if and only if

$$\max_{1\leq i\leq p} \sum_{j=1}^{m} \int |h_{ij}(t)|\, dt < \infty.$$

Proof: The proof of sufficiency involves a straightforward computation of bounds. If u is an input signal that satisfies $\|u\|_\infty < \infty$, i.e. a bounded signal, then we have

$$y(t) = \int \mathcal{H}(t-\tau)u(\tau)d\tau$$

and

$$\begin{aligned}\max_{1\le i\le p} |y_i(t)| &= \max_i \left| \int \sum_{j=1}^{m} h_{ij}(t-\tau)u_j(\tau)\,d\tau \right| \\ &\le \left[\max_i \int \sum_j |h_{ij}(t-\tau)|\,d\tau \right] \max_j \sup_t |u_j(t)|.\end{aligned}$$

It follows that

$$\|y\|_\infty = \sup_t \max_i |y_i(t)| \le \left[\max_i \sum_j \int |h_{ij}(t)|dt \right] \|u\|_\infty < \infty.$$

In order to prove the converse of the theorem, we show that if the above integral is infinite then there exists a bounded input that will be mapped to an unbounded output. Let us consider the case when $p = m = 1$, for notational simplicity (in the general case, we can still narrow the focus to a single entry of the impulse response matrix). Denote the impulse response by $h(t)$ for this scalar case. If the integral

$$\int |h(t)|\,dt$$

is unbounded then given any (large) M there exists an interval of length $2T$ such that

$$\int_{-T}^{T} |h(t)|\,dt > M.$$

Now by taking the input $u_M(t)$ as

$$u_M(t) = \begin{cases} sgn(h(-t)) & -T \le t \le T \\ 0 & |t| > T \end{cases}$$

we obtain an output $y_M(t)$ that satisfies

$$\begin{aligned}\sup_t |y_M(t)| \ge y_M(0) &= \int_{-T}^{T} h(0-\tau)u_M(\tau)\,d\tau \\ &= \int_{-T}^{T} |h(0-\tau)|\,d\tau \\ &> M.\end{aligned}$$

In other words, for any $M > 0$, we can have an input whose maximum magnitude is 1 and whose corresponding output is larger than M. Therefore, there is no finite constant C such that the inequality (24.3) holds.

Further reflection on the proof of Theorem 15.1 reveals that the constant $\|\mathcal{H}\|_1$ defined by

$$\|\mathcal{H}\|_1 = \max_i \sum_j \int |h_{ij}(t)|dt$$

is the smallest constant C that satisfies the inequalty (24.3) when $p = \infty$. This number is called the ℓ_1-norm of $\mathcal{H}(t)$. In the scalar case, this number is just the ℓ_1-norm of $h(\,\cdot\,)$, regarded as a signal.

The discrete-time case is quite similar to continuous-time where we start with a pulse response *matrix*, $\mathcal{H}(\,\cdot\,)$, whose (i,j)th entry $h_{ij}(\,\cdot\,)$ is the pulse response from the jth input to the ith output. The input-output relation is given by

$$y(t) = \sum_\tau \mathcal{H}(t-\tau)u(\tau)\,.$$

Theorem 15.2 A DT LTI system with m inputs, p outputs, and pulse response matrix $\mathcal{H}(t)$ is BIBO stable if and only if

$$\max_{1\le i\le p} \sum_{j=1}^{m} \sum_t |h_{ij}(t)| < \infty.$$

In addition, the constant $\|\mathcal{H}\|_1$ defined by

$$\|\mathcal{H}\|_1 = \max_i \sum_j \sum_t |h_{ij}(t)|$$

is the smallest constant C that satisfies the inequalty (24.3) when $p = \infty$. We leave the proof of these facts to the reader.

Application to finite-dimensional State-Space Models

Now consider the application to the following causal CT LTI system in state-space form (and hence of finite order):

$$\dot{x} = Ax + Bu \tag{15.17}$$
$$y = Cx + Du \tag{15.18}$$

The impulse response of this system is given by

$$\mathcal{H}(t) = Ce^{At}B + D\delta(t) \;\text{ for } t \ge 0$$

which has Laplace transform

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

The system (15.18) is BIBO stable if and only if the poles of $H(s)$ are in the open left half plane. (We leave the proof to you.) This is in turn guaranteed if the system is asymptotically stable, i.e. if A has all its eigenvalues in the open left half plane.

Example 15.5 BIBO Stability Doesn't Imply Asymptotic Stability

It is possible that a system be BIBO stable and not asymptotically stable. Consider the system

$$\begin{aligned} \dot{x} &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 1 \end{pmatrix} u \\ y &= \begin{pmatrix} 1 & -1 \end{pmatrix} x \end{aligned}$$

This system is not stable since A has an eigenvalue at 1. Nevertheless, thanks to a pole-zero cancellation, the only pole that $H(s)$ has is at -1, so the system is BIBO stable. We shall have much more to say about such cancellations in the context of reachability, observability, and minimality (the example here turns out to be unobservable).

Marginal stability of an LTI system, i.e., stability in the sense of Lyapunov but without asymptotic stability, is not sufficient to guarantee BIBO stability. For instance, consider a simple integrator, whose transfer function is $1/s$.

Time-Varying and Nonlinear Systems

Although there are results connecting Lyapunov stability with I/O stability for general time-varying and nonlinear systems, they are not as powerful as the linear time-invariant case. In particular, systems may be I/O stable with respect to one norm and not stable with respect to another. Below are some examples illustrating these facts.

Example 15.6 A Time-Varying System

Consider the time-varying DT system given by:

$$y(t) = H(u)(t) = u(0).$$

H is obviously ∞-stable with gain less than 1. However, it is not 2-stable.

Example 15.7 A Nonlinear System

Consider the nonlinear system given by:

$$\dot{x} = -x + e^x u, \qquad y = x.$$

The unforced system is linear and is asymptotically stable. On the other hand the system is not I/O stable. To see this, consider the input $u(t) = 1$. Since $e^x > x$, $\dot{x}$ is always strictly positive, indicating that x is strictly increasing. Hence, for a bounded input, the output is not bounded.

15.3.2 p-Stability of LTI Systems (optional)

In this section we will continue our analysis of the p-stability of systems described through input-output relations. Let us start with the continuous-time case, and restrict ourselves to single-input single-output. The input $u(t)$ is related to the output $y(t)$ by

$$y(t) = \int h(t-\tau)u(\tau)d\tau$$

where $h(t)$ is the impulse response. The following theorem shows that the constant C in 24.3 is always bounded above by $\|h\|_1$.

Theorem 15.3 If $\|h\|_1 < \infty$ and $\|u\|_p < \infty$ then $\|y\|_p < \infty$ and furthermore

$$\|y\|_p \leq \|h\|_1 \|u\|_p \ .$$

Proof: In Theorem 15.1 we have already established this result for $p = \infty$. In what follows $p = 1$ 2. The output $y(t)$ satisfies

$$|y(t)|^p = |(h * u)(t)|^p = \left| \int_{-\infty}^{\infty} h(t-\tau)u(\tau)\ d\tau \right|^p \leq \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ |u(\tau)|\ d\tau \right)^p$$

therefore,

$$\|h * u\|_p^p = \int_{-\infty}^{\infty} |(h * u)(t)|^p\ dt \leq \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ |u(\tau)|\ d\tau \right)^p\ dt \ .$$

Next we analyze the inner integral

$$\begin{aligned} \int_{-\infty}^{\infty} |h(t-\tau)|\ |u(\tau)|\ d\tau &= \int_{-\infty}^{\infty} |h(t-\tau)|^{1/q}\ |h(t-\tau)|^{1/p}\ |u(\tau)|\ d\tau \\ &\leq \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ d\tau \right)^{1/q} \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ |u(\tau)|^p\ d\tau \right)^{1/p} \end{aligned}$$

where the last inequality follows from Minkowski's inequalities, and $\frac{1}{p} + \frac{1}{q} = 1$. Hence,

$$\begin{aligned} \|h * u\|_p^p &\leq \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ d\tau \right)^{p/q} \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ |u(\tau)|^p\ d\tau \right)\ dt \\ &= \int_{-\infty}^{\infty} (\|h\|_1)^{p/q} \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ |u(\tau)|^p\ d\tau \right)\ dt \\ &= \|h\|_1^{p/q} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |h(t-\tau)|\ |u(\tau)|^p\ d\tau\ dt \\ &= \|h\|_1^{p/q} \int_{-\infty}^{\infty} |u(\tau)|^p \left(\int_{-\infty}^{\infty} |h(t-\tau)|\ dt \right)\ d\tau \\ &= \|h\|_1^{p/q+1} \int_{-\infty}^{\infty} |u(\tau)|^p\ d\tau \\ &= \|h\|_1^p\ \|u\|_p^p \end{aligned}$$

Therefore

$$\|h * u\|_p \leq \|h\|_1 \, \|u\|_p \, .$$

Recall that when $p = \infty$, $\|h\|_1$ was the smallest constant for which the inequality $\|y\|_p \leq C\|u\|_p$ for all u. This is not the case for $p = 2$, and we will see later that a smaller constant can be found. We will elaborate on these issues when we discuss systems' norms later on in the course. The discrete-time case follows in exactly the same fashion.

Example 15.8 For a finite-dimensional state-space model, a system H is p-stable if and only if all the poles of of $H(s)$ are in the LHP. This coincides with BIBO stability.

Exercises

Exercise 15.1 Non-causal Systems In this chapter, we only focused on causal operators, although the results derived were more general. As an example, consider a particular CT LTI system with a bi-lateral Laplace transform:

$$G(s) = \frac{s+2}{(s-2)(s+1)}\,.$$

(a) Check the *p-stability* and *causality* of the system in the following cases:

(i) the ROC (Region of Convergence) is $R_1 = \{s \in \mathbb{C} |\ Re(s) < -1\}$ where $Re(s)$ denotes the real part of s;

(ii)the ROC is $R_2 = \{s \in \mathbb{C} |\ -1 < Re(s) < 2\}$;

(iii) the ROC is $R_3 = \{s \in \mathbb{C} |\ Re(s) > 2\}$.

(b) In the cases where the system is not *p-stable* for $p = 2$ and $p = \infty$, find a bounded input that makes the output unbounded, i.e., find an input $u \in L_p$ that produces an output $y \notin L_p$, for $p = 2,\ \infty$.

Exercise 15.2 In nonlinear systems, p-stability may be satisfied in only a local region around zero. In that case, a system will be locally p-stable if:

$$\|Gu\|_p \leq C\|u\|_p, \qquad \text{for all } u \text{ with } \|u\|_p \leq \delta$$

Consider the system:

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

Where g is a continuous function on $[-T, T]$. Which of the following systems is p-stable, locally p-stable or unstable for $p \geq 1$:

(a) $g(x) = \cos x$.

(b) $g(x) = \sin x$.

(c) $g(x) = \mathrm{Sat}(x)$ where

$$\mathrm{Sat}(x) = \begin{cases} x & |x| \leq 1 \\ 1 & |x| \geq 1 \end{cases}$$

Chapter 17

Interconnected Systems and Feedback: Well-Posedness, Stability, and Performance

17.1 Introduction

Feedback control is a powerful approach to obtaining systems that are stable and that meet performance specifications, despite system disturbances and model uncertainties. To understand the fundamentals of feedback design, we will study system interconnections and some associated notions such as well-posedness and external stability. Unless otherwise noted, our standing *assumption* for the rest of the course — and a natural assumption in the control setting — will be that *all our models* for physical systems *have outputs that depend* **causally** *on their inputs.*

17.2 System Interconnections

Interconnections are very common in control systems. The system or process that is to be controlled — commonly referred to as the **plant** — may itself be the result of interconnecting various sorts of subsystems in series, in parallel, and in feedback. In addition, the plant is interfaced with sensors, actuators and the control system. Our model for the overall system represents all of these components in some idealized or nominal form, and will also include components introduced to represent uncertainties in, or neglected aspects of the nominal description.

We will start with the simplest feedback inteconnection of a plant with a controller, where the outputs from the plant are fed into a controller whose own outputs are in turn fed

back as inputs to the plant. A diagram of this prototype feedback control configuration is shown in Figure 17.1.

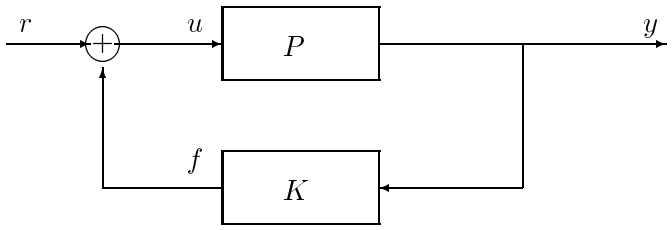


Figure 17.1: Block diagram of the prototype feedback control configuration.

The plant P and controller K could in general be nonlinear, time-varying, and infinite-dimensional, but we shall restrict attention almost entirely to **interconnections of finite-order LTI components**, whether described in state-space form or simply via their input-output transfer functions. Recall that the transfer functions of such finite-order state-space models are *proper rationals*, and are in fact *strictly proper* if there is no direct feedthrough from input to output. We shall use the notation of CT systems in the development that follows, although everything applies equally to DT systems.

The plant and controller should evidently have compatible input/output dimensions; if not, then they cannot be tied together in a feedback loop. For example, if $P(s)$ is the $p \times m$ transfer function matrix of the (nominal LTI model of the) plant in Figure 17.1, then the transfer function $K(s)$ of the (LTI) controller should be an $m \times p$ matrix.

All sorts of other feedback configurations exist; two alternatives can be found in Figures 17.2 and 17.3. For our purposes in this chapter, the differences among these various configurations are not important.

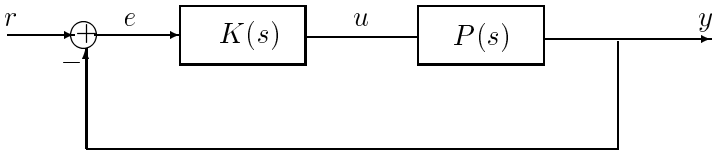


Figure 17.2: A ("servo") feedback configuration where the tracking error between the command r and output y is directly applied to the controller.

Our discussion for now will focus on the arrangement shown in Figure 17.4, which is an elaboration of Figure 17.1 that represents some additional signals of interest. Interpretations for the various (vector) signals depicted in the preceding figures are normally as follows:

- u — control inputs to plant

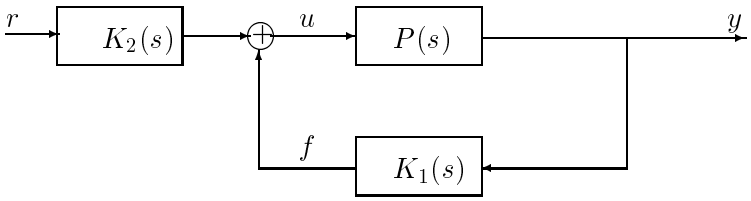


Figure 17.3: A two-parameter-compensator feedback scheme.

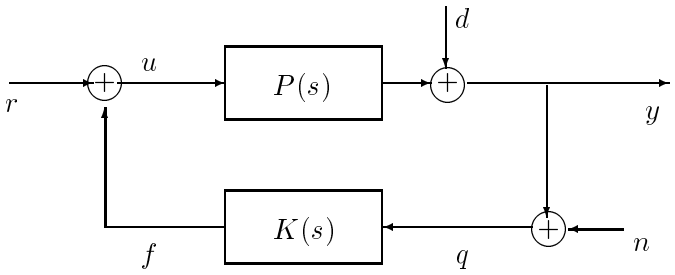


Figure 17.4: Including plant disturbances d and measurement noise n.

- y — measured outputs of plant
- d — plant disturbances, represented as acting at the output
- n — noise in the output measurements used by the feedback controller
- r — reference or command inputs
- e — tracking error $r - y$.
- f — output of feedback compensator

Transfer Functions

We now show how to obtain the transfer functions of the mappings relating the various signals found in Figure 17.4; the transform argument, s, is omitted for notational simplicity. We also depart temporarily from our convention of denoting transforms by capitals, and mark the transforms of all signals by lower case, saving upper case for transfer function matrices (i.e. transforms of impulse responses); this distinction will help the eye make its way through the expressions below, and should cause no confusion if it is kept in mind that *all quantities below are transforms.* To begin by relating the plant output to the various input signals, we can

write

$$\begin{aligned} y &= Pu + d \\ &= P[r + K(y + n)] + d \\ (I - PK)y &= Pr + PKn + d \\ y &= (I - PK)^{-1}Pr + (I - PK)^{-1}PKn + (I - PK)^{-1}d \end{aligned}$$

Similarly, the control input to the plant can be written as

$$\begin{aligned} u &= r + K(y + n) \\ &= r + K(Pu + d + n) \\ (I - KP)u &= r + Kn + Kd \\ u &= (I - KP)^{-1}r + (I - KP)^{-1}Kn + (I - KP)^{-1}Kd \end{aligned}$$

The map $u \longrightarrow f$ (with the feedback loop open and $r = 0$, $n = 0$, $d = 0$) is given by $L = KP$, and is called the *loop transfer function.*

The map $d \longrightarrow y$ (with $n = 0$, $r = 0$) is given by $S_o = (I - PK)^{-1}$ and is called the *output sensitivity function.*

The map $n \longrightarrow y$ (with $d = 0$, $r = 0$) is given by $T = (I - PK)^{-1}PK$ and is called the *complementary sensitivity function.*

The map $r \longrightarrow u$ (with $d = 0$, $n = 0$) is given by $S_i = (I - KP)^{-1}$ and is called the *input sensitivity function.*

The map $r \longrightarrow y$ ($d = 0$, $n = 0$) is given by $(I - PK)^{-1}P$ is called the *system response function* .

The map $d \longrightarrow u$ (with $n = 0$, $r = 0$) is given by $(I - KP)^{-1}K$.

Note that the transfer function $(I - KP)^{-1}K$ can also be written as $K(I - PK)^{-1}$, as may be proved by rearranging the following identity:

$$(I - KP)K = K(I - PK) \quad ,$$

Similarly the transfer function $(I - PK)^{-1}P$ can be written as $P(I - KP)^{-1}$.

Note also that the output sensitivity and input sensitivity functions are different, because, except for the case when P and K are both single-input, single-output (SISO), we have

$$(I - KP)^{-1} \neq (I - PK)^{-1}.$$

17.3 Well-Posedness

We will restrict attention to the feedback structure in Figure 17.5. Our assumption is that H_1 and H_2 have some underlying state-space descriptions with inputs u_1, u_2 and outputs y_1, y_2, so their transfer functions $H_1(s)$ and $H_2(s)$ are *proper*, i.e. $H_1(\infty)$, $H_2(\infty)$ are finite. It is possible (and in fact typical for models of physical systems, since their response falls off to zero as one goes higher in frequency) that the transfer function is in fact *strictly* proper.

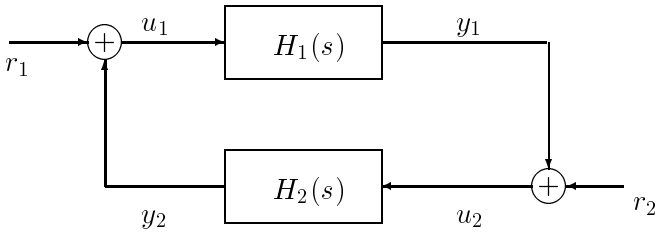


Figure 17.5: Feedback Interconnection.

The closed-loop system in Figure 17.5 can now be described in state-space form by writing down state-space descriptions for $H_1(s)$ (with input u_1 and output y_1) and $H_2(s)$ (with input u_2 and output y_2), and combining them according to the interconnection constraints represented in Figure 17.5. Suppose our state-space models for H_1 and H_2 are

$$H_1 \sim \left[\begin{array}{cc} A_1 & B_1 \\ C_1 & D_1 \end{array}\right], \quad H_2 \sim \left[\begin{array}{cc} A_2 & B_2 \\ C_2 & D_2 \end{array}\right]$$

with respective state vectors, inputs, and outputs (x_1, u_1, y_1) and (x_2, u_2, y_2), so

$$\begin{aligned} \dot{x}_1 &= A_1 x_1 + B_1 u_1 \\ y_1 &= C_1 x_1 + D_1 u_1 \\ \dot{x}_2 &= A_2 x_2 + B_2 u_2 \\ y_2 &= C_2 x_2 + D_2 u_2 \,. \end{aligned} \tag{17.1}$$

Note that $D_1 = H_1(\infty)$ and $D_2 = H_2(\infty)$. The interconnection constraints are embodied in the following set of equations:

$$\begin{aligned} u_1 &= r_1 + y_2 = r_1 + C_2 x_2 + D_2 u_2 \\ u_2 &= r_2 + y_1 = r_2 + C_1 x_1 + D_1 u_1, \end{aligned}$$

which can be rewritten compactly as

$$\left[\begin{array}{cc} I & -D_2 \\ -D_1 & I \end{array}\right] \left[\begin{array}{c} u_1 \\ u_2 \end{array}\right] = \left[\begin{array}{cc} 0 & C_2 \\ C_1 & 0 \end{array}\right] \left[\begin{array}{c} x_1 \\ x_2 \end{array}\right] + \left[\begin{array}{cc} I & 0 \\ 0 & I \end{array}\right] \left[\begin{array}{c} r_1 \\ r_2 \end{array}\right]. \tag{17.2}$$

We shall label the interconnected system **well-posed** if the internal signals of the feedback loop, namely u_1 and u_2, are *uniquely* defined for *every* choice of the system state variables x_1, x_2 and external inputs r_1, r_2. (Note that the other internal signals, y_1 and y_2, will be uniquely defined under these conditions if and only if u_1 and u_2 are, so we just focus on the latter pair.) It is evident from (17.2) that the condition for this is the invertibility of the matrix

$$\begin{bmatrix} I & -D_2 \\ -D_1 & I \end{bmatrix}. \tag{17.3}$$

This matrix is invertible if and only if

$$I - D_1 D_2 \text{ or equivalently } I - D_2 D_1 \text{ is invertible.} \tag{17.4}$$

This result follows from the fact that if X, Y, W, and Z are matrices of compatible dimensions, and X is invertible then

$$\det \begin{bmatrix} X & Y \\ Z & W \end{bmatrix} = \det(X) \det(W - ZX^{-1}Y) \tag{17.5}$$

A *sufficient* condition for (17.4) to hold is that either H_1 or H_2 (or both) be *strictly* proper; that is, either $D_1 = 0$ or $D_2 = 0$.

The significance of well-posedness is that once we have solved (17.2) to determine u_1 and u_2 in terms of x_1, x_2, r_1 and r_2, we can eliminate u_1 and u_2 from (17.1) and arrive at a state-space description of the closed-loop system, with state vector

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

We leave you to write down this description explicitly. Without well-posedness, u_1 and u_2 would not be well-defined for arbitrary x_1, x_2, r_1 and r_2, which would in turn mean that there could not be a well-defined state-space representation of the closed-loop system.

The condition in (17.4) is equivalent to requiring that

$$\Big(I - H_1(s)H_2(s)\Big)^{-1} \text{ or equivalently } \Big(I - H_2(s)H_1(s)\Big)^{-1} \text{ exists and is proper.} \tag{17.6}$$

Example 17.1 Consider a discrete-time system with $H_1(z) = 1$ and $H_2(z) = 1 - z^{-1}$ in (the DT version of) Figure 17.5. In this case $(1 - H_1(\infty)H_2(\infty)) = 1 - 1 = 0$, and thus the system is **ill-posed**. Note that the transfer function from r_1 to y_1 for this system is

$$(1 - H_1 H_2)^{-1} H_1 = (1 - 1 + z^{-1})^{-1} = z$$

which is not proper — it actually corresponds to the noncausal input-output relation

$$y_1(k) = r_1(k+1) \quad ,$$

which cannot be modeled by a state-space description.

Example 17.2 Again consider Figure 17.4, with $H_1(s) = \frac{s+1}{s+2}$ and $H_2(s) = \frac{s+2}{s+1}$. The expression $(1 - H_1(\infty)H_2(\infty)) = 0$, which implies that the interconnection is **ill-posed**. In this case notice that,

$$\begin{aligned}(1 - H_1(s)H_2(s)) &= 1 - 1 \\ &= 0 \quad \forall\ s \in \mathbb{C} \quad !\end{aligned}$$

Since the inverse of $(1 - H_1H_2)$ does not exist, the transfer functions relating external signals to internal signals cannot be written down.

17.4 External Stability

The inputs in Figure 17.5 are related to the signals y_1, and y_2 as follows:

$$\begin{aligned}y_1 &= H_1(y_2 + r_1) \\ y_2 &= H_2(y_1 + r_2),\end{aligned}$$

which can be written as

$$\begin{bmatrix} I & -H_1 \\ -H_2 & I \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix} \begin{bmatrix} r_1 \\ r_2 \end{bmatrix} \tag{17.7}$$

We assume that the interconnection in Figure 17.5 is *well-posed*. Let the map $\mathcal{T}(H_1, H_2)$ be defined as follows:

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \mathcal{T}(H_1, H_2) \begin{pmatrix} r_1 \\ r_2 \end{pmatrix}.$$

From the relations 17.7 the form of the map $\mathcal{T}(H_1, H_2)$ is given by

$$\mathcal{T}(H_1, H_2) = \begin{bmatrix} (I - H_1H_2)^{-1}H_1 & (I - H_1H_2)^{-1}H_1H_2 \\ (I - H_2H_1)^{-1}H_2H_1 & (I - H_2H_1)^{-1}H_2 \end{bmatrix}$$

We term the interconnected system **externally p-stable** if the map $\mathcal{T}(H_1, H_2)$ is p-stable. In our finite-order LTI case, what this requires is precisely that the poles of all the entries of the rational matrix

$$\mathcal{T}(H_1, H_2) = \begin{bmatrix} (I - H_1H_2)^{-1}H_1 & (I - H_1H_2)^{-1}H_1H_2 \\ (I - H_2H_1)^{-1}H_2H_1 & (I - H_2H_1)^{-1}H_2 \end{bmatrix}$$

be in the open left half of the complex plane.

External stability guarantees that bounded inputs r_1, and r_2 will produce bounded responses y_1, y_2, u_1, and u_2. External stability is guaranteed by asymptotic stability (or **internal stability**) of the state-space description obtained through the process described in our discussion of well-posedness. However, as noted in earlier chapters, it is possible to have external stability of the interconnection without asymptotic stability of the state-space description

(because of hidden unstable modes in the system — an issue that will be discussed much more in later chapters). On the other hand, external stability is stronger than input/output stability of the mapping $(I - H_1H_2)^{-1}H_1$ between r_1 and y_1, because this mapping only involves a subset of the exposed or external variables of the interconnection.

Example 17.3 Assume we have the configuration in Figure 17.5, with $H_1 = \frac{s-1}{s+1}$ and $H_2 = -\frac{1}{s-1}$. The transfer function relating r_1 to y_1 is

$$\begin{aligned} \frac{H_1}{1 - H_1H_2} &= \frac{s-1}{s+1}\left(1 + \frac{1}{s+1}\right)^{-1} \\ &= \left(\frac{s-1}{s+1}\right)\left(\frac{s+1}{s+2}\right) \\ &= \frac{s-1}{s+2} \quad . \end{aligned}$$

Since the only pole of this transfer function is at $s = -2$, the input/output relation between r_1 and y_1 is stable. However, consider the transfer function from r_2 to u_1, which is

$$\begin{aligned} \frac{H_2}{1 - H_1H_2} &= \frac{1}{s-1}\left(\frac{1}{1 + \frac{1}{s+1}}\right) \\ &= \frac{s+1}{(s-1)(s+2)} \quad . \end{aligned}$$

This transfer function is unstable, which implies that the closed-loop system is externally unstable.

17.5 A More General Description

There are at least two reasons for going to a more general system description than those shown up to now. First, our assessment of the performance of the system may involve variables that are not among the measured/fed-back output signals of the plant. Second, the disturbances affecting the system may enter in more general ways than indicated previously. We do still want our system representation to separate out the controller portions of the system (the K's or K_1, K_2 of the earlier figures), as these are the portions that we will be designing. In this section we will introduce a general plant description that organizes the different types of inputs and outputs, and their interaction with a controller. A block diagram for a general plant description is shown in Figure 17.6.

The different signals in Figure 17.6 can be classified as follows.

- Inputs:

  1. Control input vector u, which contains the actuator signals driving the plant and generated by a controller.

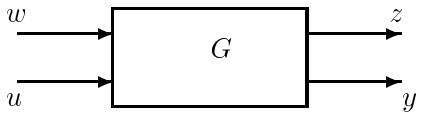


Figure 17.6: General plant description.

2. Exogeneous input vector w, which contains all other external signals, such as references and disturbances.

- Outputs:

1. Measured output vector y, which contains the signals that are available to the controller. These are based on the outputs of the sensor devices, and form the input to the controller.
2. Regulated output vector z, which contains the signals that are important for the specific application. The regulated outputs usually include the actuator signals, the tracking error signals, and the state variables that must be manipulated.

Let the transfer function matrix

$$G = \begin{bmatrix} G_{zw} & G_{zu} \\ G_{yw} & G_{yu} \end{bmatrix},$$

have the state-space realization

$$\begin{aligned} \dot{x} &= Ax + B_1 w + B_2 u \\ z &= C_1 x + D_{11} w + D_{12} u \\ y &= C_2 x + D_{21} w + D_{22} u \end{aligned}$$

Example 17.4 Consider the unity feedback system in Figure 17.7, where P is a SISO plant, K is a scalar controller, y' is the output, u is the control input, v is a reference signal, and d is an external disturbance that is "shaped" by the filter H before it is injected into the measured output. The controller is driven by the difference $e = v - y'$ (the "tracking error"). The signals v and d can be taken to constitute the exogeneous input, so

$$w = \begin{bmatrix} v \\ d \end{bmatrix}.$$

In such a configuration we typically want to keep the tracking error e small, and to put a cost on the control action. We can therefore take the regulated output z to be

$$z = \begin{bmatrix} e \\ u \end{bmatrix}.$$

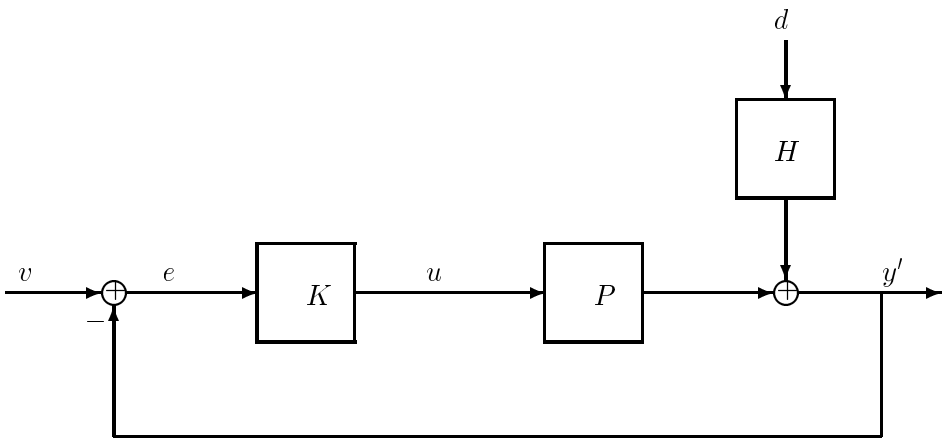


Figure 17.7: Example of a unity feedback system.

The input to the controller is e, therefore we set the measured output y to be equal to e. With these choices, the generalized plant transfer function G, which relates z and y to w and u, can be obtained from

$$\begin{aligned} z &= \begin{bmatrix} -Pu - Hd + v \\ u \end{bmatrix} = \begin{bmatrix} -P \\ 1 \end{bmatrix} u + \begin{bmatrix} 1 & -H \\ 0 & 0 \end{bmatrix} w \\ y &= -Pu + \begin{bmatrix} 1 & -H \end{bmatrix} w. \end{aligned}$$

Let us suppose that $P = \frac{1}{s-1}$ and $H = \frac{1}{s+1}$. Then a state-space realization of G is easily obtained:

$$\begin{aligned} \frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} w + \begin{bmatrix} 1 \\ 0 \end{bmatrix} u \\ z &= \begin{bmatrix} -1 & -1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} w + \begin{bmatrix} 0 \\ 1 \end{bmatrix} u \\ y &= \begin{bmatrix} -1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \end{bmatrix} w + 0u\,. \end{aligned}$$

If we close the loop, the general plant/controller structure takes the form shown in Figure 17.8.

The plant transfer matrix G is a 2×2 block matrix mapping the inputs w, u to the outputs z, y, where the part of the plant that interacts directly with the controller is just G_{yu}. The map (or transfer function) of interest in performance specifications is the map from w to z, denoted by Φ, and easily seen to be given by the following expression:

$$\Phi = G_{zw} + G_{zu}(I - KG_{yu})^{-1}KG_{yw} \tag{17.8}$$

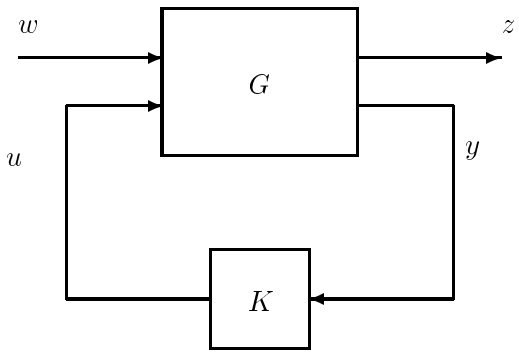


Figure 17.8: A general feedback configuration.

In this new settup we would like to determine under what conditions the closed-loop system in Figure 17.9 is **well-posed** and **externally stable**. For these purposes we inject signals r and v as shown in Figure 17.9, which is similar to what we did in the previous sections. Note that by defining the signals

$$r_1 = \begin{pmatrix} w \\ r \end{pmatrix} \quad r_2 = \begin{pmatrix} 0 \\ v \end{pmatrix}$$
$$y_1 = \begin{pmatrix} z \\ y \end{pmatrix} \quad y_2 = \begin{pmatrix} 0 \\ f \end{pmatrix}$$

this structure is equivalent to the structure in Figure 17.5. This is illustrated in Figure 17.10, with

$$H_1 = \begin{bmatrix} G_{zw} & G_{zu} \\ G_{yw} & G_{yu} \end{bmatrix}$$
$$H_2 = \begin{bmatrix} 0 \\ I \end{bmatrix} K \begin{bmatrix} 0 & I \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & K \end{bmatrix}$$

This interconnection is **well-posed** if and only if

$$\left(I - \begin{pmatrix} G_{zw}(\infty) & G_{zu}(\infty) \\ G_{yw}(\infty) & G_{yu}(\infty) \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & K(\infty) \end{pmatrix} \right)$$

is invertible. This is the same as requiring that

$$(I - K(s)G_{yu}(s))^{-1} \text{ or equivalently } (I - G_{yu}(s)K(s))^{-1} \text{ exists and is proper}$$

The inputs in Figure 17.9 are related to the signals z, u and y as follows:

$$\begin{bmatrix} I & -G_{zu} & 0 \\ 0 & I & -K \\ 0 & -G_{yu} & I \end{bmatrix} \begin{bmatrix} z \\ u \\ y \end{bmatrix} = \begin{bmatrix} G_{zw} & 0 & 0 \\ 0 & I & K \\ G_{yw} & 0 & 0 \end{bmatrix} \begin{bmatrix} w \\ r \\ v \end{bmatrix} \tag{17.9}$$

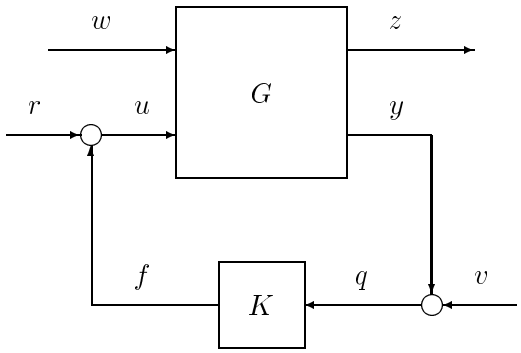


Figure 17.9: A more general feedback configuration.

Let the map $\mathcal{T}(P,K)$ be defined as follows:

$$\begin{pmatrix} z \\ u \\ y \end{pmatrix} = \mathcal{T}(P,K) \begin{pmatrix} w \\ r \\ v \end{pmatrix}$$

The interconnected system is **externally p-stable** if the map from r_1, r_2 to y_1, y_2 is p-stable, see Figure 17.10. This is equivalent to requiring that the map $\mathcal{T}(P,K)$ is p-stable.

17.6 Obtaining Stability and Performance: A Preview

In the lectures ahead we will be concerned with developing analysis and synthesis tools for studying stability and performance in the presence of plant uncertainty and system disturbances.

Stabilization

Stabilization is the first requirement in control design — without stability, one has nothing! There are two relevant notions of stability:

(a) nominal stability (stability in the absence of modeling errors), and

(b) robust stability (stability in the presence of some modeling errors).

In the previous sections, we have shown that stability analysis of an interconnected feedback system requires checking the stability of the closed-loop operator, $\mathcal{T}(P,K)$. In the case where

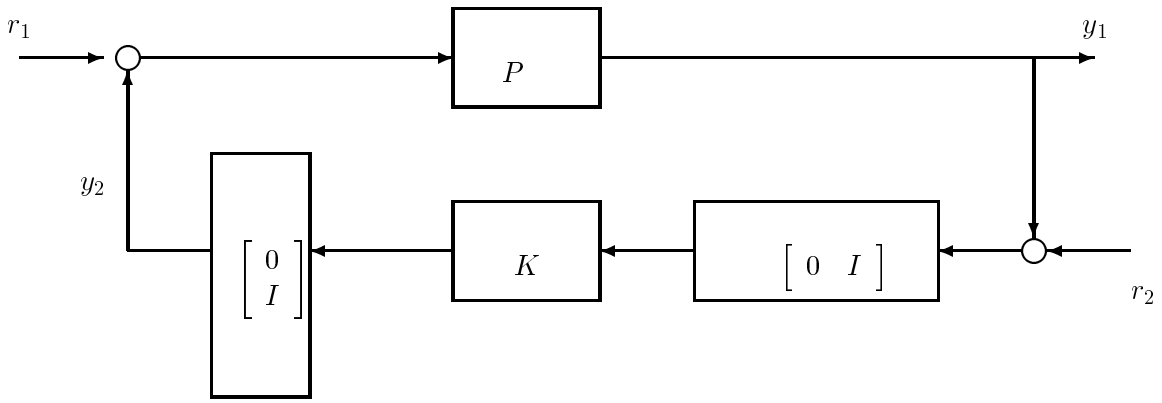


Figure 17.10: A more general feedback configuration.

modeling errors are present, such a check has to be done for every possible perturbation of the system. Efficient methods for performing this check for specified classes of modeling errors are necessary.

Meeting Performance Specifications

Performance specifications (once stability has been ensured) include disturbance rejection, command following (*i.e.,* tracking), and noise rejection. Again, we consider two notions of performance:

(a) nominal performance (performance in the absence of modeling errors), and

(b) robust performance (performance in the presence of modeling errors).

Many of the performance specifications that one may want to impose on a feedback system can be classified under the following two types of specifications:

1. Disturbance Rejection. This corresponds to minimizing the effect of the exogenous inputs w on the regulated variables z in the general 2-input 2-output description, when the exogenous inputs are only partially known. To address this problem, it is necessary to provide a model for the exogenous variables. One possibility is to assume that w has finite energy but is otherwise unknown. If we desire to minimize the energy in the z produced by this w, we can pose the performance task as involving the minimization of

$$\sup_{w \neq 0} \frac{\|\Phi w\|_2}{\|w\|_2}$$

where Φ is the map relating w to z. This is just the square root of the energy-energy gain, and is measured by the $\mathcal{H}_\infty$-norm of Φ.

Alternatively, if w is assumed to have finite peak magnitude, and we are interesed in the peak magnitude of the regulated output z, then the measure of performance is given by the peak-peak gain of the system, which is measured by the $\ell_1/\mathcal{L}_1$-norm of Φ. Other alternatives such as power-power amplification can be considered.

A rather different approach, and one that is quite powerful in the linear setting, is to model w as a stochastic process (e.g, white noise process). By measuring the variance of z, we obtain a peformance measure on Φ.

2. Fixed-Input Specifications. These specifications are based on a specific command or nominal trajectory. One can, for instance, specify a template in the time-domain within which the output is required to remain for a given class of inputs. Familiar specifications such as overshoot, undershoot, and settling time for a step input fall in this category.

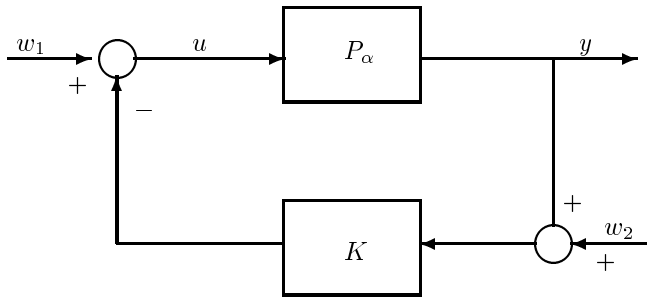
Finally, conditions for checking whether a system meets a given performance measure in the presence of prescribed modeling errors have to be developed. These topics will be revisited later on in this course.

Exercises

Exercise 17.1 Let $P(s) = e^{-2s} - 1$ be connected in a unity feedback configuration. Is this system well-posed?

Exercise 17.2 Assume that P_α and K in the diagram are given by:

$$P_\alpha(s) = \begin{pmatrix} \frac{s}{s+1} & \frac{-\alpha}{s+1} \\ \frac{1}{(s+1)} & \frac{1}{s+1} \end{pmatrix}, \quad \alpha \in \mathbb{R}, \qquad K(s) = \begin{pmatrix} \frac{s+1}{s(s+5)} & 0 \\ -\frac{s+1}{s(s+5)} & \frac{s+1}{s+5} \end{pmatrix}.$$



1. Is the closed loop system stable for all $\alpha > 0$?
2. Is the closed loop system stable for $\alpha = 0$?

Exercise 17.3 Consider the standard servo loop, with

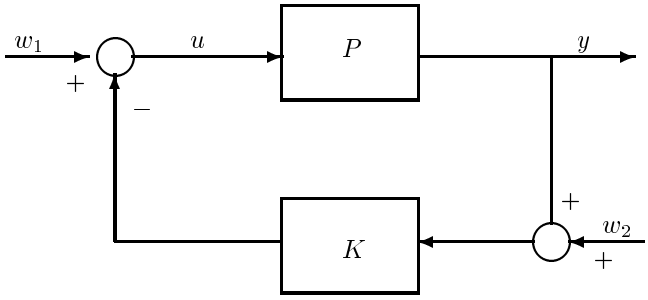
$$P(s) = \frac{1}{10s + 1}, \quad K(s) = k$$

but with no measurement noise. Find the least positive gain such that the following are *all* true:

- The feedback system is internally stable.
- With no disturbance at the plant output ($d(t) \equiv 0$), and with a unit step on the command signal $r(t)$, the error $e(t) = r(t) - y(t)$ settles to $|e(\infty)| \leq 0.1$.
- Show that the $\mathcal{L}_2$ to $\mathcal{L}_\infty$ induced norm of a SISO system is given by $\mathcal{H}_2$ norm of the system.
- With zero command ($r(t) \equiv 0$), $\|y\|_\infty \leq 0.1$ for all $d(t)$ such $\|d\|_2 \leq 1$. [ADD NEW Problem]

Exercise 17.4 Parametrization of Stabilizing Controllers

Consider the diagram shown below where P is a given stable plant. We will show a simple way of parametrizing all stabilizing controllers for this plant. The plant as well as the controllers are finite dimensional.



1. Show that the feedback controller

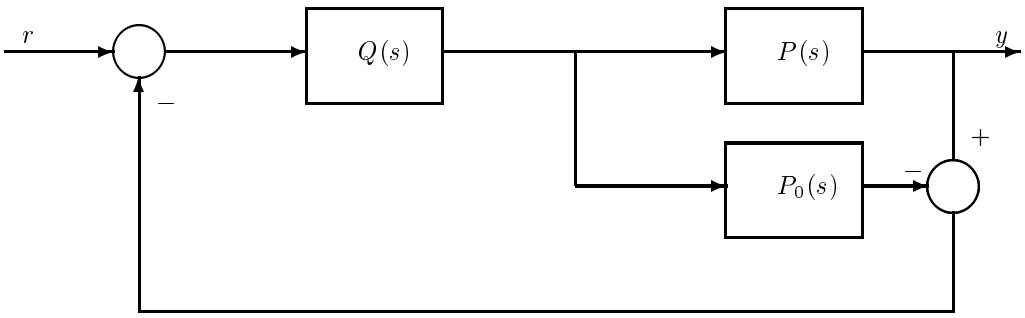
$$K = Q(I - PQ)^{-1} = (I - QP)^{-1}Q$$

for any stable rational Q is a stabilizing controller for the closed loop system.

2. Show that every stabilizing controller is given by $K = Q(I - PQ)^{-1}$ for some stable Q. (Hint: Express Q in terms of P and K).

3. Suppose P is SISO, w_1 is a step, and $w_2 = 0$. What conditions does Q have to satisfy for the steady state value of u to be zero. Is it always possible to satisfy this condition?

Exercise 17.5 Consider the block diagram shown in the figure below.

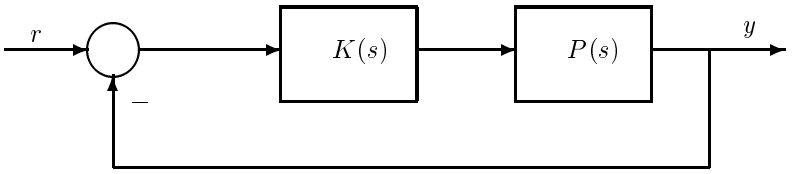


(a) Suppose $P(s) = \dfrac{2}{s-1}$, $P_0(s) = \dfrac{1}{s-1}$ and $Q = 2$. Calculate the transfer function from r to y.

(b) Is the above system internally stable?

(c) Now suppose that $P(s) = P_0(s) = H(s)$ for some $H(s)$. Under what conditions on $H(s)$ is the system internally stable for any *stable* (but otherwise arbitrary) $Q(s)$?

Exercise 17.6 Consider the system shown in the figure below.



The plant transfer function is known to be given by:

$$P(s) = \begin{bmatrix} \frac{s-1}{s+1} & 1 \\ 0 & \frac{s+1}{s+2} \end{bmatrix}$$

A control engineer designed the controller $K(s)$ such that the closed-loop transfer function from r to y is:

$$H(s) = \begin{bmatrix} \frac{1}{s+4} & 0 \\ 0 & \frac{1}{s+4} \end{bmatrix}$$

(a) Compute $K(s)$.

(b) Compute the poles and zeros (with associated input zero directions) of $P(s)$ and $K(s)$.

(c) Are there pole/zero cancellations between $P(s)$ and $K(s)$?

(d) Is the system internally stable? Verify your answer.

Exercise 17.7 An engineer wanted to estimate the peak-to-peak gain of a closed loop system h (the input-output map). The controller was designed so that the system tracks a step input in the steady state. The designer simulated the step response of the system and computed the amount of overshoot (e_1) and undershoot (e_2) of the response. He/She immediately concluded that

$$\|h\|_1 \geq 1 + 2e_1 + 2e_2.$$

Is this a correct conclusion? Verify.

Chapter 18

Performance of Feedback Systems

18.1 Introduction

It is now time to turn to issues of **performance**. As noted in earlier chapters, performance specifications typically involve the *closed-loop* relations between the exogenous inputs w and the regulated outputs z. These relationships are typically captured through the use of the signal and system norms. The *analysis* of a given controlled system usually involves evaluating the appropriate norms. The *synthesis* of a controller is a harder problem, as it involves picking a feedback compensator K for which the closed-loop performance specifications are attained.

We begin our discussions with the single-input, single-output (SISO) case, and then move on to study multi-input, multi-output (MIMO) extensions. Much of what we present for the SISO case actually echoes what is done in "classical feedback control", although our perspective is somewhat more modern (or neo-classical or post-modern or ...!).

18.2 SISO Loop Shaping

The Classical Viewpoint

The standard "servo" or *tracking* configuration of classical feedback control is shown in Figure 18.1. In this arrangement, the controller K is fed by an error signal e, which is the difference between a reference r and the measured output y of the plant P. The measurement is perhaps corrupted by noise n. The output of the controller is the input u to the plant. In addition, external disturbances may drive the plant, and are represented here via the signal d added in at the output of the plant. In a typical classical control design, the compensator K would be picked as the lowest-order system that ensures the following:

1. the closed-loop system is *stable*;

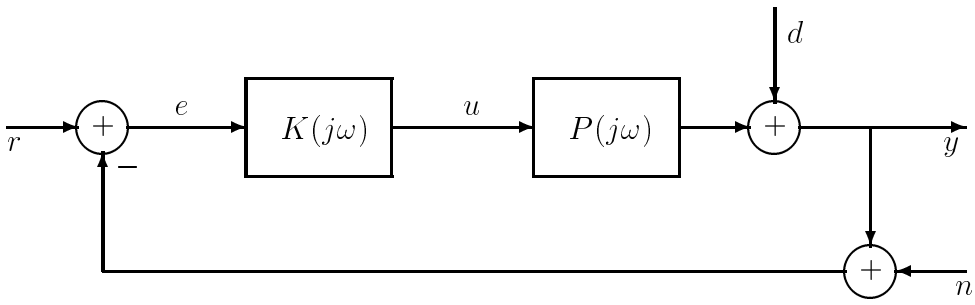


Figure 18.1: Standard feedback configuration with noise, disturbance, and reference inputs.

2. the **loop gain** $P(j\omega)K(j\omega)$ has *large* magnitude at frequencies (low frequencies, typically) where the power of the plant disturbance d or reference input r is concentrated;

3. the loop gain has *small* magnitude at frequencies (high frequencies, typically) where the power of the measurement noise n is concentrated.

The need for the first requirement is clear. The origins of the second and third requirements will be explained below. In order to simultaneously attain all three objectives, it is most convenient to have a criterion for closed-loop stability that is stated in terms of the (open-loop) loop gain, and this is provided by the *Nyquist stability criterion*.

The reasons for the second and third requirements above lie in the *sensitivities* of the closed-loop system to plant disturbances, reference signals, and measurement noise. Let S denote the transfer function that maps a disturbance d to the output y in the closed-loop system. This S is termed the (output) **sensitivity function**, and for the arrangement in Figure 18.1 it is given by

$$S = (1 + PK)^{-1} \quad . \tag{18.1}$$

Speaking informally for the moment, if $|P(j\omega)K(j\omega)|$ is large at frequencies where (in some sense) the power of d is concentrated, then $|S(j\omega)|$ will be small there, so the effect of the disturbance on the output will be attenuated. Since plant disturbances are typically concentrated around the low end of the frequency spectrum, one would want $|P(j\omega)K(j\omega)|$ to be large at low frequencies. Thus, *disturbance rejection* is a key motivation behind classical control's low-frequency specification on the loop gain.

Note that (in the SISO case) S is also the transfer function from r to e. If we want y to track r with good accuracy, then we want a small response of the error signal e to the driving signal r. This again leads us to ask for $|S(j\omega)|$ to be small — or equivalently for $|P(j\omega)K(j\omega)|$ to be large — at frequencies where the power of the reference signal r is concentrated. Fortunately, in many (if not most) control applications, the reference signal is slowly varying, so this requirement again reduces to asking for $|P(j\omega)K(j\omega)|$ to be large at low frequencies. Thus, *tracking accuracy* is another motivatoin behind classical control's low-frequency specification on the loop gain.

In contrast, the motivation behind classical control's high frequency specification is *noise rejection*. Let T denote the transfer function that maps the noise input n to the output y. Given the arrangement in Figure 18.1,

$$T = PK(1+PK)^{-1} \quad . \tag{18.2}$$

This T is termed the **complementary sensitivity function**, because

$$T + S = 1 \, . \tag{18.3}$$

Note that T is also the transfer function from r to y. If $|P(j\omega)K(j\omega)|$ is small at frequencies where the power in n is concentrated, then $|T(j\omega)|$ will be small there, so the effect of the noise on the output will be attenuated. Measurement noise tends to occur at higher frequencies, so to minimize its effects on the output, we typically specify that $|P(j\omega)K(j\omega)|$ be small at high frequencies. This constraint fortunately does not conflict with the low-frequency constraints imposed above by typical d and r. Also, the constraint is well matched to the inevitable fact that the gain of physical systems will eventually fall off with frequency.

The picture of the control design task that emerges from the above discussion is the following: Given the plant P, one typically needs to pick the compensator K so as to obtain a loop gain magnitude $|P(j\omega)K(j\omega)|$ that is large at low frequencies, "rolls off" to low values at high frequencies, and varies in such a way that the Nyquist stability criterion is satisfied. [For the special case of open-loop stable plants and compensators, the stability condition can be stated in alternative forms that are easy to check using Bode plots rather than Nyquist plots, and this can be more convenient. The standard rule of thumb focuses on the roll-off around the *crossover frequency* ω_c, defined as the frequency where the loop gain magnitude is unity; this frequency is a crude measure of closed-loop bandwidth. The specification is that the roll-off of the loop gain magnitude around ω_c should be no steeper than -20dB/decade. Furthermore, ω_c should be picked below frequencies where the loop gain is significantly affected by any right-half-plane zeros of the loop transfer function PK; this provides an initial indication that right-half-plane zeros can limit the attainable closed-loop performance.]

A Modern Viewpoint

The challenge now is to translate the above classical control design approach into something more precise and systematic, and more likely to have a natural MIMO extension. The following example points the way, and makes free use of the signal and system norms that we defined in Lectures 11 and 12.

Example 18.1 (SISO Disturbance Rejection and Weighted Sensitivity)

We have already seen that the expression relating y to d in the SISO feedback configuration depicted in Figure 18.1 is

$$y = (1+PK)^{-1}d \quad . \tag{18.4}$$

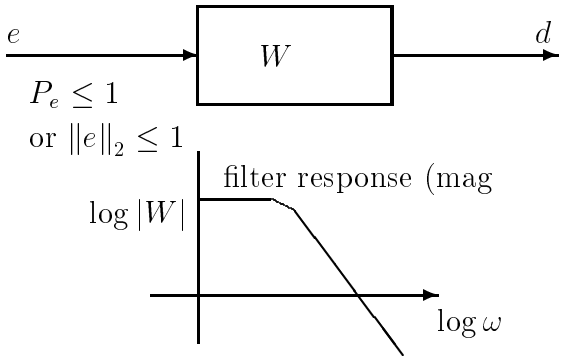


Figure 18.2: Representing the plant disturbance d as the output of a shaping filter W whose input e is an arbitrary bounded energy or bounded power signal, or possibly white noise.

Typically, d has frequency content concentrated in the low-frequency range. In order to get the requisite frequency characteristic, one might model d as the output of a *shaping filter* with transfer function W, as shown in Figure 18.2, with the input e of the filter being an arbitrary bounded energy or bounded power disturbance (or, in the stochastic setting, white noise). Thus e has no spectral "coloring", and all the coloring of d is embodied in the frequency response of W.

For the rest of this example, let us focus on the bounded energy or bounded power models for e. Suppose our goal now is to choose K to minimize the effect of the disturbance d on the output y. From Lectures 11 and 12, and given our model for d, we know that this is equivalent to minimizing the $\mathcal{H}_\infty$-gain of the transfer function from e to y, because in the case of a bounded power e this gain is the attainable or "tight" bound on the ratio of rms values at the output and input,

$$\frac{\rho_y}{\rho_e} \leq \|(1 + P(j\omega)K(j\omega))^{-1}W(j\omega)\|_\infty \quad ,$$

while in the case of a bounded energy e we again have the tight bound

$$\frac{\|y\|_2}{\|e\|_2} \leq \|(1 + P(j\omega)K(j\omega))^{-1}W(j\omega)\|_\infty \quad .$$

In terms of the sensitivity function,

$$S(j\omega) = (1 + P(j\omega)K(j\omega))^{-1} \quad ,$$

the task is to pick K to minimize the $\mathcal{H}_\infty$ norm $\|S(j\omega)W(j\omega)\|_\infty$.

If

$$\|S(j\omega)W(j\omega)\|_\infty \leq \gamma \, , \tag{18.5}$$

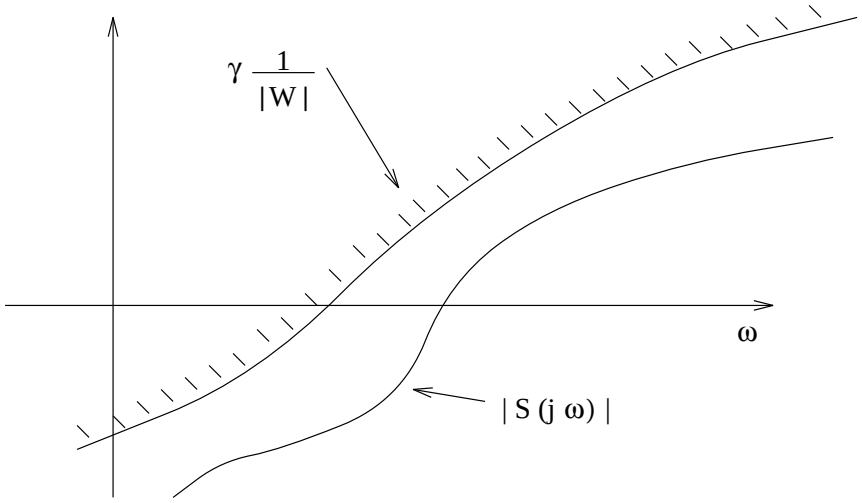


Figure 18.3: Graphical interpretation of the sensitivity function being bounded by a scaled reciprocal of the weighting filter frequency response.

then

$$|S(j\omega)|\,|W(j\omega)| \leq \gamma\ , \quad \forall\omega\ . \tag{18.6}$$

This implies that

$$|S(j\omega)| \leq \gamma \frac{1}{|W(j\omega)|} \quad , \tag{18.7}$$

which tells us that the sensitivity function is bounded by a scaled reciprocal of the weighting filter. A graphical representation of this bound is shown in Figure 18.3. From Figure 18.3 we can see that the value γ and the filter $W(j\omega)$ give us a clear picture of the constraint on the sensitivity function. This allows one to more systematically design a controller, since we directly get the closed loop characteristics. Note also that with the Q-parametrization of K, the sensitivity function S is affine in Q, and this form is much easier to work with than the fractional form that S takes as a function of K.

The major benefit of the formulation in the above example is that a MIMO version of it is quite immediate, as we see in the next section.

18.3 MIMO Loop Shaping

Let us now revisit the above example in the MIMO setting. The example will require the following facts about singular values, so we ask you to confirm these facts for yourself before proceeding:

1. $\sigma_{max}(AB) \leq \sigma_{max}(A)\sigma_{max}(B)$, and

2. If $\sigma_{max}(CD) < 1$ then $\sigma_{max}(C) < \dfrac{1}{\sigma_{min}(D)}$ assuming D is invertible.

The first statement follows from the fact that σ_{max} is the induced 2-norm, and therefore submultiplicative. To prove the second, apply the first with $A = CD$ and $B = D^{-1}$.)

Example 18.2 (MIMO Disturbance Rejection and Weighted Sensitivity)

The set-up and formulation for the MIMO case are the same as in the SISO example, with the obvious replacements of SISO subsystems by MIMO subsystems. One again arrives at the equation (18.5). However, the inference from this equation in the MIMO case is no longer (18.6) and (18.7), but rather

$$\sigma_{max}\left[(I + P(j\omega)K(j\omega))^{-1}\right] \leq \gamma \frac{1}{\sigma_{min}\left[W(j\omega)\right]} \quad .$$

This leads us to the singular value plot shown in Figure 18.4, which is the natural extension of the plot we had in the SISO example.

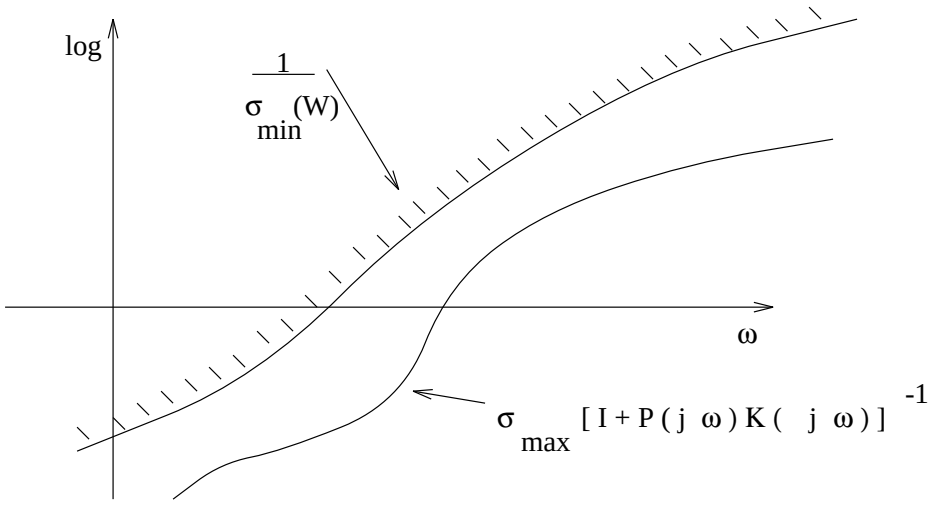


Figure 18.4: Singular value plot for a MIMO system.

With the insight provided by the above example, we can formulate a variety of MIMO performance problems in terms of appropriate weighting operators. Alternatively, having seen what sorts of modifications of the SISO statements are needed for the MIMO case, we can actually describe various MIMO control tasks in a language that is closer to that of classical SISO control, and this is what we do in the rest of this lecture. We shall return to the explicit use of weighting functions in later lectures.

Typical Closed-Loop Perfomance Constraints

Typically in control systems the disturbances d have frequency content that is concentrated in the low-frequency range. Therefore, in order to attenuate the effects of disturbances on the output, we require that $\sigma_{max}(S(j\omega))$ be small in the range of frequencies where the disturbances are active, say $0 \leq \omega \leq \omega_{sy}$. On the other hand, typically the noise input n has frequency content that is concentrated in the high-frequency range. Therefore, in order to attenuate the effect of n on the output we require that $\sigma_{max}(T(j\omega))$ be small over a frequency range of the form $\omega \geq \omega_{ty}$. The controller K should also enable the closed-loop system to track reference inputs r that are typically concentrated in the low frequency range, for example in the interval $0 \leq \omega \leq \omega_r$. This objective requires that $T(j\omega) \approx I$ for all ω in the interval $0 \leq \omega \leq \omega_r$. This requirement can be restated as

$$\begin{aligned} \sigma_{max}(T(j\omega)) &\approx 1 \\ \sigma_{min}(T(j\omega)) &\approx 1, \end{aligned}$$

in the frequency range $0 \leq \omega \leq \omega_r$.

The control signals must also generally be kept as small as possible in the presence of both disturbances d and measurement noise n. It is easy to see that

$$u = (I + KP)^{-1}Kr - (I + KP)^{-1}K(d + n) \; .$$

Therefore, in order to keep the control signal small, we must make sure that

$$\sigma_{max}\left((I + K(j\omega)P(j\omega))^{-1}K(j\omega)\right)$$

remains small in the frequency range where disturbances and/or measurement errors are effective. We can summarize these design requirements in the following table:

Design Requirement	Closed-Loop Condition	Frequency Range
Sensitivity to Disturbances	$\sigma_{max}\left((I + P(j\omega)K(j\omega))^{-1}\right) \approx 0$	Low frequency $0 \leq \omega \leq \omega_{sy}$
Noise Propagation Attenuation	$\sigma_{max}\left((I + P(j\omega)K(j\omega))^{-1}P(j\omega)K(j\omega)\right) \approx 0$	High Frequency $\omega \geq \omega_{ty}$
Tracking of Reference Signals	$\sigma_{max}\left((I + P(j\omega)K(j\omega))^{-1}P(j\omega)K(j\omega)\right) \approx 1$ $\sigma_{min}\left((I + P(j\omega)K(j\omega))^{-1}P(j\omega)K(j\omega)\right) \approx 1$	Low frequency $0 \leq \omega \leq \omega_r$
Low Control Energy	$\sigma_{max}\left((I + K(j\omega)P(j\omega))^{-1}K(j\omega)\right) \approx 0$	Frequencies where d and n are dominant

Translation to Open-Loop Constraints

Now let us relate the closed-loop requirements that are summarized in the preceding table to open-loop conditions, i.e., conditions on the singular values of the loop gain operator

PK. The first design requirement is that $\sigma_{max}\left((I+PK)^{-1}\right)$ be small in the frequency range $0 \leq \omega \leq \omega_{sy}$. The relation

$$\sigma_{max}\left((I+P(j\omega)K(j\omega))^{-1}\right) = \frac{1}{\sigma_{min}(I+P(j\omega)K(j\omega))}$$

implies that if $\sigma_{min}(P(j\omega)K(j\omega)) >> 1$ then

$$\sigma_{max}\left((I+P(j\omega)K(j\omega))^{-1}\right) \approx \frac{1}{\sigma_{min}(P(j\omega)K(j\omega))}\,. \tag{18.8}$$

Therefore, if $\sigma_{min}(P(j\omega)K(j\omega)) >> 1$ for all ω in the interval $[0, \omega_{sy}]$, then $\sigma_{max}\left((I+P(j\omega)K(j\omega))^{-1}\right)$ will be small in that interval.

For noise attenuation, consider

$$\begin{aligned}
\sigma_{max}\left(T(j\omega)\right) &= \sigma_{max}\left(I-(I+P(j\omega)K(j\omega))^{-1}\right) \\
&= \sigma_{max}\left(\left(I+(P(j\omega)K(j\omega))^{-1}\right)^{-1}\right) \\
&= \frac{1}{\sigma_{min}\left(I+(P(j\omega)K(j\omega))^{-1}\right)}.
\end{aligned}$$

Therefore, for the frequency range $\omega \geq \omega_{ty}$ we require that $\sigma_{min}\left(I+(P(j\omega)K(j\omega))^{-1}\right)$ be as large as possible. This can be guaranteed if we make $\sigma_{min}\left((P(j\omega)K(j\omega))^{-1}\right)$ as large as possible or equivalently by making $\sigma_{max}\left(P(j\omega)K(j\omega)\right)$ as small as possible.

The tracking objective can be achieved if we ensure that

$$\begin{aligned}
\sigma_{max}\left((I+P(j\omega)K(j\omega))^{-1}P(j\omega)K(j\omega)\right) &\approx 1 \\
\sigma_{min}\left((I+P(j\omega)K(j\omega))^{-1}P(j\omega)K(j\omega)\right) &\approx 1
\end{aligned}$$

over the frequency interval $[0, \omega_r]$. Since

$$I-(I+P(j\omega)K(j\omega))^{-1} = (I+P(j\omega)K(j\omega))^{-1}P(j\omega)K(j\omega)$$

the tracking objective can be achieved if we require $(I+P(j\omega)K(j\omega))^{-1}$ to be close to zero on the frequency range $[0, \omega_r]$, that is $\sigma_{max}\left((I+P(j\omega)K(j\omega))^{-1}\right)$ to be small in that interval. Equivalently, we may require $\sigma_{min}\left(I+P(j\omega)K(j\omega)\right)$ to be as large as possible on the interval $[0, \omega_r]$. This can be ensured if we require that $\sigma_{min}\left(P(j\omega)K(j\omega)\right)$ be as large as possible over the frequency range $[0, \omega_r]$.

The constraint of small control energy leads to the condition that $\sigma_{max}\left((I+K(j\omega))P(j\omega)\right)^{-1}K(j\omega))$ be made as small as possible. However, we have

$$\begin{aligned}
\sigma_{max}\left((I+K(j\omega)P(j\omega))^{-1}K(j\omega)\right) &\leq \sigma_{max}\left((I+K(j\omega)P(j\omega))^{-1}\right)\sigma_{max}(K(j\omega)) \\
&= \frac{\sigma_{max}(K(j\omega))}{\sigma_{min}\left(I+K(j\omega)P(j\omega)\right)}.
\end{aligned} \tag{18.9}$$

Note that

$$\begin{aligned}\sigma_{min}\left(I + K(j\omega)P(j\omega)\right) &\leq \sigma_{max}\left(I + K(j\omega)P(j\omega)\right) \\ &\leq 1 + \sigma_{max}(P(j\omega))\sigma_{max}(K(j\omega))\end{aligned}$$

so

$$\begin{aligned}\frac{\sigma_{max}(K(j\omega))}{\sigma_{min}\left(I + K(j\omega)P(j\omega)\right)} &\geq \frac{\sigma_{max}(K(j\omega))}{1 + \sigma_{max}(P(j\omega))\sigma_{max}(K(j\omega))} \\ &= \frac{1}{\frac{1}{\sigma_{max}(K(j\omega))} + \sigma_{max}(P(j\omega))}.\end{aligned}$$

Therefore, we can minimize the right hand side of equation 18.9 only if we make

$$\frac{1}{\sigma_{max}(K(j\omega))} + \sigma_{max}(P(j\omega))$$

large in the ranges of frequencies where d and/or n are dominant. For example, if $\sigma_{max}(P(j\omega))$ is small at a certain set of frequencies of interest then necessarily $\sigma_{max}(K(j\omega))$ must also be small on that set. Clearly this condition is not necessary or sufficient to make

$$\sigma_{max}\left((I + K(j\omega)P(j\omega))^{-1}K(j\omega)\right)$$

small. It only applies to the upper bound of $\sigma_{max}\left((I + K(j\omega)P(j\omega))^{-1}K(j\omega)\right)$, which is given by

$$\frac{\sigma_{max}(K(j\omega))}{\sigma_{min}\left(I + K(j\omega)P(j\omega)\right)}$$

and it is only necessary for the upper bound to be small.

The following table summarizes our discussion above on open-loop requirements

Design Requirement	Open-Loop Condition	Frequency Range
Sensitivity to Disturbances	$\sigma_{min}\left(P(j\omega)K(j\omega)\right)$ large	Low frequency $0 \leq \omega \leq \omega_{sy}$
Noise Propagation Attenuation	$\sigma_{max}\left(P(j\omega)K(j\omega)\right)$ small	High Frequency $\omega \geq \omega_{ty}$
Tracking of Reference Signals	$\sigma_{min}\left(P(j\omega)K(j\omega)\right)$ large	Low frequency $0 \leq \omega \leq \omega_r$
Low Control Energy	$\sigma_{max}\left(K(j\omega)\right)$ small	Frequencies where $\sigma_{max}\left(P(j\omega)\right)$ is not large enough

Figure 18.6 illustrates the open-loop conditions that we have formulated. Note that in this plot the minimum passband open-loop gain is bounded by $\sigma_{min}\left[P(j\omega)K(j\omega)\right]$, and the maximum stopband open loop gain bounded by $\sigma_{max}\left[P(j\omega)K(j\omega)\right]$.

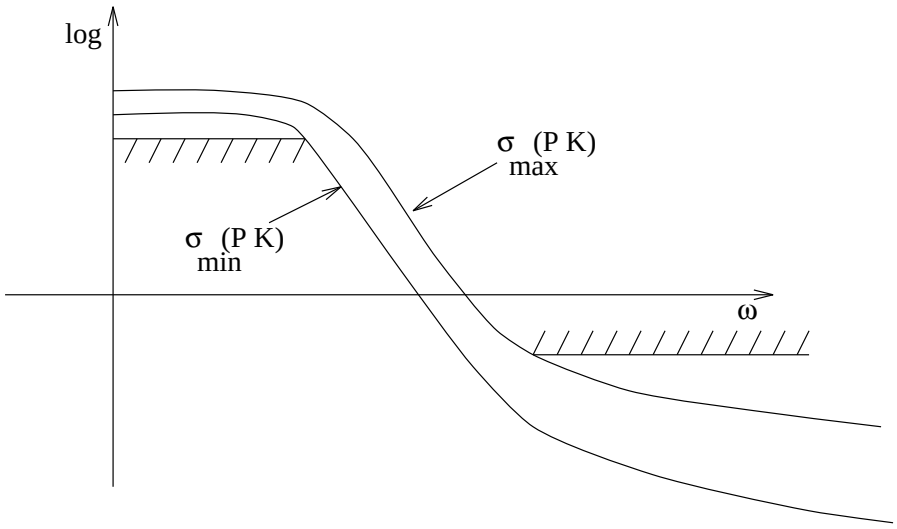


Figure 18.5: Singular value bounds for the open loop gain, $P(j\omega)K(j\omega)$.

18.4 Algebraic Constraints

In general we would like to design feedback controllers to attenuate both noise and disturbances at the output. We have examined SISO and MIMO conditions that guarantee rejection of low frequency disturbances as well as similar conditions for the rejection of high frequency noise. However, one might wonder if we can

1. minimize the influence of either noise or disturbances over all frequencies, and/or
2. minimize the influence of both noise and disturbances at the same frequency.

Let us begin this discussion by recalling the following:

- $S = (I + PK)^{-1}$ is the transfer function mapping disturbances to the output;
- $T = PK(I + PK)^{-1}$ is the transfer function mapping noise to the output.

As we mentioned earlier, in a control design it is usually desirable to make both S and T small. However, because of algebraic constraints, both goals are not simultaneously achievable at the same frequency. These constraints are as follows.

General Limitations

$S + T = I$ for all complex (Laplace domain) frequencies s. This is easily verified, since

$$\begin{aligned} S + T &= (I + PK)^{-1} + PK(I + PK)^{-1} \\ &= (I + PK)\,(I + PK)^{-1} \\ &= I \quad . \end{aligned}$$

This result implies that if $\sigma_{max}\left[S(j\omega)\right]$ is small in some frequency range, $\sigma_{max}\left[T(j\omega)\right] \sim 1$. The converse is also true.

Fortunately, we rarely need to make both of these functions small in the same frequency region.

Limitations Due to RHP Zeros and Poles

Before we discuss these limitations, we quote the following fact from complex analysis:

> Let $H(s)$ be a stable, causal, linear time-invariant continuous-time system. The *maximum modulus principle* implies that
>
> $$\sigma_{max}\left[H(s)\right] \leq \sup_{\omega} \sigma_{max}\left[H(j\omega)\right] = \|H\|_{\infty} \qquad \forall\, s \in \text{RHP} \quad .$$

In other words, a stable function, which is analytic in the RHP, achieves its maximum value over the RHP when evaluated on the imaginary axis.

Using this result, we can arrive at relationships between poles and zeros of the plant P located in the RHP and limitations on performance (*e.g.*, disturbance and noise rejection).

SISO Systems: Disturbance Rejection

Consider the stable sensitivity function $S = (1+PK)^{-1}$ for any stabilizing controller, K; then,

$$\begin{aligned} S(z_i) &= (1+P(z_i)K(z_i))^{-1} = 1 \quad \text{for all RHP zeros } z_i \text{ of } P \\ S(p_i) &= (1+P(p_i)K(p_i))^{-1} = 0 \quad \text{for all RHP poles } p_i \text{ of } P \quad . \end{aligned}$$

Since the $\mathcal{H}_\infty$ norm bounds the gain of a system over all frequencies,

$$1 = |S(z_i)| \leq \|S\|_{\infty} \quad .$$

This means that we cannot uniformly attenuate disturbances over the entire frequency range if there are zeros in the RHP.

SISO Systems: Noise Rejection

Since the transfer function relating a noise input to the output is $T = PK(1+PK)^{-1}$, an argument for T similar to S can be made, but with the roles of poles and zeros interchanged. In this case, RHP poles of the plant restrict us from uniformly attenuating noise over the entire frequency range.

MIMO Systems: Disturbance Rejection

Suppose P has a transmission zero at $\tilde{z} \in$ RHP with left input zero direction η^*. Then $\eta^* P(\tilde{z})K(\tilde{z}) = 0$, and thus

$$\eta^*(I + P(\tilde{z})K(\tilde{z}))^{-1} = \eta^* \quad .$$

Stated equivalently,

$$\eta^* S(\tilde{z}) = \eta^* \quad . \tag{18.10}$$

Also, taking the conjugate transpose of both sides,

$$S^*(\tilde{z})\eta = \eta \quad . \tag{18.11}$$

We then multiply the expressions in (18.10) and (18.11), obtaining

$$\eta^* S(\tilde{z})S^*(\tilde{z})\eta = \eta^*\eta \quad ,$$

which can be alternately written as

$$\frac{\eta^* S(\tilde{z})S^*(\tilde{z})\eta}{\eta^*\eta} = 1 \quad . \tag{18.12}$$

Applying the maximum modulus principle (*i.e.*, $\max_{s\in\text{RHP}} \sigma_{max}[S(s)]$ occurs on the imaginary axis) and observing that the left hand side of (18.12) is less than or equal to $\sigma^2_{max}[S(\tilde{z})]$, we conclude that

$$\|S\|^2_\infty \geq \frac{\eta^* S(\tilde{z})S^*(\tilde{z})\eta}{\eta^*\eta} = 1 \quad .$$

Thus, the conclusion regarding disturbance rejection for MIMO systems is the same as the conclusion we reached for SISO systems. Namely, RHP zeros make disturbance attenuation over all frequencies impossible.

18.5 Analytic Constraints: The "Waterbed Effect"

One performance limitation of LTI SISO Feedback systems (these systems have rational sensitivity transfer functions), is known as the *waterbed effect*. Loosely speaking, when one designs a controller to "push" the sensitivity function in a particular direction, another part of the sensitivity function necessarily "pulls" back in the opposite direction. This effect is due to a property of analytic functions $f(s)$ as stated by Cauchy's theroem. In words, this theorem says that the line integral of an analytic function around any simple closed contour C in a region $\mathbf{R}$ is zero, *i.e.,*

$$\int_C f(s)ds = 0.$$

for every contour C in $\mathbf{R}$.

A proof of this theorem will not be shown here but can be found in standard complex analysis textbooks. One consequence of this theorem is the following integral constraint (known as *Bode's Integral*) on the rational sensitivity transfer function $S(jw)$:

$$\int_0^\infty ln|S(jw|dw = \sum_i \pi Re(p_i)$$

where $\sum_i \pi Re(p_i)$ is the sum over the unstable open-loop poles (poles of $P(jw)K(jw)$). This result holds for all closed-loop systems as long as the product PK has relative degree two. The result implies that making $S(jw)$ small at almost all frequencies (a common performance objective) is impossible since the integrated value of $ln|S(jw)|$ over all frequencies must be constant. This constant is zero for open-loop stable systems (PK stable) and positive otherwise. Therefore, lowering the sensitivity function in one range of frequencies, increases the same function in another range-hence the name "waterbed effect." Figure 18.5 below illustrates this phenomenon.

Figure 18.6: Water-bed Effect

Constraints on Singular Value Plots

From what we have seen already, it is clear that singular value plots over all frequencies are the MIMO system analogs of Bode plots. The following fact establishes some simple bounds involving singular values of S and T:

Fact 18.5.1 *If $S = (I + PK)^{-1}$ and $T = (I + PK)^{-1}PK$ then the following hold*

$$|1 - \sigma_{max}(S)| \leq \sigma_{max}(T) \leq 1 + \sigma_{max}(S)$$

and

$$|1 - \sigma_{max}(T)| \leq \sigma_{max}(S) \leq 1 + \sigma_{max}(T).$$

Proof: Since $S + T = I$ then clearly

$$\sigma_{max}(T) = \sigma_{max}(I - S) \leq \sigma_{max}(I) + \sigma_{max}(S)$$

and therefore $\sigma_{max}(T) \leq 1 + \sigma_{max}(S)$. For any element $x \in \mathbb{C}^n$ with $\|x\|_2 = 1$ we have

$$\begin{aligned} x - Sx &= Tx \\ |\|x\|_2 - \|Sx\|_2| &\leq \|x - Sx\|_2 = \|Tx\|_2 \\ |1 - \|Sx\|_2| &\leq \sigma_{max}(T) \\ |1 - \sigma_{max}(S)| &\leq \sigma_{max}(T). \end{aligned}$$

Combining this relation with $\sigma_{max}(T) \leq 1 + \sigma_{max}(S)$, we obtain

$$|1 - \sigma_{max}(S)| \leq \sigma_{max}(T) \leq 1 + \sigma_{max}(S).$$

The other relation follows in exactly the same manner.

Exercises

Exercise 18.1 Suppose a discrete-time plant is given by

$$P = \begin{pmatrix} \frac{1-2z^{-1}}{1-.5z^{-1}} \\ \frac{1-z^{-1}}{1-.5z^{-1}} \end{pmatrix}$$

Does there exist a controller that uniformly attenuates the input sensitivity function $(I+KP)^{-1}$, i.e., $\|(I+KP)^{-1}\|_\infty < 1$. Explain.

Exercise 18.2 Let a plant be given by

$$G(s) = \begin{pmatrix} \frac{s-1}{s+1} & -5 \\ \frac{s+2}{(s+1)^2} & \frac{s-1}{s+1} \end{pmatrix}.$$

We are interested in verifying whether or not there exists a controller K such that the output sensitivity $S = (I+PK)^{-1}$ satisfies $\|S\|_\infty < 1$ (i.e., the maximum singular value is strictly less than 1 for all frequencies). If this is possible, we would like to find such a controller.

1. One engineer argued as follows: Since the transfer functions from u_1 to y_1 and u_2 to y_2 have nonminimum-phase zeros, then the sensitivity cannot be uniformly attenuated. Do you accept this argument. If so, explain her/his rationale, and if not explain why not.
2. Another engineer suggested that the controller can invert the plant and add a scaling factor, so that the sensitivity is uniformly less than 1. Again discuss this option and argue for it or against it.

Exercise 18.3 Consider the following MIMO plant $P(s)$ whose state-space description is

$$\dot{x}(t) = \begin{bmatrix} -1.5 & 1 & 0 & 1 \\ 2 & -3 & 2 & 0 \\ 0 & .5 & -2 & 1 \\ 1 & -1.5 & 0 & -5 \end{bmatrix} x(t) + \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 1 & 1 \\ 0 & 1.8 \end{bmatrix} u(t)$$

$$y(t) = \begin{bmatrix} 0 & 2.4 & -3.1 & 1 \\ 1 & 6 & -.5 & -2.8 \end{bmatrix} x(t)$$

(a) Use Matlab to compute the poles and the zeros of the plant, as well as the associated input zero directions. (The transmission zeros should turn out to be around $-.544 \pm j2.43$.)

(b) Plot the singular values of $P(j\omega)$ for $\omega \in [-10^{-2}, 10^2]$ rad/sec. Relate the shapes of the singular values to the pole and zero frequencies of $P(s)$.

(c) Compute $\|P\|_\infty$ using the Hamiltonian matrix and "gamma iteration", and compare the result to part b).

(d) Consider the standard MIMO servo feedback loop with a compensator of transfer matrix $K(s)$ preceding $P(s)$ in the forward loop. The input to the compensator is the error signal $e(t) = r(t) - y(t)$, where $r(t)$ is an external reference signal. Design $K(s)$ to have the following properties:

(i) $K(s)$ should be strictly proper, second-order (i.e. a minimal realization of it is second-order), with no transmission zeros, and with poles that exactly cancel the transmission zeros of $P(s)$ — so $P(s)K(s)$ does not have these zeros.

(ii) $\lim_{s\to 0} P(s)K(s) = 40I$

Also obtain a state-space description of $K(s)$.

(e) Plot the singular values of the open-loop frequency response $P(j\omega)K(j\omega)$, the sensitivity function $S(j\omega)$, and the closed-loop frequency response (or complementary sensitivity function) $T(j\omega) = I - S(j\omega)$.

(f) Predict the steady-state value of the output vector $y(t)$ when the reference input to the closed-loop system (which is assumed initially at rest) is the step

$$r(t) = \begin{bmatrix} 7 \\ -3 \end{bmatrix}, \qquad t \geq 0 \tag{18.13}$$

and verify by computing (with Matlab!) the transient response for the above step input. By carefully examining the transients of the control input and output signals, discuss the implications of having oscillatory poles in the compensator that cancel the plant transmissions zeros.

(g) Predict the steady-state maximum and minimum value of the tracking error $e(t)$ when the command input vector comprises unit sinusoids at a frequency of $\omega = 1$ rad/sec. Repeat for $\omega = 2.5$ rad/sec.

Chapter 19

Robust Stability in SISO Systems

19.1 Introduction

There are many reasons to use feedback control. As we have seen earlier, with the help of an appropriately designed feedback controller we can reduce the effect of noise and disturbances, and we can improve the tracking of command signals. Another very important use for feedback control is the reduction of the effects of plant uncertainty. The mathematical models that we use to describe the plant dynamics are almost never perfect. A feedback controller can be designed so as to maintain stability of the closed-loop and an acceptable level of performance in the presence of uncertainties in the plant description, i.e., so as to achieve *robust stability* and *robust performance* respectively.

For the study of robust stability and robust performance, we assume that the dynamics of the actual plant are represented by a transfer function that belongs to some uncertainty set Ω. We begin by giving mathematical descriptions of two possible uncertainty sets. Many other descriptions exist, and may be treated by methods similar to those we present for these particular types of uncertainty sets.

19.2 Additive Representation of Uncertainty

It is commonly the case that the nominal plant model is quite accurate for low frequencies but deteriorates in the high-frequency range, because of parasitics, nonlinearities and/or time-varying effects that become significant at higher frequencies. These high-frequency effects may have been left unmodeled because the effort required for system identification was not justified by the level of performance that was being sought, or they may be well-understood effects that were omitted from the nominal model because they were awkward and unwieldy to carry along during control design. This problem, namely the deterioration of nominal models at higher frequencies, is mitigated to some extent by the fact that almost all physical systems have

strictly proper transfer functions, so that the system gain begins to roll off at high frequency.

In the above situation, with a nominal plant model given by the proper transfer function $P_0(s)$, the actual plant represented by $P(s)$, and the difference $P(s) - P_0(s)$ assumed to be stable, we may be able to characterize the model uncertainty via a bound of the form

$$|P(j\omega) - P_0(j\omega)| \leq \ell_a(\omega) \tag{19.1}$$

where

$$\ell_a(\omega) = \begin{cases} \text{"Small"} & ; \quad |\omega| < \omega_c \\ \text{"Bounded"} & ; \quad |\omega| > \omega_c \end{cases} . \tag{19.2}$$

This says that the response of the actual plant lies in a "band" of uncertainty around that of the nominal plant. Notice that no phase information about the modeling error is incorporated into this description. For this reason, it may lead to conservative results.

The preceding description suggests the following simple *additive* characterization of the uncertainty set:

$$\Omega_a = \{P(s) \mid P(s) = P_0(s) + W(s)\Delta(s)\} \tag{19.3}$$

where Δ is an arbitrary *stable* transfer function satisfying the norm condition

$$\|\Delta\|_\infty = \sup_\omega |\Delta(j\omega)| \leq 1, \tag{19.4}$$

and the *stable* proper rational weighting term $W(s)$ is used to represent any information we have on how the accuracy of the nominal plant model varies as a function of frequency. Figure 19.1 shows the additive representation of uncertainty in the context of a standard servo loop, with K denoting the compensator.

When the modeling uncertainty increases with frequency, it makes sense to use a weighting function $W(j\omega)$ that looks like a high-pass filter: small magnitude at low frequencies, increasing but bounded at higher frequencies.

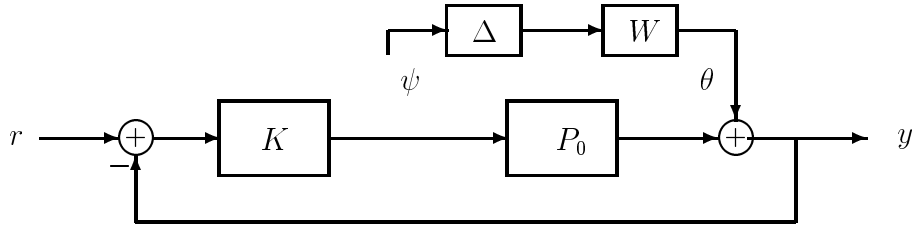


Figure 19.1: Representation of the actual plant in a servo loop via an additive perturbation of the nominal plant.

Caution: The above formulation of an additive model perturbation should *not* be interpreted as saying that the actual or perturbed plant is the *parallel combination* of the nominal system $P_0(s)$ and a system with transfer function $W(s)\Delta(s)$. Rather, the actual plant should be considered as being a *minimal realization* of the transfer function $P(s)$, which happens to be written in the additive form $P_0(s) + W(s)\Delta(s)$.

Some features of the above uncertainty set are worth noting:

- The *unstable* poles of all plants in the set are precisely those of the nominal model. Thus, our modeling and identification efforts are assumed to be careful enough to accurately capture the unstable poles of the system.
- The set includes models of arbitrarily large order. Thus, if the uncertainties of major concern to us were *parametric uncertainties*, i.e. uncertainties in the values of the parameters of a particular (e.g. state-space) model, then the above uncertainty set would greatly overestimate the set of plants of interest to us.

The control design methods that we shall develop will produce controllers that are guaranteed to work for *every member* of the plant uncertainty set. Stated slightly differently, our methods will treat the system as though *every* model in the uncertainty set is a possible representation of the plant. To the extent that not all members of the set are possible plant models, our methods will be conservative.

Suppose we have a set of possible plants Π such that the true plant is a member of that set. We can try to embed this set in an additive perturbation structure. First let $P_0 \in \Pi$ be a certain nominal plant in Π. For any other plant $P \in \Pi$ we write,

$$P(j\omega) = P_0(j\omega) + W(j\omega)\Delta(j\omega).$$

The weight $|W(j\omega)|$ satisfies

$$\begin{aligned} |W(j\omega)| &\geq |W(j\omega)\Delta(j\omega)| = |P(j\omega) - P_0(j\omega)| \\ |W(j\omega)| &\geq \max_{P\in\Pi} |P(j\omega) - P_0(j\omega)| = \ell_a(j\omega). \end{aligned}$$

With the knowledge of the lower bound $\ell_a(j\omega)$, we find a stable system $W(s)$ such that $|W(j\omega)| \geq \ell_a(j\omega)$

19.3 Multiplicative Representation of Uncertainty

Another simple means of representing uncertainty that has some nice analytical properties is the *multiplicative perturbation*, which can be written in the form

$$\Omega_m = \{P \mid P = P_0(1 + W\Delta),\ \|\Delta\|_\infty \leq 1\}. \tag{19.5}$$

W and Δ are stable. As with the additive representation, models of arbitrarily large order

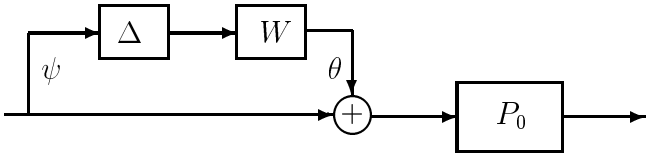


Figure 19.2: Representation of uncertainty as multiplicative perturbation at the plant input.

are included in the above sets.

The caution mentioned in connection with the additive perturbation bears repeating here: the above multiplicative characterizations should *not* be interpreted as saying that the actual plant is the *cascade combination* of the nominal system P_0 and a system $1 + W\Delta$. Rather, the actual plant should be considered as being a *minimal realization* of the transfer function $P(s)$, which happens to be written in the multiplicative form.

Any unstable poles of P are poles of the nominal plant, but not necessarily the other way, because unstable poles of P_0 may be cancelled by zeros of $I + W\Delta$. In other words, the actual plant is allowed to have fewer unstable poles than the nominal plant, but all its unstable poles are confined to the same locations as in the nominal model. In view of the caution in the previous paragraph, such cancellations do *not* correspond to unstable hidden modes, and are therefore not of concern.

As in the case of additive perturbations, suppose we have a set of possible plants Π such that the true plant is a member of that set. We can try to embed this set in a multiplicative perturbation structure. First let $P_0 \in \Pi$ a certain nominal plant in Π. For any other plant $P \in \Pi$ we have,

$$P(j\omega) = P_0(j\omega)(1 + W(j\omega)\Delta(j\omega)).$$

The weight $|W(j\omega)|$ satisfies

$$\begin{aligned} |W(j\omega)| &\geq |W(j\omega)\Delta(j\omega)| = \left|\frac{P(j\omega) - P_0(j\omega)}{P_0(j\omega)}\right| \\ |W(j\omega)| &\geq \max_{P\in\Pi}\left|\frac{P(j\omega) - P_0(j\omega)}{P_0(j\omega)}\right| = \ell_m(j\omega). \end{aligned}$$

With the knowledge of the envelope $\ell_m(j\omega)$, we find a stable system $W(s)$ such that $|W(j\omega)| \geq \ell_m(j\omega)$

Example 19.1 Uncertain Gain

Suppose we have a plant $P = k\bar{P}(s)$ with an uncertain gain k that lies in the interval $k_1 \leq k \leq k_2$. We can write $k = \alpha(1 + \beta x)$ such that

$$\begin{aligned} k_1 &= \alpha(1 - \beta) \\ k_2 &= \alpha(1 + \beta). \end{aligned}$$

Therefore $\alpha = \frac{k_1+k_2}{2}$, $\beta = \frac{k_2-k_1}{k_2+k_1}$, and we can express the set of plants as

$$\Pi = \left\{P(s)|P(s) = \frac{k_1 + k_2}{2}\bar{P}(s)\left(1 + \frac{k_2 - k_1}{k_2 + k_1}x\right), -1 \leq x \leq 1\right\}.$$

We can embed this Π in a multiplicative structure by enlarging the uncertain elements x which are real numbers to complex $\Delta(j\omega)$ representing dynamic perturbations. This results in the following set

$$\Omega_m = \left\{P(s)|P(s) = \frac{k_1 + k_2}{2}\bar{P}(s)\left(1 + \frac{k_2 - k_1}{k_2 + k_1}\Delta\right), \|\Delta\|_\infty \leq 1\right\}.$$

Note that in this representation $P_0 = \frac{k_1+k_2}{2}\bar{P}$, and $W = \frac{k_2-k_1}{k_2+k_1}$.

Example 19.2 Uncertain Delay

Suppose we have a plant $P = e^{-ks}P_0(s)$ with an uncertain delay $0 \le k \le k_1$. We want to represent this family of plants in a multiplicative perturbation structure. The weight $W(s)$ should satisfy

$$\begin{aligned} |W(j\omega)| &\geq \max_{0\le k\le k_1} \left| \frac{e^{-j\omega k}P_0(j\omega) - P_0(j\omega)}{P_0(j\omega)} \right| \\ &= \max_{0\le k\le k_1} |e^{-j\omega k} - 1| \\ &= \begin{cases} |1 - e^{-j\omega k_1}| & \omega < \frac{\pi}{k_1} \\ 0 & \omega \geq \frac{\pi}{k_1} \end{cases} \\ &= \ell_m(\omega). \end{aligned}$$

A stable weight that satisfies the above relation can be taken as

$$W(s) = \alpha \frac{2\pi k_1 s}{\pi k_1 s + 1}.$$

where $\alpha > 1$. The reader should verify that this weight will work by ploting $|W(j\omega)|$ and $\ell_m(\omega)$, and showing that $\ell_m(\omega)$ is below the curve $|W(j\omega)|$ for all ω.

19.4 The Nyquist Criterion

Before we analyze the stability of feedback loops where the plant is uncertain, we will review the Nyquist criterion. Consider the feedback structure in Figure 19.3. The transfer function

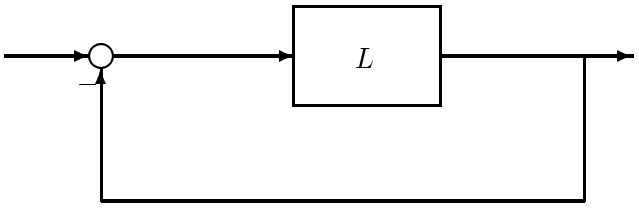


Figure 19.3: Unity Feedback Confuguration.

L is called the open-loop transfer function. The condition for the stability of the system in 19.3 is assured if the zeros of $1 + L$ are all in the left half of the complex plane. The agrument principle from complex analysis gives a criterion to calculate the difference between the number of zeros and the number of poles of an analytic function in a certain domain, $\mathcal{D}$ in the complex plane. Suppose the domain is as shown in Figure 19.4, and the boundary of $\mathcal{D}$, denoted by $\delta\mathcal{D}$, is oriented clockwise. We call this oriented boundary of $\mathcal{D}$ the Nyquist contour.

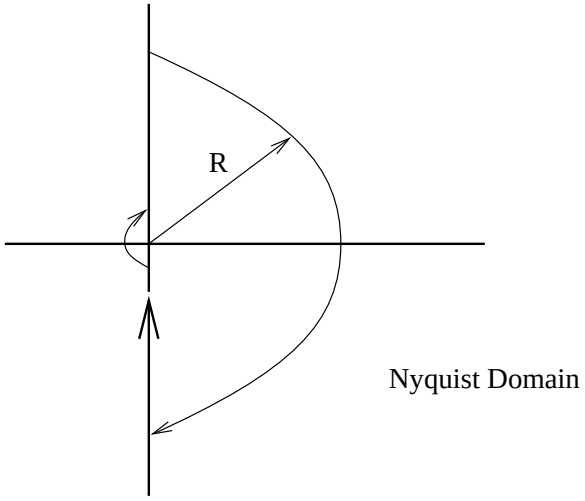


Figure 19.4: Nyquist Domain.

As the radius of the semicircle in Figure 19.4 goes to infinty the domain covers the right half of the complex plane. The image of $\delta\mathcal{D}$ under L is called a Nyquist plot, see Figure 19.5. Note that if L has poles at the $j\omega$ axis then we indent the Nyquist contour to avoid these poles, as shown in Figure 19.4. Define

$$\begin{aligned}\pi_{ol} &= \text{Open}-\text{loop poles} = \text{Number of poles of } L \text{ in } \mathcal{D} = \text{Number of poles of } 1+L \text{ in } \mathcal{D}\\ \pi_{cl} &= \text{Closed}-\text{loop poles} = \text{Number of zeros of } 1+L \text{ in } \mathcal{D}.\end{aligned}$$

From the argument principle it follows that

$$\pi_{cl}-\pi_{ol} = \text{The number of clockwise encirclements that the Nyquist Plot makes of the point } -1.$$

Using this characterization of the difference of the number of the closed-loop poles and the open-loop poles we arrive at the following theorem for the stability of Figure 19.3

Theorem 19.1 *The closed-loop system in Figure 19.3 is stable if and only if the Nyquist plot*

- *does not pass through the origin,*
- *makes π_{ol} counter-clockwise encirclements of -1.*

19.5 Robust Stability

In this section we will show how we can analyze the stability of a feedback system when the plant is uncertain and is known to belong to a set of the form that we described earlier. We will start with the case of additive pertubations. Consider the unity feedback configuration in Figure 19.1. The open-loop transfer function is $L(s) = (P_0(s) + W(s)\Delta(s))K(s)$, and the

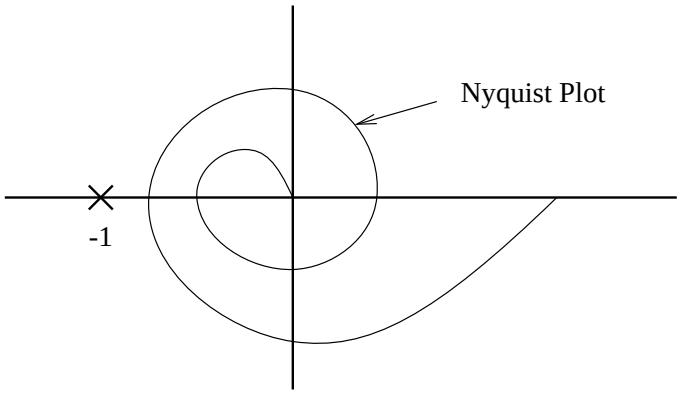


Figure 19.5: Nyquist Plot.

nominal open-loop transfer function is $L_0(s) = P_0(s)K(s)$. The nominal feedback system with the nominal open-loop transfer function L_0 is stable, and we want to know whether the feeback system remains stable for all $\Delta(s)$ satisfying $|\Delta(j\omega)| \leq 1$ for all $\omega \in \mathbb{R}$. We will assume that the nominal open-loop system is stable. This causes no loss of generality and the result holds in the general case. From the Nyquist criterion, we have that the Nyquist plot of L_0 does not encircle the point -1. For the perturbed system, we have that

$$\begin{aligned} 1 + L(j\omega) &= 1 + P(j\omega)K(j\omega) \\ &= 1 + (P_0(j\omega) + W(j\omega)\Delta(j\omega))K(j\omega) \\ &= 1 + L_0(j\omega) + W(j\omega)\Delta(j\omega)K(j\omega) \end{aligned}$$

From the Figure 19.6, it is clear that $L(\jmath\omega)$ will not encircle the point -1 if the following condition is satisfied,

$$|W(j\omega)K(j\omega)| < |1 + L_0(j\omega)|,$$

which can be written as

$$\left| \frac{W(j\omega)K(j\omega)}{1 + L_0(j\omega)} \right| < 1. \tag{19.6}$$

A Small Gain Argument

Next we will present a different derivation of the above result that does not rely on the Nyquist criterion, and will be the basis for the multivariable generalizations of the robust stability results. Since the nominal feedback system is stable, the zeros of $1 + L_0(s)$ are all in the left half of the complex plane. Therefore, by the continuity of zeros, the perturbed system

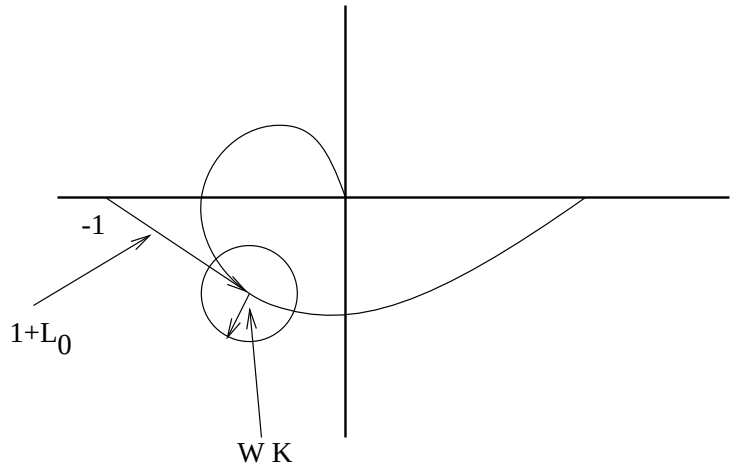


Figure 19.6: Nyquist Plot Illustrating Robust Stability.

will be stable if and only if

$$|1 + (P_0(j\omega) + W(j\omega)\Delta(j\omega))K(j\omega)| > 0$$

for all $\omega \in \mathbb{R}$, $\|\Delta\|_\infty \leq 1$. By rearranging the terms, the perturbed system is stable if and only if

$$\min_{|\Delta(j\omega)|\leq 1} \left|1 + \frac{W(j\omega)K(j\omega)}{1 + P_0(j\omega)K(j\omega)}\Delta(j\omega)\right| > 0 \quad \text{for all } \omega \in \mathbb{R}$$

The following lemma will help us to transform this condition to the one given earlier.

Lemma 19.1 *The following are equivalent*

1.

$$\min_{|\Delta(j\omega)|\leq 1} \left|1 + \frac{W(j\omega)K(j\omega)}{1 + P_0(j\omega)K(j\omega)}\Delta(j\omega)\right| > 0 \quad \textit{for all } \omega \in \mathbb{R}$$

2.

$$1 - \left|\frac{W(j\omega)K(j\omega)}{1 + P_0(j\omega)K(j\omega)}\right| > 0 \quad \textit{for all } \omega \in \mathbb{R}$$

Proof. First we show that 2) implies 1), which is a consequence of the following inequalities

$$\begin{aligned}\left|1 + \frac{W(j\omega)K(j\omega)}{1 + P_0(j\omega)K(j\omega)}\Delta(j\omega)\right| &\geq 1 - \left|\frac{W(j\omega)K(j\omega)}{1 + P_0(j\omega)K(j\omega)}\Delta(j\omega)\right| \\ &\geq 1 - \left|\frac{W(j\omega)K(j\omega)}{1 + P_0(j\omega)K(j\omega)}\right|.\end{aligned}$$

For the converse suppose 2) is violated, that is there exists ω_0 such that

$$\left| \frac{W(j\omega_0)K(j\omega_0)}{1 + P_0(j\omega_0)K(j\omega_0)} \right| \geq 1.$$

Write

$$\frac{W(j\omega_0)K(j\omega_0)}{1 + P_0(j\omega_0)K(j\omega_0)} = a e^{j\phi},$$

and let $\bar{\Delta}(j\omega_0) = \frac{1}{a} e^{-j\phi - j\pi}$. Clearly, $|\bar{\Delta}(j\omega_0)| \leq 1$ and

$$1 + \frac{W(j\omega_0)K(j\omega_0)}{1 + P_0(j\omega_0)K(j\omega_0)} \bar{\Delta}(j\omega_0) = 0.$$

Now select a real rational perturbation $\bar{\Delta}(s)$ as

$$\bar{\Delta}(s) = \pm \frac{1}{a} \frac{s - \alpha}{s + \alpha},$$

such that $\pm \frac{j\omega_0 - \alpha}{\omega_0 + \alpha} = e^{-j\phi - j\pi}$.

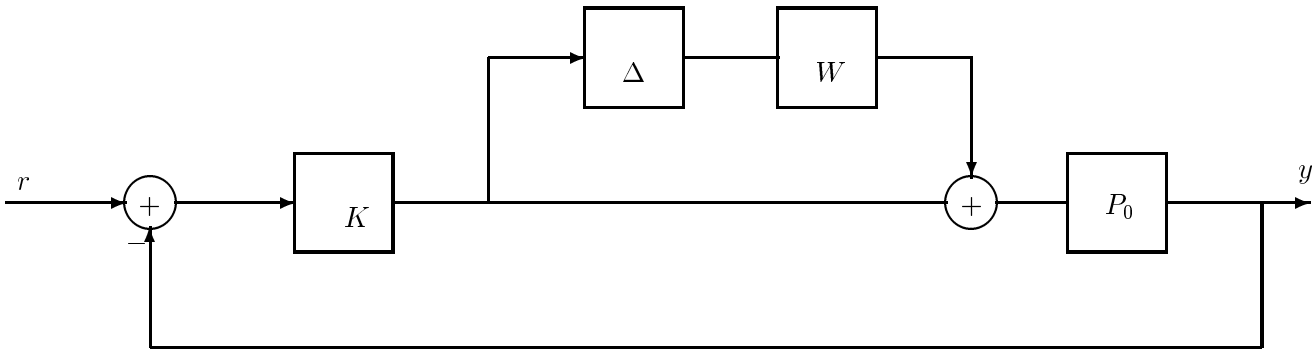


Figure 19.7: Representation of the actual plant in a servo loop via a multiplicative perturbation of the nominal plant.

A similar set of results can be obtained for the case of multiplicative perturbations. In particular, a robust stability of the configuration in Figure 19.7 can be guaranteed if the system is stable for the nominal plant P_0 and

$$\left| \frac{W(j\omega)P_0(j\omega)K(j\omega)}{1 + P_0(j\omega)K(j\omega)} \right| < 1. \quad \text{for all } \omega \in \mathbb{R}. \tag{19.7}$$

Example 19.3 Stabilizing a Beam

We are interested in deriving a controller that stabilizes the beam in Figure 19.8 and tracks a step input (with good properties). The rigid body model from torque input to the tip deflection is given by

$$P_0(s) = \frac{6.28}{s^2}$$

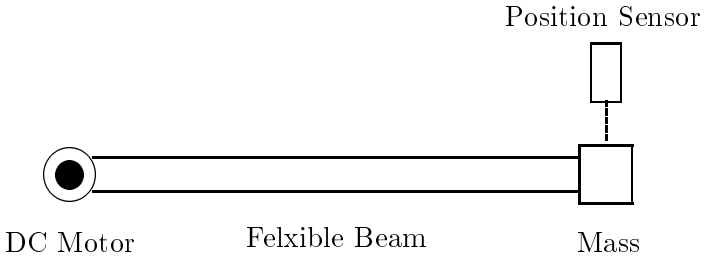


Figure 19.8: Flexible Beam.

Consider the controller

$$K_0(s) = \frac{500(s+10)}{s+100}$$

The loop gain is given by

$$P_0(s)K_0(s) = \frac{3140(s+10)}{s^2(s+100)}$$

and is shown in Figure 19.9. The closed loop poles are located at -49.0, -28.6, -22.4, and the nominal Sensitivity function is given by

$$S_0(s) = \frac{1}{1+P_0(s)K_0(s)} = \frac{s^2(s+100)}{s^3+100s^2+3140s+31400}$$

and is shown in Figure 19.10. It is evident from this that the system has good disturbance rejection and tracking properties. The closed loop step response is show in Figure 19.11

While this controller seems to be an excellent design, it turns out that it performs quite poorly in practice. The bandwidth of this controller (which was never constrained) is large enough to excite the flexible modes of the beam, which were not taken into account in the model. A more complicated model of the beam is given by

$$P_1(s) = \underbrace{\frac{6.28}{s^2}}_{\text{nominal plant}} + \underbrace{\frac{12.56}{s^2+0.707s+28}}_{\text{flexible mode}}$$

If K_0 is connected to this plant, then the closed loop poles are -1.24, 0.29, 0.06, -0.06, which implies instability.

Instead of using the new model to redesign the controller, we would like to use the nominal model P_0, and account for the flexible modes as unmodeled dynamics with a certain frequency concentration. There are several advantages in this. For

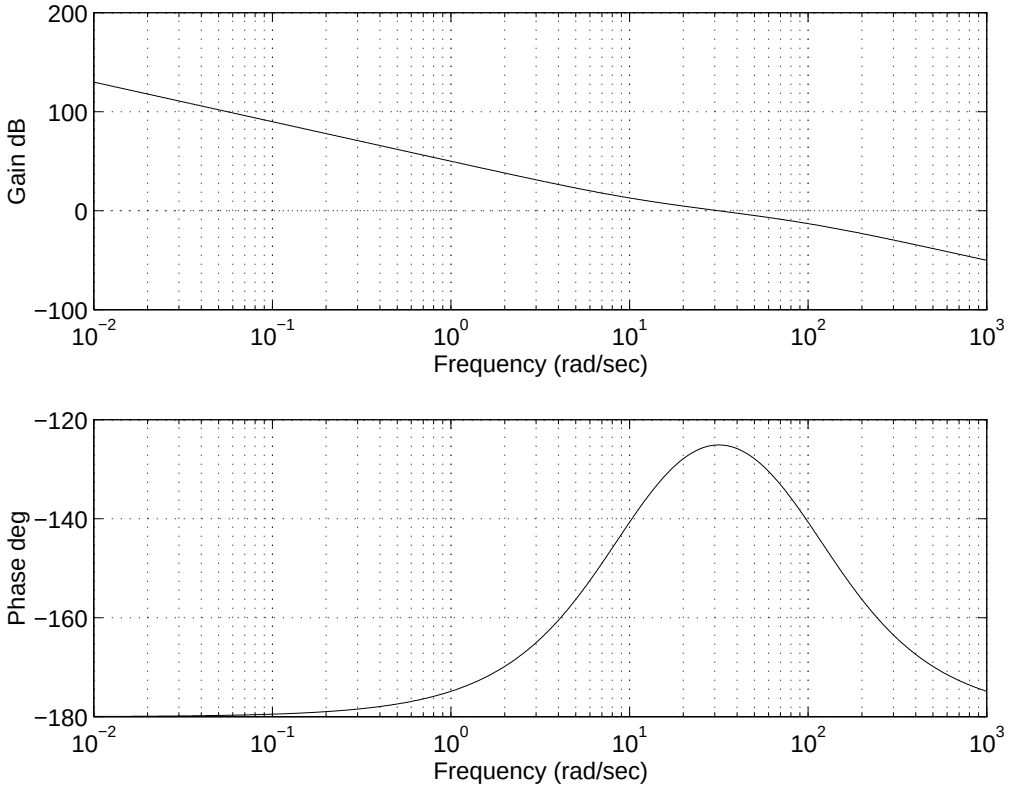


Figure 19.9: Open-loop Bode Plot

one, the design is based on a simpler nominal model and hence may result in a simpler controller. This approach also allows us to acomodate additional flexible modes without increasing the complexity of the description. And finally, it enables us to tradeoff performance for robustness.

Consider the set of plants:

$$\Omega = \{P = P_0(1+\Delta);\ \ |\Delta(j\omega)| \leq \ell(\omega), \Delta \text{ is stable}\}$$

where

$$\ell(\omega) \leq 2 \left| \frac{\omega^2}{28 - \omega^2 + 0.707 j\omega} \right|$$

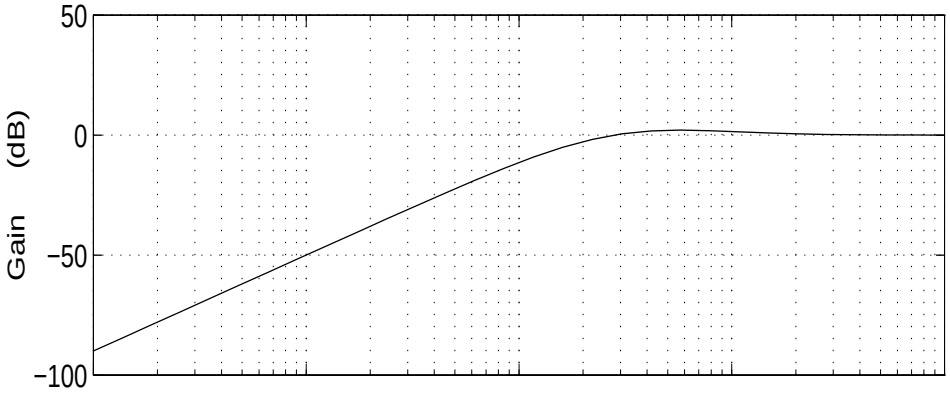


Figure 19.10: Nominal Sensitivity

This set includes the model P_1. The stability Robustness Condition is given by:

$$|T(j\omega)| < \frac{1}{\ell(\omega)}$$

Where T is the nominal closed loop map with any controller K. First, consider the stability analysis of the initial controller $K_0(s)$. Figure 19.12 shows both the frequency response for $|T_0(j\omega)|$ and $[\ell(\omega)]^{-1}$. It is evident that the Stability robustness condition is violated since

$$|T_0(j\omega)| \not< \frac{1}{\ell(\omega)}, \quad 3 \leq \omega \leq 70 \text{ rad/sec}$$

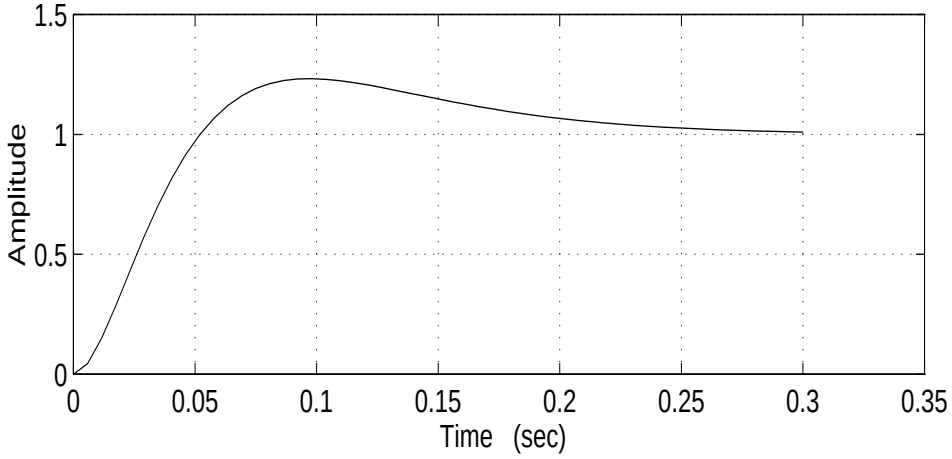


Figure 19.11: Step Response

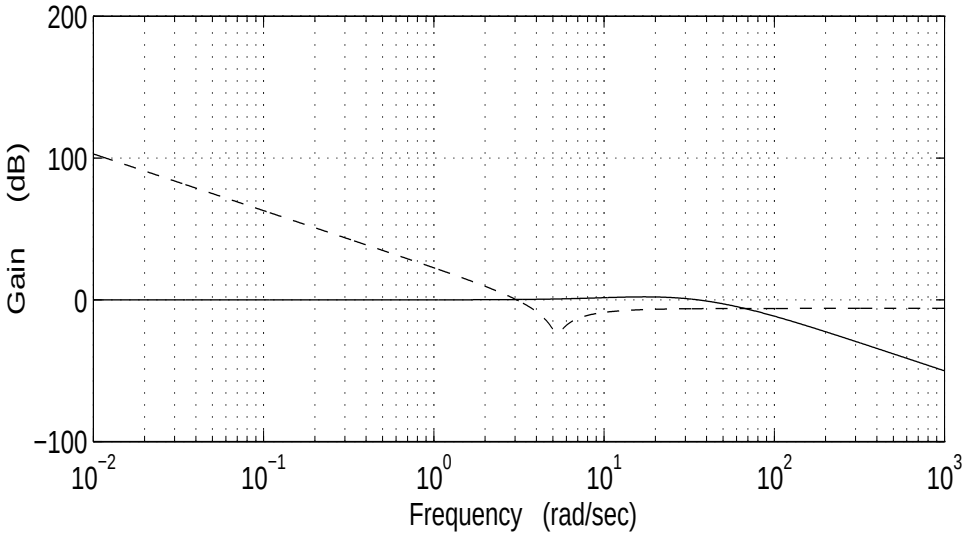


Figure 19.12: $|T_0(j\omega)|$ and $[\ell(\omega)]^{-1}$

Let's try a new design with a different controller

$$K_1(s) = \frac{(5 \times 10^{-4})(s + 0.01)}{s + 0.1}$$

The new loop-gain is

$$P_0(s)K_1(s) = \frac{(3.14 \times 10^{-3})(s + 0.01)}{s^2(s + 0.1)}$$

which is shown in the Figure 19.13 We first check the robustness condition with the new controller. T_1 is given by

$$T_1(s) = \frac{P_0(s)K_1(s)}{1 + P_0(s)K_1(s)}$$

Figure 19.14 depicts both $|T_1(j\omega)|$ and $[\ell(\omega)]^{-1}$. It is clear that the condition is satisfied. Figure 19.15 shows the new nominal step response of the system. Observe that the response is much slower than the one derived by the controller K_0. This is essentially due to the limited bandwidth of the new controller, which was necessary to prevent instability.

Exercises

Exercise 19.1 Suppose $P(s) = \frac{a}{s}$ is connected with a controller $K(s)$ in a unity feedback configuration. Does there exists a K such that the system is stable for both $a = 1$ and $a = -1$.

Exercise 19.2 For $P(s)$ and $K(s)$ given by

$$P(s) = \frac{1}{(s+2)(s+a)}, \qquad K(s) = \frac{1}{s},$$

find the range of a such that the closed loop system with P and K is stable.

Exercise 19.3 Let P be given by:

$$P(s) = (1 + W(s)\Delta(s))P_0,$$

where

$$P_0(s) = \frac{1}{s-1}, \qquad W(s) = \frac{2}{s+10},$$

and Δ is arbitrary stable with $\|\Delta\|_\infty \leq 2$. Find a controller $K(s) = k$ (constant) gain such that the system is stable. Compute all possible such gains.

Exercise 19.4 Find the stability robustness condition for the set of plant described by:

$$P = \{\frac{P_0}{1 + \Delta W P_0}, \qquad \|\Delta\|_\infty \leq 1\}.$$

Assume WP_0 is strictly proper for well posedness.

Exercise 19.5 Suppose

$$P(s) = \frac{1}{s-a} \text{and} \qquad \text{K(s)} = 10,$$

are connected in standard feedback configuration. While it is easy in this case to compute the exact stability margin as a changes, in general, such problems are hard to solve when there are many parameters. One approach is to embed the problem in a robust stabilization problem with unmodeled dynamics and derive the appropriate stability robustness condition. Clearly, the later provides a conservative bound on a for which the system remains stable.

(a) Find the exact range of a for which the system is stable.

(b) Assume the nominal plant is $P_0 = \frac{1}{s}$. Show that P belongs to the set of plants:

$$\Omega = \{P = \frac{P_0}{1 + W\Delta P_0}, \|\Delta\|_\infty \leq 1\}$$

and $W = -a$.

(c) Derive a condition on the closed loop system that guarantees the stability of the set Ω. How does this condition constrain a? Is this different than part (a)?

(d) Repeat with nominal plant $P_0 = \frac{1}{s+100}$.

Exercise 19.6 Let a model be given by the stable plant:

$$P_0(z) = \frac{1}{z^{-1} - (1 + a_0)}, \quad 1 >> a_0 > 0.$$

Consider the class of plants given by:

$$\Omega = \left\{ (z) = \frac{1}{z^{-1} - (1 + b)} | \ -2a_0 \leq b \leq 2a_0 \right\}.$$

1. Can the set Ω be embedded in a set of additive or multiplicative norm bounded perturbations, with nominal plant P_0? Show how or explain your answer.
2. If your answer to the previous part is NO, show that the class Ω can be embedded in some other larger set characterized by norm-bounded perturbations. Give a sufficient condition for stability using the small gain theorem.
3. Improve your earlier condition so that it captures the fact that the unknown is a real parameter. (The condition does not have to be necessary, but should still take into consideration the phase information!).

Exercise 19.7 Consider Exercise 17.4. Suppose that due to implementation problems (e.g. quantization effects), the actual controller can be modeled as:

$$K_a = (I - KW\Delta)^{-1}K$$

where W is a fixed stable filter, and is a stable perturbation of $\mathcal{H}_\infty$-norm less than 1, but otherwise arbitrary. Provide a non-conservative condition for the stability robustness of the closed loop system. Use the parametrization of K in terms of Q to express your condition as a function of P and Q.

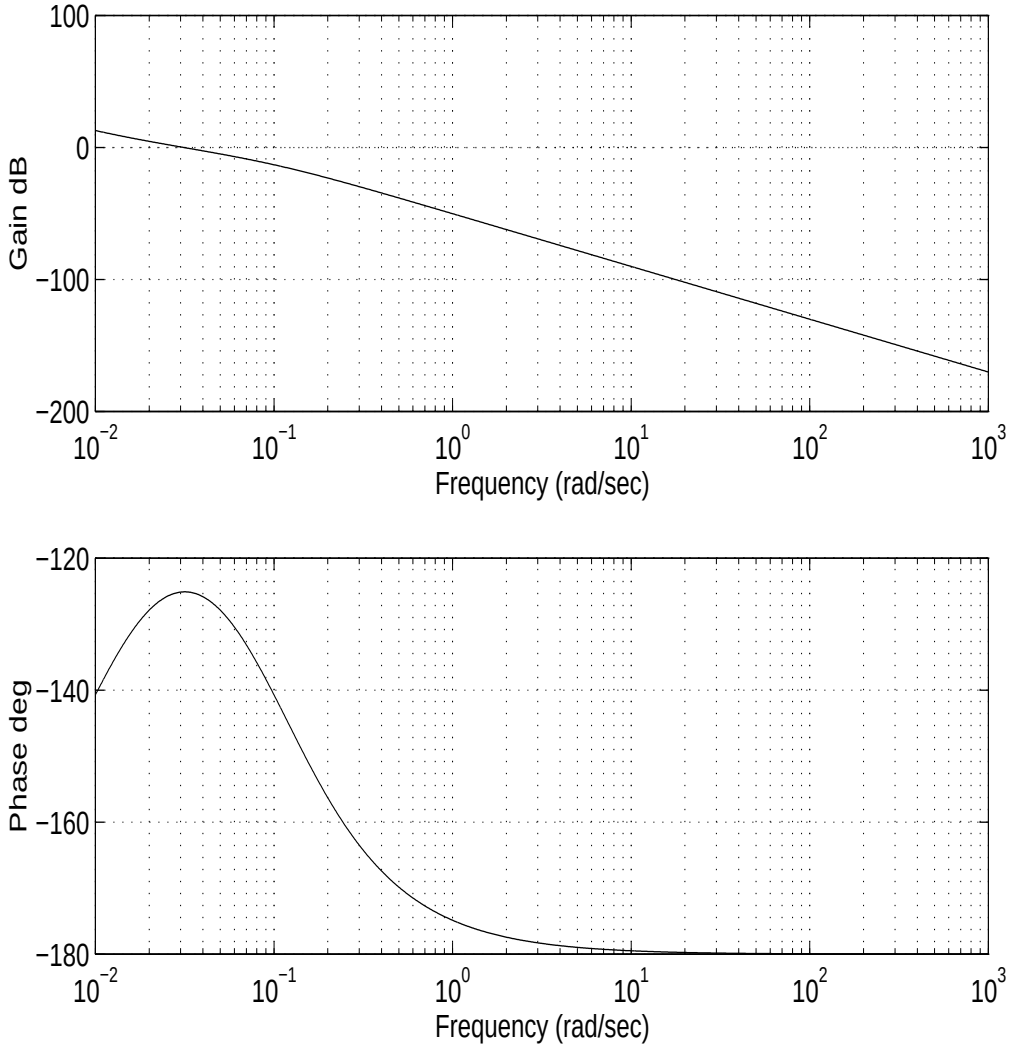


Figure 19.13: Loop Gain P_0K_1

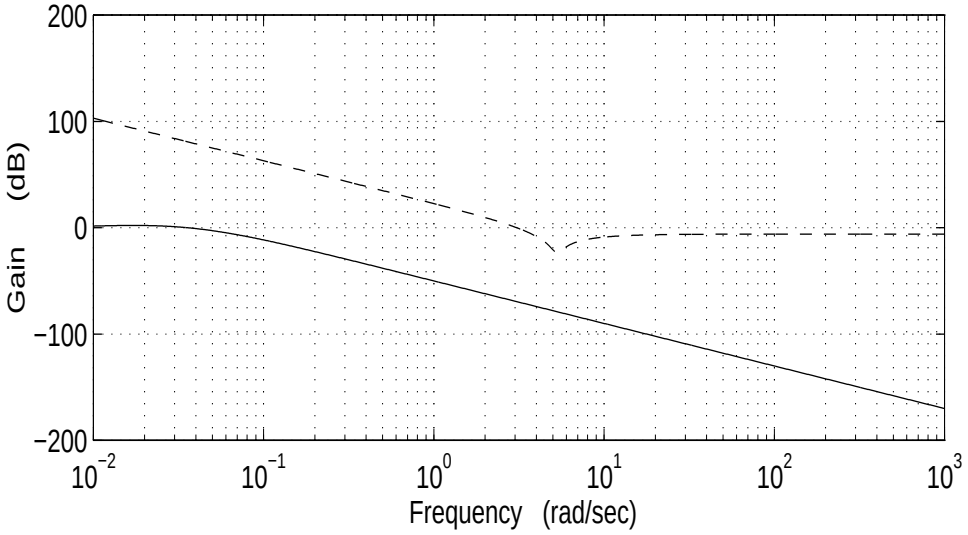


Figure 19.14: $|T_1(j\omega)|$ and $[\ell(\omega)]^{-1}$.

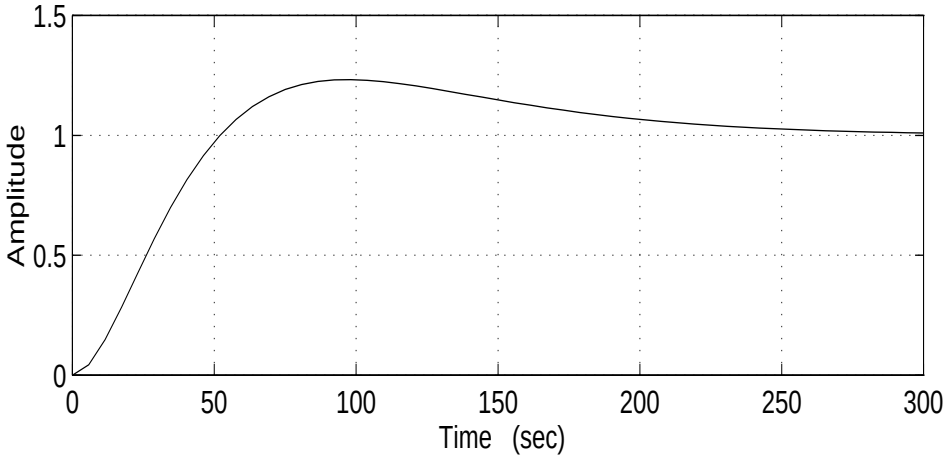


Figure 19.15: New Nominal Closed-loop Step Response

Chapter 20

Stability Robustness

20.1 Introduction

Last chapter showed how the Nyquist stability criterion provides conditions for the stability robustness of a SISO system. It is possible to provide an extension of those conditions by generalizing the Nyquist criterion for MIMO systems. This, however, turns out to be unnecessary and a direct derivation is possible through the small gain theorem, which will be presented in this chapter.

20.2 Additive Representation of Uncertainty

It is commonly the case that the nominal plant model is quite accurate for low frequencies but deteriorates in the high-frequency range, because of parasitics, nonlinearities and/or time-varying effects that become significant at higher frequencies. These high-frequency effects may have been left unmodeled because the effort required for system identification was not justified by the level of performance that was being sought, or they may be well-understood effects that were omitted from the nominal model because they were awkward and unwieldy to carry along during control design. This problem, namely the deterioration of nominal models at higher frequencies, is mitigated to some extent by the fact that almost all physical systems have strictly proper transfer functions, so that the system gain begins to roll off at high frequency.

In the above situation, with a nominal plant model given by the proper rational matrix $P_0(s)$, the actual plant represented by $P(s)$, and the difference $P(s) - P_0(s)$ assumed to be stable, we may be able to characterize the model uncertainty via a bound of the form

$$\sigma_{max}\left[P(j\omega) - P_0(j\omega)\right] \leq \ell_a(\omega) \tag{20.1}$$

where

$$\ell_a(\omega) = \begin{cases} \text{“Small”} & ; \quad |\omega| < \omega_c \\ \text{“Bounded”} & ; \quad |\omega| > \omega_c \end{cases} \tag{20.2}$$

This says that the response of the actual plant lies in a "band" of uncertainty around that of the nominal plant. Notice that no phase information about the modeling error is incorporated into this description. For this reason, it may lead to conservative results.

The preceding description suggests the following simple *additive* characterization of the uncertainty set:

$$\Omega = \{P(s) \mid P(s) = P_0(s) + W(s)\Delta(s)\} \tag{20.3}$$

where Δ is an arbitrary *stable* transfer matrix satisfying the norm condition

$$\|\Delta\|_\infty = \sup_\omega \sigma_{max}(\Delta(j\omega)) \leq 1 \tag{20.4}$$

and the *stable* proper rational (matrix or scalar) weighting term $W(s)$ is used to represent any information we have on how the accuracy of the nominal plant model varies as a function of frequency. Figure 20.1 shows the additive representation of uncertainty in the context of a standard servo loop, with K denoting the compensator.

When the modeling uncertainty increases with frequency, it makes sense to use a weighting function $W(j\omega)$ that looks like a high-pass filter: small magnitude at low frequencies, increasing but bounded at higher frequencies. In the case of a matrix weight, a variation on the use of the additive term $W\Delta$ is to use a term of the form $W_1\Delta W_2$; we leave you to examine how the analysis in this lecture will change if such a two-sided weighting is used.

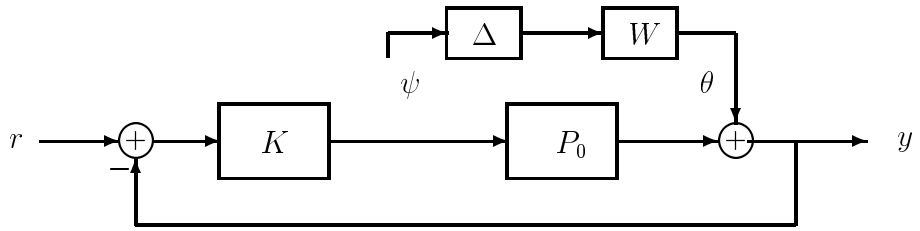


Figure 20.1: Representation of the actual plant in a servo loop via an additive perturbation of the nominal plant.

Caution: The above formulation of an additive model perturbation should *not* be interpreted as saying that the actual or perturbed plant is the *parallel combination* of the nominal system $P_0(s)$ and a system with transfer matrix $W(s)\Delta(s)$. Rather, the actual plant should be considered as being a *minimal realization* of the transfer function $P(s)$, which happens to be written in the additive form $P_0(s) + W(s)\Delta(s)$.

Some features of the above uncertainty set are worth noting:

- The *unstable* poles of all plants in the set are precisely those of the nominal model. Thus, our modeling and identification efforts are assumed to be careful enough to accurately capture the unstable poles of the system.

- The set includes models of arbitrarily large order. Thus, if the uncertainties of major concern to us were *parametric uncertainties*, i.e. uncertainties in the values of the

parameters of a particular (e.g. state-space) model, then the above uncertainty set would greatly overestimate the set of plants of interest to us.

The control design methods that we shall develop will produce controllers that are guaranteed to work for *every member* of the plant uncertainty set. Stated slightly differently, our methods will treat the system as though *every* model in the uncertainty set is a possible representation of the plant. To the extent that not all members of the set are possible plant models, our methods will be conservative.

20.3 Multiplicative Representation of Uncertainty

Another simple means of representing uncertainty that has some nice analytical properties is the *multiplicative perturbation*, which can be written in the form

$$\Omega = \{P \mid P = P_0(I + W\Delta),\ \|\Delta\|_\infty \leq 1\}. \tag{20.5}$$

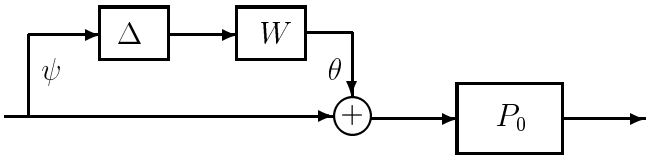


Figure 20.2: Representation of uncertainty as multiplicative perturbation at the plant input.

An alternative to this input-side representation of the uncertainty is the following output-side representation:

$$\Omega = \{P \mid P = (I + W\Delta)P_0,\ \|\Delta\|_\infty \leq 1\}. \tag{20.6}$$

In both the multiplicative cases above, W and Δ are stable. As with the additive representation, models of arbitrarily large order are included in the above sets. Still other variations may be imagined; in the case of matrix weights, for instance, the term $W\Delta$ can be replaced by $W_1\Delta W_2$.

The caution mentioned in connection with the additive perturbation bears repeating here: the above multiplicative characterizations should *not* be interpreted as saying that the actual plant is the *cascade combination* of the nominal system P_0 and a system $I + W\Delta$. Rather, the actual plant should be considered as being a *minimal realization* of the transfer function $P(s)$, which happens to be written in the multiplicative form.

Any unstable poles of P are poles of the nominal plant, but not necessarily the other way, because unstable poles of P_0 may be cancelled by zeros of $I + W\Delta$. In other words, the actual plant is allowed to have fewer unstable poles than the nominal plant, but all its unstable poles are confined to the same locations as in the nominal model. In view of the caution in the previous paragraph, such cancellations do *not* correspond to unstable hidden modes, and are therefore not of concern.

20.4 More General Representation of Uncertainty

Consider a nominal interconnected system obtained by interconnecting various (reachable and observable) nominal subsystems. In general, our representation of the uncertainty regarding any nominal subsystem model such as P_0 involves taking the signal ψ at the input or output of the nominal subsystem, feeding it through an "uncertainty block" with transfer function $W\Delta$ or $W_1\Delta W_2$, where each factor is stable and $\|\Delta\|_\infty \leq 1$, and then adding the output θ of this uncertainty block to either the input or output of the nominal subsystem. The one additive and two multiplicative representations described earlier are special cases of this construction, but the construction actually yields a total of three additional possibilities with a given uncertainty block. Specifically, if the uncertainty block is $W\Delta$, we get the following additional *feedback representations* of uncertainty:

- $P = P_0(I - W\Delta P_0)^{-1}$;
- $P = P_0(I - W\Delta)^{-1}$;
- $P = (I - W\Delta)^{-1}P_0$.

A useful feature of the three uncertainty representations itemized above is that the unstable poles of the actual plant P are not constrained to be (a subset of) those of the nominal plant P_0.

Note that in all six representations of the perturbed or actual system, the signals ψ and θ become *internal* to the actual subsystem model. This is because it is the combination of P_0 with the uncertainty model that constitutes the representation of the actual model P, and the actual model is only accessed at its (overall) input and output.

In summary, then, perturbations of the above form can be used to represent many types of uncertainty, for example: high-frequency unmodeled dynamics, unmodeled delays, unmodeled sensor and/or actuator dynamics, small nonlinearities, parametric variations.

20.5 A Linear Fractional Description

We start with a given a nominal plant model P_0, and a feedback controller K that stabilizes P_0. The robust stability question is then: under what conditions will the controller stabilize *all* $P \in \Omega$? More generally, we assume we have an interconnected system that is nominally internally stable, by which we mean that the transfer function from an input added in at any subsystem input to the output observed at any subsystem output is always stable in the nominal system. The robust stability question is then: under what conditions will the interconnected system remain internally stable for all possible perturbed models.

If the plant uncertainty is specified (additively, multiplicatively, or using a feedback representation) via an uncertainty block of the form $W\Delta$, where W and Δ are stable, then the actual (closed-loop) system can be mapped into the very simple feedback configuration

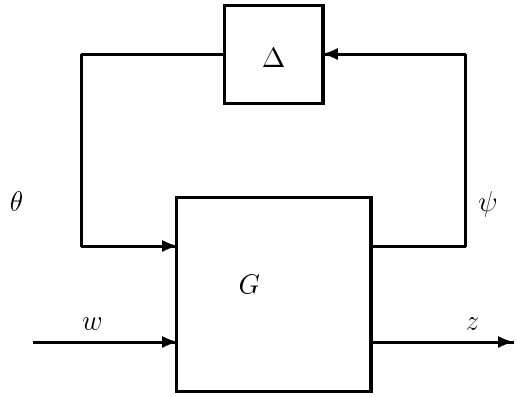


Figure 20.3: Standard model for uncertainty.

shown in Figure 20.3. (The generalization to an uncertainty block of the form $W_1 \Delta W_2$ is trivial, and omitted here to avoid additional notation.)
As in the previous subsection, the signals ψ and θ respectively denote the input and output of the uncertainty block. The input w is added in at some arbitrary *accessible point* of the interconnected system, and z denotes an output taken from an arbitrary accessible point. An accessible point in our terminology is simply some subsystem input or output in the *actual* or perturbed system; the input ψ and output θ of the uncertainty block would *not* qualify as accessible points.

If we remove the perturbation block Δ in Fig. 20.3, we are left with the nominal closed-loop system, which is stable by hypothesis (since the compensator K has been chosen to stabilize the nominal plant and is lumped in G). Stability of the nominal system implies that the transfer functions relating the outputs ψ and z of the nominal system to the inputs θ and w are all stable. Thus, in the transfer function representation

$$\begin{pmatrix} \Psi(s) \\ Z(s) \end{pmatrix} = \begin{pmatrix} M(s) & N(s) \\ J(s) & L(s) \end{pmatrix} \begin{pmatrix} \Theta(s) \\ W(s) \end{pmatrix} \tag{20.7}$$

each of the transfer matrices M, N, J, and L is stable.

Now incorporating the constraint imposed by the perturbation, namely

$$\Theta = (\Delta)\, \Psi \tag{20.8}$$

and solving for the transfer function relating z to w in the *perturbed* system, we obtain

$$G_{wz}(s) = L + J\Delta(I - M\Delta)^{-1}N. \tag{20.9}$$

Note that M is the transfer function "seen" by the perturbation Δ, from the input θ that it imposes on the rest of the system, to the output ψ that it measures from the rest of the system. Recalling that w and z denoted arbitrary inputs and outputs at the accessible points of the actual closed-loop system, we see that internal stability of the actual (i.e. perturbed) closed-loop system requires the above transfer function be stable for all allowed Δ.

20.6 The Small-Gain Theorem

Since every term in G_{wz} other than $(I-M\Delta)^{-1}$ is known to be stable, we shall have stability of G_{wz}, and hence guaranteed stability of the actual closed-loop system, *if* $(I-M\Delta)^{-1}$ is stable for all allowed Δ. In what follows, we will arrive at a condition — the *small-gain condition* — that guarantees the stability of $(I-M\Delta)^{-1}$. It can also be shown (see Appendix) that if this condition is violated, then there is a stable Δ with $\|\Delta\|_\infty \leq 1$ such that $(I-M\Delta)^{-1}$ and $\Delta(I-M\Delta)^{-1}$ are unstable, and G_{wz} is unstable for some choice of z and w.

Theorem 20.1 ("Unstructured" Small-Gain Theorem) *Define the set of stable perturbation matrices* $\mathbb{\Delta} \stackrel{\triangle}{=} \{\Delta \mid \|\Delta\|_\infty \leq 1\}$. *If* M *is stable, then* $(I-M\Delta)^{-1}$ *and* $\Delta(I-M\Delta)^{-1}$ *are stable for each* Δ *in* $\mathbb{\Delta}$ *if and only if* $\|M\|_\infty < 1$.

Proof. The proof of necessity (see Appendix) is by construction of an allowed Δ that causes $(I-M\Delta)^{-1}$ and $\Delta(I-M\Delta)^{-1}$ to be unstable if $\|M\|_\infty \geq 1$, and ensures that G_{wz} is unstable.

For here, we focus on the proof of sufficiency. We need to show that if $\|M\|_\infty < 1$ then $(I-M\Delta)^{-1}$ has no poles in the closed right half-plane for any $\Delta \in \mathbb{\Delta}$, or equivalently that $I-M\Delta$ has no zeros there. For arbitrary $x \neq 0$ and any s_+ in the closed right half-plane (CRHP), and using the fact that both M and Δ are well-defined throughout the CRHP, we can deduce that

$$\begin{aligned} \|[I-M(s_+)\Delta(s_+)]x\|_2 &\geq \|x\|_2 - \|M(s_+)\Delta(s_+)x\|_2 \\ &\geq \|x\|_2 - \sigma_{max}[M(s_+)\Delta(s_+)]\,\|x\|_2 \\ &\geq \|x\|_2 - \|M\|_\infty\,\|\Delta\|_\infty\|x\|_2 \\ &> 0 \end{aligned} \tag{20.10}$$

The first inequality above is a simple application of the triangle inequality. The third inequality above results from the *Maximum Modulus Theorem* of complex analysis, which says that the largest magnitude of a complex function over a region of the complex plane is found on the boundary of the region, if the function is analytic inside and on the boundary of the region. In our case, both $q'M'Mq$ and $q'\Delta'\Delta q$ are stable, and therefore analytic, in the CRHP, for unit vectors q; hence their largest values over the CRHP are found on the imaginary axis. The final inequality in the above set is a consequence of the hypotheses of the theorem, and establishes that $I-M\Delta$ is nonsingular — and therefore has no zeros — in the CRHP.

20.7 Stability Robustness Analysis

Next, we present a few examples to illustrate the use of the small-gain theorem in stability robustness analysis.

Example 20.1 (Additive Perturbation)

For the configuration in Figure 20.1, it is easily seen that

$$M = -K(I + P_0K)^{-1}W = -(I + KP_0)^{-1}KW$$

Example 20.2 (Multiplicative Perturbation)

A multiplicative perturbation of the form of Figure 20.2 can be inserted into the closed-loop system at either the plant input or output. The procedure is then identical to Example 20.1, except that M becomes a different function. Again it is easily verified that for a multiplicative perturbation at the plant input,

$$M = -(I + KP_0)^{-1}KP_0W, \tag{20.11}$$

while a perturbation at the output yields

$$M = -(I + P_0K)^{-1}P_0KW. \tag{20.12}$$

What the above examples show is that stability robustness requires ensuring the weighted versions of certain familiar transfer functions have $\mathcal{H}_\infty$ norms that are less than 1. For instance, with a multiplicative perturbation at the output as in the last example, what we require for stability robustness is $\|TW\|_\infty < 1$, where T is the complementary sensitivity function associated with the nominal closed-loop system. This condition evidently has the same flavor as the conditions we discussed earlier in connection with nominal performance of the closed-loop system.

The small-gain theorem fails to take advantage of any special structure that there might be in the uncertainty set $\mathbf{\Delta}$, and can therefore be very conservative. As examples of the kinds of situations that arise, consider the following two examples.

Example 20.3

Suppose we have a system that is best represented by the model of Figure 20.4. When this system is reduced to the standard form, Δ will have a block-diagonal

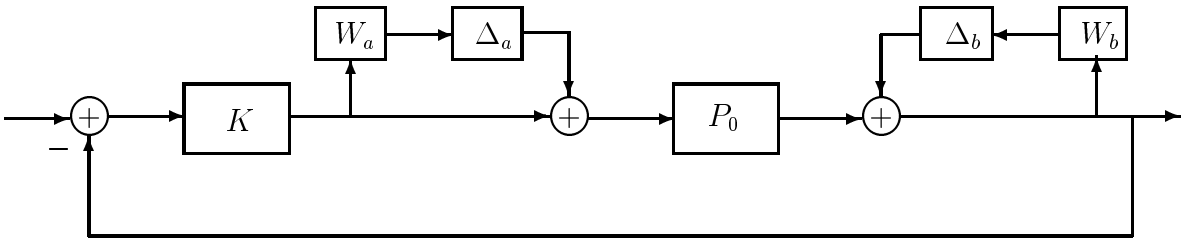


Figure 20.4: Plant with multiple uncertainties.

structure, since the two perturbations enter at different points in the system:

$$\Delta = \begin{bmatrix} \Delta_a & 0 \\ 0 & \Delta_b \end{bmatrix} \tag{20.13}$$

Thus, there is some added information about the plant uncertainty that cannot be captured by the unstructured small-gain theorem, and in general, even if $\|M\|_\infty \geq 1$ for the M that corresponds to the Δ above, there may be no admissible perturbation that will result in unstable $(I - M\Delta)^{-1}$.

Example 20.4

Suppose that in addition to norm bounds on the uncertainty, we know that the phase of the perturbation remains in the sector $[-30°, 30°]$. Again, even if $\|M\|_\infty \geq 1$ for the M that corresponds to the Δ for this system, there may be no admissible perturbation that will result in unstable $(I - M\Delta)^{-1}$.

In both of the preceding two examples, the unstructured small-gain theorem gives conservative results.

Relating Stability Robustness to the (SISO) Nyquist Criterion

Suppose we have a SISO nominal plant with a multiplicative perturbation, and a nominally stabilizing controller K. Then $P = P_0(1 + W\Delta)$, and the compensated open-loop transfer function is

$$PK = P_0K + P_0KW\Delta. \tag{20.14}$$

Since P_0, K, and W are known and $|\Delta| \leq 1$ with arbitrary phase, we may deduce from (20.14) that the "real" Nyquist plot at any given frequency ω_0 is contained in a region delimited by a circle centered at $P_0(j\omega_0)K(j\omega_0)$, with radius $|P_0KW(j\omega_0)|$. This is illustrated in Figure 20.5(a). Clearly, if the circle of uncertainty ever includes -1, there is the possibility that the "real" Nyquist plot has an extra encirclement, and hence is unstable. We may relate this to the robust stability problem as follows. From Example 20.2, the SISO system is robustly stable by the small gain theorem if

$$\left|\frac{P_0K}{1 + P_0K}W\right| < 1, \qquad \forall\, \omega. \tag{20.15}$$

Equivalently,

$$|P_0KW| < |1 + P_0K|. \tag{20.16}$$

The right-hand side of (20.16) is the magnitude of a translation of the Nyquist plot of the nominal loop transfer function. In Figure 20.5(b), because of the translation, encirclement of zero will destabilize the system. Clearly, this cannot happen if (20.16) is satisfied. This makes the relationship of robust stability to the SISO Nyquist criterion clear.

Performance as Stability Robustness

Suppose that, for some plant model P, we wish to design a feedback controller that not only stabilizes the plant (first order of priority!), but also provides some performance benefits, such as improved output regulation in the presence of disturbances. Given that something is known

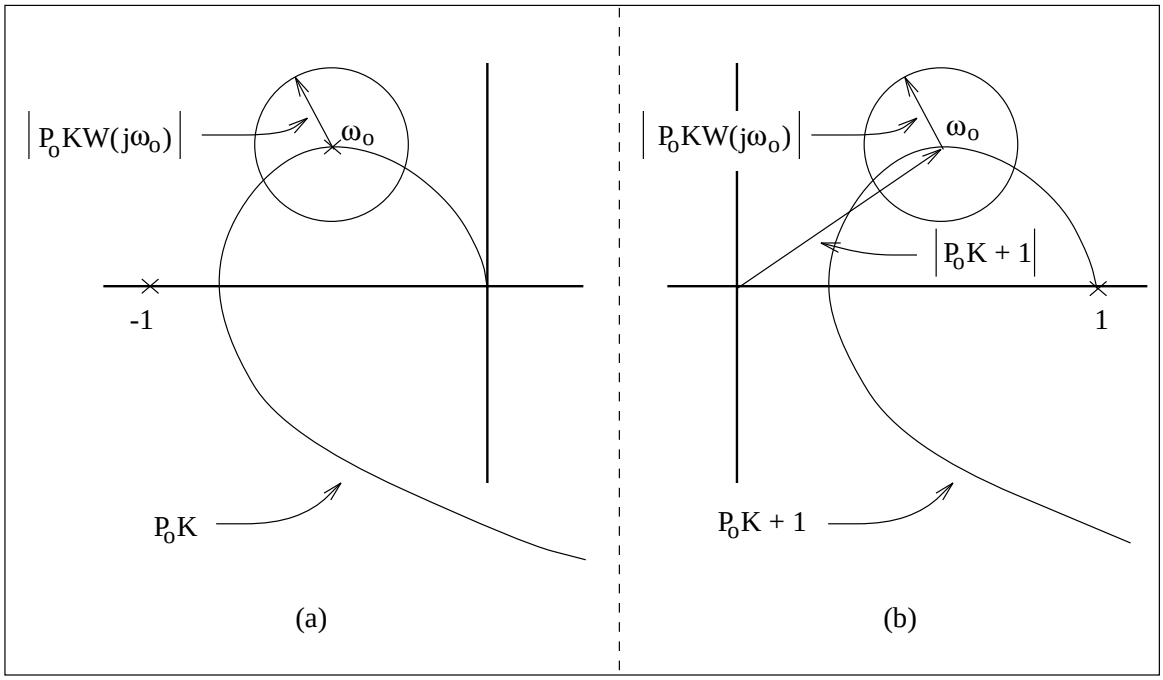


Figure 20.5: Relation of Nyquist criterion and robust stability.

about the frequency spectrum of such disturbances, the system model might look like Figure 20.6, where $\|\xi\|_2 < 1$, and the modeling filter W can be constructed to capture frequency characteristics of the disturbance. Calculating the transfer function of this loop from ξ to y, we have that $y = (I + PK)^{-1}W\xi$. We assume that the performance specification will be met if $\|(I + PK)^{-1}W\|_\infty < 1$, which does not restrict the problem, since W can always be scaled to reflect the actual magnitude of the disturbance or performance specification. This formulation looks analogous to a robust stability problem, and indeed, it can be verified that the small-gain theorem applied to the system of Figure 20.7 captures the identical constraint on the system transfer function. By mapping this system into the standard form of Figure 20.3, we find that $M = (I+PK)^{-1}W$, which is exactly the M that is needed if the small-gain condition is to yield the desired condition.

Finally, plant uncertainty has to be brought into the picture simultaneously with the

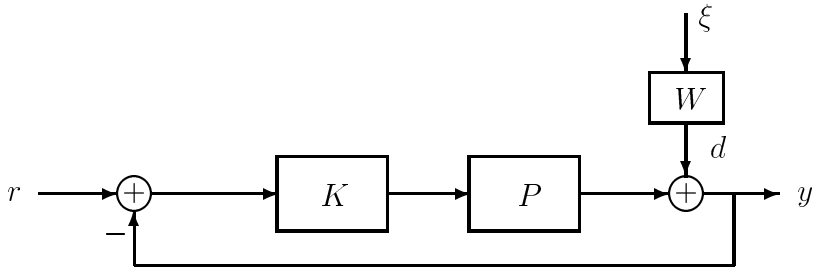


Figure 20.6: Plant with disturbance.

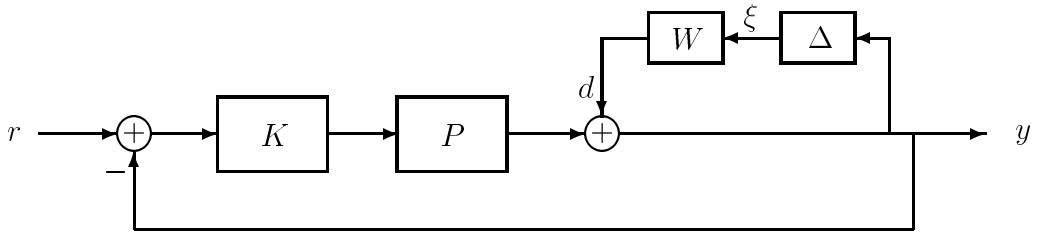


Figure 20.7: Mapping performance specifications into a stability problem.

performance constraints. This is necessary to formulate the *performance robustness* problem. It should be evident that this will lead to situations with block-diagonal Δ, as was obtained in the context of the last example in the previous subsection. The treatment of this case will require the notion of structured singular values, as we shall see in the next lecture.

Appendix

Necessity of the small gain condition for robust stability can be proved by showing that if $\sigma_{max}[M(j\omega_0)] > 1$ for some ω_0, we can construct a Δ of norm less than one, such that the resulting closed-loop map G_{zv} is unstable. This is done as follows. Take the singular value decomposition of $M(j\omega_0)$,

$$M(j\omega_0) = U\Sigma V' = U \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \end{bmatrix} V'. \tag{20.17}$$

Since $\sigma_{max}[M(j\omega_0)] > 1$, $\sigma_1 > 1$. Then $\Delta(j\omega_0)$ can be constructed as:

$$\Delta(j\omega_0) = V \begin{bmatrix} 1/\sigma_1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} U' \tag{20.18}$$

Clearly, $\sigma_{max}\Delta(j\omega_0) < 1$. We then have

$$
\begin{aligned}
(I - M\Delta)^{-1}(j\omega_0) = \quad & I - U \begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_n \end{bmatrix} V'V \begin{bmatrix} 1/\sigma_1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} U' \\
& = U \left[I - \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \right] U' \\
& = U \begin{bmatrix} 0 & & & \\ & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} U'
\end{aligned} \tag{20.19}
$$

which is singular. Only one problem remains, which is that $\Delta(s)$ must be legitimate as the transfer function of a *stable system*, evaluating to the proper value at $s = j\omega_0$, and having its maximum singular value over all ω bounded below 1. The value of the destabilizing perturbation at ω_0 is given by

$$\Delta_0(j\omega_0) = \frac{1}{\sigma_{max}(M(j\omega_0))} v_1 u_1'$$

Write the vectors v_1 and u_1' as

$$
v_1 = \begin{bmatrix} \pm|a_1|e^{j\theta_1} \\ \pm|a_2|e^{j\theta_2} \\ \vdots \\ \pm|a_n|e^{j\theta_n} \end{bmatrix}, \quad u_1' = \begin{bmatrix} \pm|b_1|e^{j\phi_1} & \pm|b_2|e^{j\phi_2} & \cdots & \pm|b_n|e^{j\phi_n} \end{bmatrix}, \tag{20.20}
$$

where θ_i and ϕ_i belong to the interval $[0, \pi)$. Note that we used $\pm$ in the representation of the vectors v_1 and u_1' so that we can restrict the angles θ_i and ϕ_i to the interval $[0, \pi)$. Now we can choose the nonnegative constants $\alpha_1,\ \alpha_2,\ \cdots,\ \alpha_n$ and $\beta_1,\ \beta_2,\ \cdots,\ \beta_n$ such that the phase of the function $\frac{s-\alpha_i}{s+\alpha_i}$ at $s = j\omega_0$ is θ_i, and the phase of the function $\frac{s-\beta_i}{s+\beta_i}$ at $s = j\omega_0$ is ϕ_i. Now the destabilizing $\Delta(s)$ is given by

$$\Delta(s) = \frac{1}{\sigma_{max}(M(j\omega_0))} g(s) h^T(s) \tag{20.21}$$

where

$$
g(s) = \begin{bmatrix} \pm|a_1|\frac{s-\alpha_1}{s+\alpha_1} \\ \pm|a_2|\frac{s-\alpha_2}{s+\alpha_2} \\ \vdots \\ \pm|a_n|\frac{s-\alpha_n}{s+\alpha_n} \end{bmatrix}, \quad h(s) = \begin{bmatrix} \pm|b_1|\frac{s-\beta_1}{s+\beta_1} \\ \pm|b_2|\frac{s-\beta_2}{s+\beta_2} \\ \vdots \\ \pm|b_n|\frac{s-\beta_n}{s+\beta_n} \end{bmatrix}. \tag{20.22}
$$

Exercises

Exercise 20.1 Consider a plant described by the transfer function matrix

$$P_\alpha(s) = \begin{pmatrix} \frac{\alpha}{s-1} & \frac{1}{s-1} \\ \frac{2s-1}{s(s-1)} & \frac{1}{s-1} \end{pmatrix}$$

where α is a real but uncertain parameter, confined to the range $[0.5\,,\,1.5]$. We wish to design a feedback compensator $K(s)$ for robust stability of a standard servo loop around the plant.

(a) We would like to find a value of α, say $\tilde{\alpha}$, and a scalar, stable, proper rational $W(s)$ such that the set of possible plants $P_\alpha(s)$ is contained within the "uncertainty set"

$$P_{\tilde{\alpha}}(s)[I + W(s)\Delta(s)]$$

where $\Delta(s)$ ranges over the set of stable, proper rational matrices with $\|\Delta\|_\infty \leq 1$. Try and find (no assurances that this is possible!) a suitable $\tilde{\alpha}$ and $W(s)$, perhaps by keeping in mind that what we really want to do is guarantee

$$\sigma_{max}\{P_{\tilde{\alpha}}^{-1}(j\omega)[P_\alpha(j\omega) - P_{\tilde{\alpha}}(j\omega)]\} \leq |W(j\omega)|$$

What specific choice of $\Delta(s)$ yields the plant $P_1(s)$ (i.e. the plant with $\alpha = 1$) ?

(b) Repeat part (a), but now working with the uncertainty set

$$P_{\tilde{\alpha}}(s)[I + W_1(s)\Delta(s)W_2(s)]$$

where $W_1(s)$ and $W_2(s)$ are column and row vectors respectively, and $\Delta(s)$ is scalar. Plot the upper bound on

$$\sigma_{max}\{P_{\tilde{\alpha}}^{-1}(j\omega)[P_\alpha(j\omega) - P_{\tilde{\alpha}}(j\omega)]\}$$

that you obtain in this case.

(c) For each of the cases above, write down a sufficient condition for robust stability of the closed-loop system, stated in terms of a norm condition involving the nominal complementary sensitivity function $T = (I + KP_{\tilde{\alpha}})^{-1}KP_{\tilde{\alpha}}$ and W — or, in part (b), W_1 and W_2.

Exercise 20.2 It turns out that the small gain theorem holds for nonlinear systems as well. Consider a feedback configuration with a stable system M in the forward loop and a stable, unknown perturbation in the feedback loop. Assume that the configuration is well-posed. Verify that the closed loop system is stable if $\|M\|\|\Delta\| < 1$. Here the norm is the gain of the system over *any* p-norm. (This result is also true for both DT and CT systems; the same proof holds).

Exercise 20.3 The design of a controller should take into consideration quantization effects. Let us assume that the only variable in the closed loop which is subject to quantization is the output of the plant. Two very simple schemes are proposed:

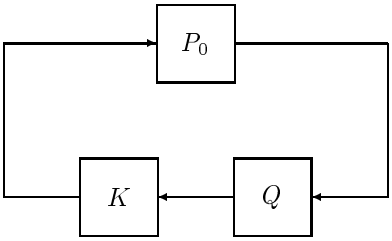


Figure 20.8: Quantization in the Closed Loop.

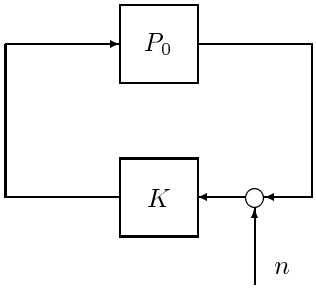


Figure 20.9: Quantization Modeled as Bounded Noise.

1. Assume that the output is passed through a quantization operator Q defined as:

$$Q(x) = a\lfloor \frac{|x|}{.5+a} \rfloor sgn(x), \quad a > 0$$

   where $\lfloor r \rfloor$ denotes the largest integer smaller than r. The output of this operator feeds into the controller as in Figure 20.8. Derive a sufficient condition that guarantees stability in the presence of Q.

2. Assume that the input of the controller is corrupted with an unknown but bounded signal, with a small bound as in Figure 20.9. Argue that the controller should be designed so that it does not amplify this disturbance at its input.

   Compare the two schemes, i.e., do they yield the same result? Is there a difference?

Chapter 21

Robust Performance and Introduction to the Structured Singular Value Function

---

21.1 Introduction

As discussed in Lecture 20, a process is better described in terms of a set of plants centered around a nominal model. The robust stabilization problem is concerned with finding non conservative conditions on the stable nominal closed loop system that guarantee the stability of all possible closed loop systems. An equally important problem is the robust performance problem which is concerned with finding non conservative conditions on the nominal closed loop system that guarnatee that the performance is met for all possible closed loop systems.

21.2 Robust Disturbance Rejection

We will focus our discussion on one prototype problem, namely, the robust disturbance rejection problem shown in Figure 21.1. This motivates the following problem:

Robust Disturbance Rejection Problem (RP)

Find conditions on the nominal closed-loop system (P_o, K) such that

1. K robustly stabilizes all $P \in \Omega$, where $\Omega = \{P \mid P = (I + \Delta_1 W_1)P_o, \quad \|\Delta\|_\infty < 1\}$.
2. $\|(I + PK)^{-1} W_2\|_\infty \leq 1$ for all $P \in \Omega$.

From Lecture 20, a performance objective in terms of the $\mathcal{H}_\infty$-norm of some closed loop map between some exogenous input w, to a regulated variable z, is mathematically equivalent to a robust stabilization problem with a perturbation block mapping the regulated output z to the exogenous input w. Obviously, the new perturbed system is stable if and only if $\|T_{zw}\|_\infty \leq 1$, which is the performance

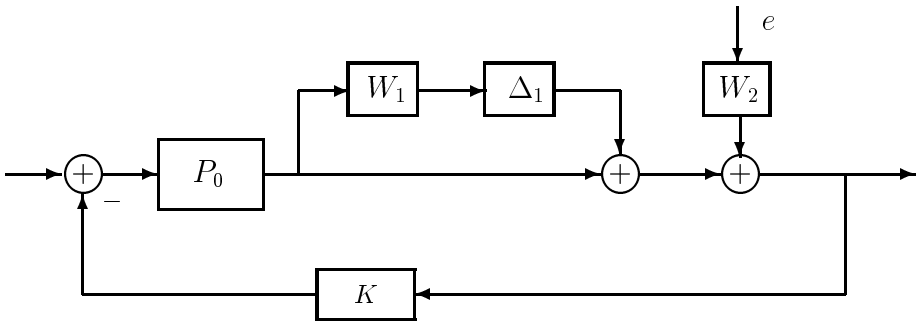


Figure 21.1: Uncertain Plant with Disturbance

objective. Notice that if the performance objective consists of several closed loop maps, then several perturbation blocks can be introduced in exactly the same fashion.

Proceeding for **RP**, we can "wrap" a frequency-weighted perturbation from the output to the input of interest, which results in the model of Figure 21.2. Next, we can re-arrange the system into the

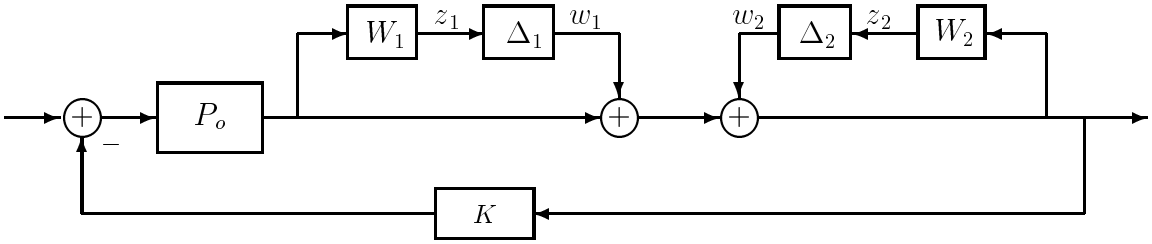


Figure 21.2: Robust Performance Model

M-Δ feedback form (a nominal stable M in feedback with the perturbation Δ) as in Figure 21.3. In this case, however, there are multiple inputs and outputs to consider. We use the following procedure to generate M and Δ:

1. Define $w_i,\ z_i$ to be the output and input, respectively, of the perturbation Δ_i.
2. For a total of m perturbations, compute the matrix transfer function M as the map from

$$w = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix} \quad \text{to} \quad z = \begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix}. \tag{21.1}$$

   In other words, all the Δ blocks are removed, and the transfer functions "seen" by the blocks from each input w_j to each output z_i are calculated and used as the $(i,j)^{\text{th}}$ element of M.

3. The perturbation matrix Δ will have the structure

$$\Delta = \begin{bmatrix} \Delta_1 & & \\ & \ddots & \\ & & \Delta_m \end{bmatrix}, \qquad \|\Delta_i\|_\infty < 1. \tag{21.2}$$

   For a SISO system, each $\Delta_i(j\omega)$ is a scalar, so that Δ becomes a diagonal matrix with complex entries. In the MIMO case, Δ is block-diagonal.

Example 21.1 (Robust Disturbance Rejection)

Applying the robust performance procedure to Figure 21.2 yields:

$$M = \begin{bmatrix} -W_1(I+P_0K)^{-1}P_0K & -W_1(I+P_0K)^{-1}P_0K \\ W_2(I+P_0K)^{-1} & W_2(I+P_0K)^{-1} \end{bmatrix}. \tag{21.3}$$

The transfer functions on the diagonal are identical to those in the single-block robust stability and disturbance-rejection problems, respectively, while the off-diagonal terms account for the interaction between the two constraints. Having found the appropriate M and Δ, we have thereby reduced the robust performance problem to a stability problem for the system of Figure 21.3.

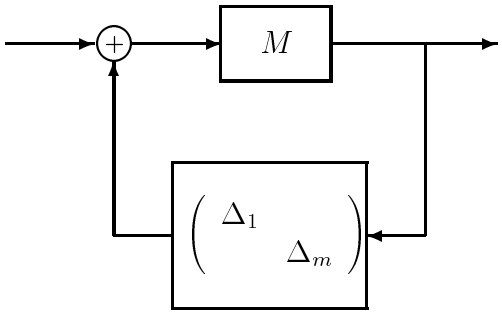


Figure 21.3: M-Δ Feedback Form

A sufficient condition for robust stability is given by the small gain theorem, namely,

$$\sigma_{\max}[M(jw)]\sigma_{\max}[\Delta(jw)] \leq \gamma < 1, \qquad \text{for all } w.$$

Since Δ is norm bounded by one, this condition translates to $\|M\|_\infty \leq \gamma$. This condition, however, is far from necessary since Δ has a block diagonal structure.

21.3 The Structured Singular Value

For an unstructured perturbation, the supremum of the maximum singular value of M (*i.e.* $\|M\|_\infty$) provides a clean and numerically tractable method for evaluating robust stability. Recall that, for the standard M-Δ loop, the system fails to be robustly stable if there exists an admissible Δ such that $(I - M\Delta)$ is singular. What distinguishes the current situation from the unstructured case is that we have placed constraints on the set $\mathbb{\Delta}$. Given this more limited set of admissible perturbations, we desire a measure of robust stability similar to $\|M\|_\infty$. This can be derived from the *structured singular value* $\mu(M)$.

Definition 21.1 The structured singular value of a complex matrix M with respect to a class of perturbations $\mathbb{\Delta}$ is given by

$$\mu(M) \triangleq \frac{1}{\inf\{\sigma_{max}(\Delta) \mid \det(I - M\Delta) = 0\}}, \qquad \Delta \in \mathbb{\Delta}. \tag{21.4}$$

If $det(I - M\Delta) \neq 0$ for all $\Delta \in \mathbb{\Delta}$, then $\mu(M) = 0$.

Theorem 21.1 The M-Δ System is stable for all $\Delta \in \mathbb{\Delta}$ with $\|\Delta\|_\infty < 1$ if and only if

$$\sup_\omega \mu(M(j\omega)) \leq 1.$$

Proof: Immediate, from the definition. Clearly, if $\mu \leq 1$, then the norm of the smallest allowable destabilizing perturbation Δ must by definition be greater than 1.

21.4 Properties of the Structured Singular Value

It is important to note that μ is a function that depends on the perturbation class $\mathbb{\Delta}$ (sometimes, this function is denoted by $\mu_{\mathbb{\Delta}}$ to indicate this dependence). The following are useful properties of such a fucntion.

1. $\mu(M) \geq 0$.
2. If $\mathbb{\Delta} = \{\lambda I \mid \lambda \in \mathbb{C}\}$, then $\mu(M) = \rho(M)$, the spectral radius of M (which is equal to the magnitude of the eigenvalue of M with maximum magnitude).
3. If $\mathbb{\Delta} = \{\Delta \mid \Delta$ is an arbitrary complex matrix$\}$ then $\mu = \sigma_{\max}(M)$, from which $\sup_\omega \mu = \|M\|_\infty$.

Property 2 shows that the spectral radius function is a particular μ function with respect to a perturbation class consisting of matrices of the form of scaled identity. Property 3 shows that the maximum singular value function is a particular μ function with respect to a perturbation class consisting of arbitrary norm bounded perturbations (no structural constraints).

4. If $\mathbb{\Delta} = \{\text{diag}(\Delta_1, \ldots, \Delta_n) \mid \Delta_i$ complex$\}$, then $\mu(M) = \mu(D^{-1}MD)$ for any $D = \text{diag}(d_1, \ldots, d_n),\ |d_i| > 0$. The set of such scales is denoted $\mathcal{D}$.

This can be seen by noting that $\det(I - AB) = \det(I - BA)$, so that $\det(I - D^{-1}MD\Delta) = \det(I - MD\Delta D^{-1}) = \det(I - M\Delta)$. The last equality arises since the diagonal matrices Δ and D commute.

5. If $\mathbb{\Delta} = \text{diag}(\Delta_1, \ldots, \Delta_n),\ \Delta_i$ complex, then $\rho(M) \leq \mu(M) \leq \sigma_{\max}(M)$.

This property follows from the following observation: If $\mathbb{\Delta}_1 \subset \mathbb{\Delta}_2$, then $\mu_1 \leq \mu_2$. It is clear that the class of perturbations consisting of scaled identity matrices is a subset of $\mathbb{\Delta}$ which is a subset of the class of all unstructured perturbations.

6. From 4 and 5 we have that $\mu(M) = \mu(D^{-1}MD) \leq \inf_{D\in\mathcal{D}} \sigma_{\max}(D^{-1}MD)$.

21.5 Computation of μ

In general, there is no closed-form method for computing μ. Upper and lower bounds may be computed and refined, however. In these notes we will only be concerned with computing the upper bound. If $\mathbb{\Delta} = \text{diag}(\Delta_1, \ldots, \Delta_n)$, then the upper bound on μ is something that is easy to calculate. Furthermore, property 6 above suggests that by infimizing $\sigma_{\max}(D^{-1}MD)$ over all possible diagonal scaling matrices, we obtain a better approximation of μ. This turns out to be a convex optimization problem at each

frequency, so that by infimizing over $\mathcal{D}$ at each frequency, the tightest upper bound over the set of $\mathcal{D}$ may be found for μ.

We may then ask when (if ever) this bound is tight. In other words, when is it truly a least upper bound. The answer is that for three or fewer Δ's, the bound is tight. The proof of this is involved, and is beyond the scope of this class. Unfortunately, for four or more perturbations, the bound is not tight, and there is no known method for computing μ exactly for more than three perturbations.

21.6 Robust Disturbance Rejection (SISO)

As shown earlier, the disturbance rejection requirement could be converted to a robust stability problem with two blocks of uncertainty, as in Figure 21.2, where Δ_1 and Δ_2 are SISO stable systems. Hence $\mathbb{\Delta}$ is the set of 2×2 diagonal complex matrices (which result from evaluating Δ at each frequency).

Now, since this is a two-block problem, it should be possible to find μ by infimizing $\sigma_{\max}(D^{-1}MD)$. We have $D = \mathrm{diag}(d_1, d_2)$, so that

$$\mu(M(j\omega)) = \inf_{d_1,d_2>0} \left\{ \sigma_{\max} \underbrace{\begin{bmatrix} -\frac{W_1 P_0 K}{1+P_0 K}(j\omega) & -\frac{d_2}{d_1}\frac{W_1 K}{1+P_0 K}(j\omega) \\ \frac{d_1}{d_2}\frac{W_2 P_0}{1+P_0 K}(j\omega) & \frac{W_2}{1+P_0 K}(j\omega) \end{bmatrix}}_{A(\alpha)} \right\}, \tag{21.5}$$

with the "pure" robust stability requirement occupying the upper left diagonal, and the nominal performance requirement on the lower right. Setting $\alpha = d_2/d_1$ and fixing ω, and taking the definition of $A(\alpha)$ from (21.5), we have

$$\mu(M(j\omega)) = \inf_{|\alpha|>0} \{\lambda_{\max}^{1/2}(A^*(\alpha)A(\alpha))\}. \tag{21.6}$$

Now, for nominal performance, we require that

$$\left| \frac{W_2}{1+P_0 K}(j\omega) \right| \leq 1. \tag{21.7}$$

For robust stability, we nee

$$\left| \frac{W_1 P_0 K}{1+P_0 K}(j\omega) \right| \leq 1. \tag{21.8}$$

For robust performance, the necessary and sufficient condition is

$$\mu(M(j\omega)) \leq 1. \tag{21.9}$$

A bit of algebra yields

$$\lambda_{\max}(A^*A) = |\alpha|^2 \left| \frac{W_1 K}{1+P_0 K}(j\omega) \right|^2 + \left| \frac{W_2}{1+P_0 K}(j\omega) \right|^2 \tag{21.10}$$

$$+ \left| \frac{W_1 K P_0}{1+P_0 K}(j\omega) \right|^2 + \frac{1}{|\alpha|^2} \left| \frac{W_2 P_0}{1+P_0 K}(j\omega) \right|^2 \tag{21.11}$$

from which we have

$$\inf_{\alpha} \lambda_{\max}(A^*A) = \left(\left| \frac{W_1 P_0 K}{1+P_0 K}(j\omega) \right| + \left| \frac{W_2}{1+P_0 K}(j\omega) \right| \right)^2. \tag{21.12}$$

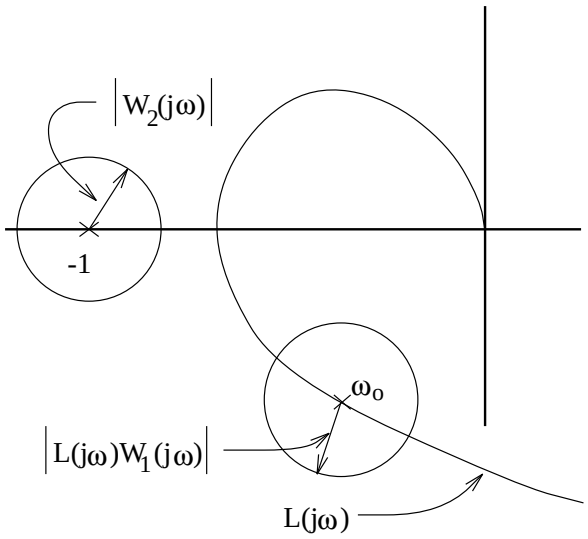


Figure 21.4: Robust Performance/Nyquist Criterion

This minimum occurs at

$$|\alpha|^2 = \frac{|W_2 P_0|}{|W_1 K|} \tag{21.13}$$

which is not equal to 1 in general, so that $\sup_\omega \mu \leq \|M\|_\infty$. In other words, μ is a less conservative measure than $\|\cdot\|_\infty$ in this case.

Once again, there is a graphical interpretation of the SISO robust disturbance rejection problem, in terms of the Nyquist criterion. From (21.12), we have

$$\mu(M(j\omega)) \leq 1 \Longleftrightarrow \left|\frac{W_1 P_0 K}{1+P_0 K}(j\omega)\right| + \left|\frac{W_2}{1+P_0 K}(j\omega)\right| \leq 1. \tag{21.14}$$

Letting $L(j\omega)$ represent the nominal loop gain $P_0 K(j\omega)$, this can be rewritten as:

$$|W_1 L(j\omega)| + |W_2| \leq |1 + L(j\omega)|. \tag{21.15}$$

Graphically, we can represent this at each frequency ω as a circle centered at -1 of radius $|W_2|$, and a second circle centered at $L(j\omega)$ of radius $|W_1 L(j\omega)|$. Robust performance will be achieved as long as the two circles never intersect.

Loop-shaping Revisited

Loop-shaping is a well-established method of control design that concentrates on the frequency-domain characteristics of the open-loop transfer function $L = P_0 K$. Based primarily on design experience, there are certain characteristics of the loop transfer function that translate into desirable control performance. Other open-loop characteristics are known by experience to result in undesirable or unpredictable behavior. This method differs from μ-synthesis and $\mathcal{H}_\infty$ methods, which concentrate on optimizing the characteristics of the closed-loop transfer function. Since, presumably, a controller with good behavior designed by loop-shaping should be similar in some way to a controller designed by more recent methods, it is of interest to look for parallels in the heuristic rules of loop-shaping and the more methodical methods of μ-synthesis and $\mathcal{H}_\infty$.

Identifying the sensitivity and complementary sensitivity functions from (21.14), we can write the RP requirement as

$$|W_1(j\omega)T(j\omega)| + |W_2(j\omega)S(j\omega)| \leq 1. \tag{21.16}$$

Model uncertainty typically increases with frequency, so it is important that the complementary sensitivity function decreases with increasing frequency. For disturbance rejection, which is typically most critical over a low frequency range, we require that $S(j\omega)$ remain small. The weighting functions W_1 and W_2 are designed to reflect this, and so might take on the form of Figure 21.5. Normally, at low

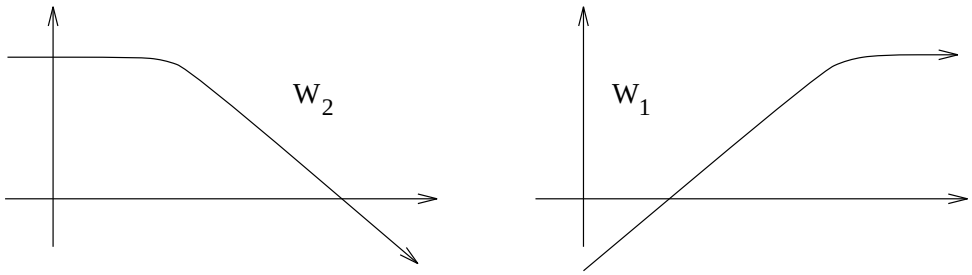


Figure 21.5: Typical Weighting Functions

frequency, $L(j\omega) >> 1$ and at high frequency, $L(j\omega) << 1$. Now,

$$T_0 = \frac{L}{1+L}, \quad S_0 = \frac{1}{1+L} \tag{21.17}$$

so that at low frequency, $T_0 \approx 1$ and $S_0 \approx 1/L$. Thus we can approximate the RP requirement at the low end as:

$$|W_1| + \left|W_2 \frac{1}{L}\right| \leq 1 \quad \Longrightarrow \quad |L| \geq \frac{|W_2|}{1-|W_1|} \tag{21.18}$$

At high frequency, the approximation is $T_0 \approx L$ and $S_0 \approx 1$, which leads to:

$$|W_1 L| + |W_2| \leq 1, \quad \Longrightarrow \quad |L| \leq \frac{1-|W_2|}{|W_1|}. \tag{21.19}$$

These constraints are summarized in Figure 21.6, which also notes another design rule, which is that the 0 dB crossing should occur at a slope no more negative than -40 dB per decade. If W_1 and W_2 do not overlap significantly in frequency, then the upper and lower bounds reduce to $|W_2|$ and $1/|W_1|$, respectively.

Example 21.2 (Loop Shaping)

Assume P_0 is minimum phase stable with relative degree 1. Designing a controller by shaping the loop gain $L = P_0 K$ is not affected by P_0; just the relative degree is needed.

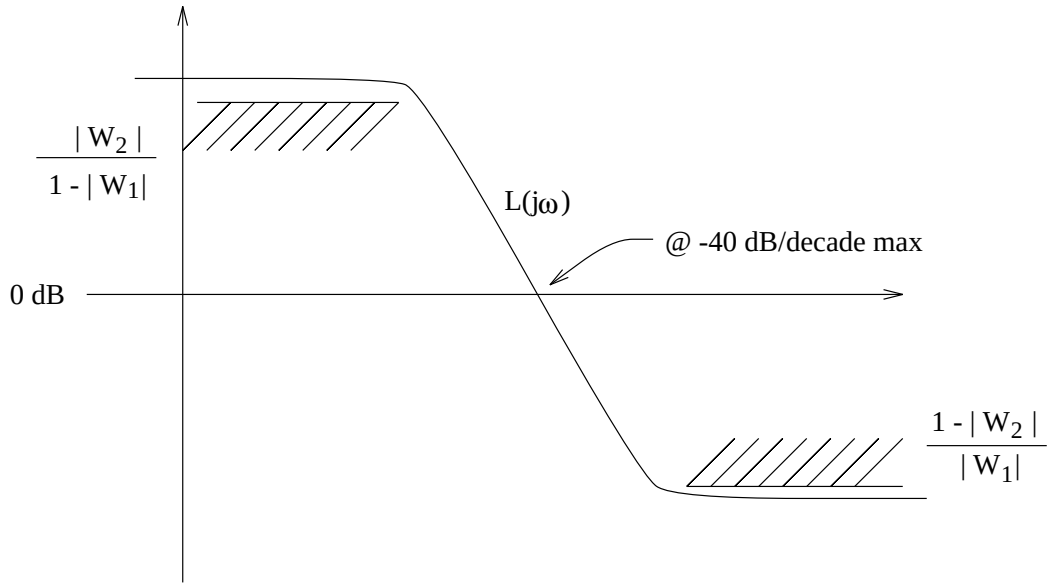


Figure 21.6: Typical Loop-shaping Problem

Suppose the multiplicative uncertainty is described by

$$W_1 = \frac{s+1}{20(0.01s+1)},$$

i.e., the multiplicative perturbations of the plant are upper bounded by $W_1(j\omega)$ at each frequency.

The objective is to track sinusoidal signals at the reference input in the frequency range $[0, 1]$ rad/s. We would like to make the tracking error small; however, we do not know yet by how much. Let $W_2(j\omega)$ have the following frequency response

$$|W_2(j\omega)| = \begin{cases} a & 0 \le \omega \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that this may not correspond to a stable $W_2(s)$; however, this does not affect the resulting loop shape. We are going to exhibit the design by trial and error. Let

$$L(s) = \frac{b}{cs+1}.$$

At high frequency, $\omega \ge 20$,

$$L \le \frac{1-|W_2|}{|W_1|} = \frac{1}{|W_1|} \qquad \omega \ge 20.$$

If we pick $c = 1$, then the largest value for b such that the above is satisfied is $b = 20$. Hence

$$L(s) = \frac{20}{s+1}.$$

At low frequency, $\omega \le 1$,

$$|L| \ge \frac{|W_2|}{1-|W_1|} = \frac{a}{1-|W_1|}.$$

Since $|L(j\omega)|$ is decreasing and $|W_1(j\omega)|$ is increasing in the range $[0,1]$, the largest a can be solved for:

$$|L(j1)| = \frac{a}{1-|W_1(j1)|},$$

which implies that $a = 13.15$. Checking the RP condition

$$|W_2 S(j\omega)| + |W_1 T(j\omega)| \leq 0.92 \quad \forall\omega$$

which implies RP is achieved and the tracking error is smaller than $1/13.15$ in the range $[0,1]$. If a better performance is desired, a possibly more complicated L needs to be used.

The discussion in this chapter has focused on perturbations that are arbitrary dynamic systems. This alowed us to think of any class of structured perurbations as sets of arbitrary (structured) matrices at each frequency point. These matrices correspond to evaluating the dynamic system at a given frequency.

In practical applications, some perturbations may be static and not dynamic. These arise in problems with real parameter uncertainties. We can still proceed as before and transform such problems to the general M-Δ diagram. In this case, Δ will have a combination of both static and dynamic perturbations. μ for such a class can be defined as before, and it will provide a necessary and sufficient condition for robust stability.

The main issue here is computing a good upper bound for μ. Of course, we can always embed this class of perturbations in a larger class containing dynamic perturbations and use D-scaling to obtain an upper bound. This, however, gives conservative conditions. Computing non-conservative upper bounds of μ for such perturbations remains an active area of research.

21.7 Rank-One μ

Although we do not have methods for computing μ exactly, there is one particular situation where this is possible. This situation occurs if M has rank 1, *i.e.*

$$M = ab^*$$

where $a, b \in \mathbb{C}^n$. Then it follows that μ with respect to $\mathbf{\Delta}$ containing complex diagonal perturbations is given by

$$\frac{1}{\mu(M)} = \inf_{\Delta\in\mathbf{\Delta}} \{\sigma_{\max}(\Delta) \mid \det(I - M\Delta) = 0\}.$$

However,

$$\begin{aligned}
\det(I - M\Delta) &= \det(I - ab^*\Delta) \\
&= \det(I - b^*\Delta a) \\
&= \det\left(I - [\Delta_1 \cdots \Delta_n] \begin{bmatrix} \bar{b_1} a_1 \\ \bar{b_2} a_2 \\ \vdots \\ \bar{b_n} a_n \end{bmatrix} \right) \\
&= 1 - [\Delta_1 \cdots \Delta_n] \begin{bmatrix} \bar{b_1} a_1 \\ \bar{b_2} a_2 \\ \vdots \\ \bar{b_n} a_n \end{bmatrix},
\end{aligned}$$

and $\sigma_{\max}(\Delta) = \max_i |\Delta_i|$. Hence,

$$\frac{1}{\mu(M)} = \inf_{\Delta_1,\ldots,\Delta_n} \left\{ \max_i |\Delta_i| \;\middle|\; [\Delta_1 \cdots \Delta_n] \begin{bmatrix} \bar{b_1} a_1 \\ \bar{b_2} a_2 \\ \vdots \\ \bar{b_n} a_n \end{bmatrix} = 1 \right\}.$$

Optimizing the RHS, it follows that (verify)

$$\frac{1}{\mu(M)} = \frac{1}{\sum_{i=1}^n |\bar{b_i} a_i|} \leftrightarrow \mu(M) = \sum_{i=1}^n |\bar{b_i} a_i|.$$

Notice that the SISO robust disturbance rejection problem is a rank-one problem. This follows since

$$M = \begin{bmatrix} -W_1 K \\ W_2 \end{bmatrix} [\frac{P_0}{1+P_0K} \quad \frac{1}{1+P_0K}].$$

Then

$$\mu(M(j\omega)) = \left| \frac{W_1 P_0 K}{1+P_0K}(j\omega) \right| + \left| \frac{W_2}{1+P_0K}(j\omega) \right|$$

which is the condition we derived before.

Coprime Factor Perturbations

Consider the class of SISO systems

$$\Omega = \left\{ \frac{N(s)}{D(s)} \;\middle|\; N = N_0 + \Delta_1 W_1, D = D_0 + \Delta_2 W_2,\ \|\Delta_i\| < 1 \right\}$$

where the nominal plant is N_0/D_0 with the property that both N_0 and D_0 are stable with no common zeros in the RHP. Assume that K stabilizes N_0/D_0. This block diagram is shown in Figure 21.7.

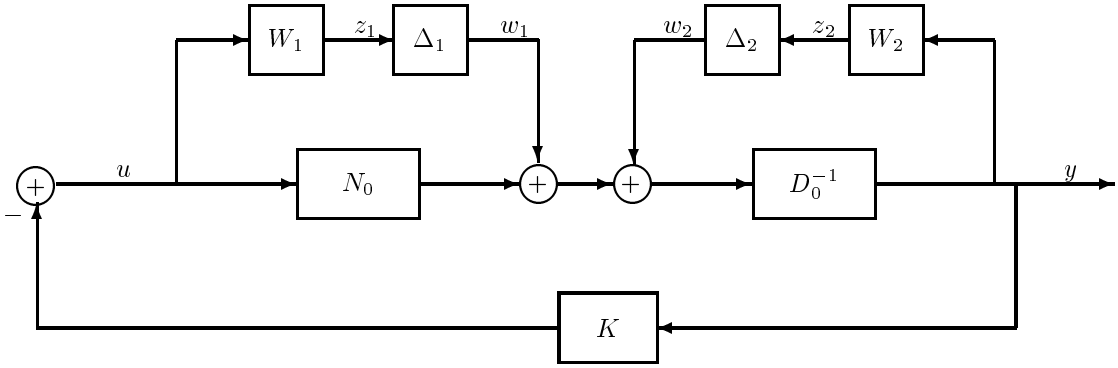


Figure 21.7: Coprime Factor Perturbation Model

The closed loop block diagram can be mapped to the M-Δ diagram where

$$\begin{aligned} M &= \begin{bmatrix} -\frac{W_1 K}{D_0+N_0 K} & -\frac{W_1 K}{D_0+N_0 K} \\ \frac{W_2}{D_0+N_0 K} & \frac{W_2}{D_0+N_0 K} \end{bmatrix} \\ &= \begin{bmatrix} -\frac{W_1 K}{D_0+N_0 K} \\ \frac{W_2}{D_0+N_0 K} \end{bmatrix} [1 \quad 1]. \end{aligned}$$

Hence, M has rank 1 and

$$\mu(M(j\omega)) = \left|\frac{W_1 K}{D_0+N_0 K}\right| + \left|\frac{W_2}{D_0+N_0 K}\right|.$$

Robust Hurwitz Stability of Polynomials with Complex Perturbations

Another application of the structured singular value with rank one matrices is the robust stability of a family of polynomials with complex perturbations of the coefficients. In this case let $\delta = \begin{bmatrix} \delta_{n-1} & \delta_{n-2} & \dots & \delta_0 \end{bmatrix}^T$ and consider the polynomial family

$$P(s,\delta) = s^n + (a_{n-1} + \gamma_{n-1}\delta_{n-1})s^{n-1} + \ldots + (a_0 + \gamma_0\delta_0),$$

where a_i, γ_i, and $\delta_i \in \mathbb{C}$ and $|\delta_i| \leq 1$. We want to obtain a condition that is both necessary and sufficient for the Hurwitz stability of the entire family of polynomials $P(s,\delta)$. We can write the polynomials in this family as

$$\begin{aligned} P(s,\delta) &= P(s,0) + \tilde{P}(s,\delta) && (21.20) \\ &= \left(s^n + a_{n-1}s^{n-1} + \ldots + a_0\right) + \left(\gamma_{n-1}\delta_{n-1}s^{n-1} + \ldots + \gamma_0\delta_0\right), && (21.21) \end{aligned}$$

which can also be rewritten as

$$P(s,\delta) = P(s,0) + \begin{bmatrix} 1 & 1 & \dots 1 \end{bmatrix} \begin{bmatrix} \delta_{n-1} & 0 & 0 & \dots & 0 \\ 0 & \delta_{n-2} & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ & & & \delta_1 & 0 \\ 0 & 0 & \dots & 0 & \delta_0 \end{bmatrix} \begin{bmatrix} \gamma_{n-1}s^{n-1} \\ \gamma_{n-2}s^{n-2} \\ \vdots \\ \gamma_1 s \\ \gamma_0 \end{bmatrix}.$$

We assume that the center polynomial $P(s,0)$ is Hurwitz stable. This implies that the stability of the entire family $P(s,\delta)$ is equivalent to the condition that

$$1 + \frac{1}{P(j\omega,0)} \begin{bmatrix} 1 & 1 & \dots 1 \end{bmatrix} \begin{bmatrix} \delta_{n-1} & 0 & 0 & \dots & 0 \\ 0 & \delta_{n-2} & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ & & & \delta_1 & 0 \\ 0 & 0 & \dots & 0 & \delta_0 \end{bmatrix} \begin{bmatrix} \gamma_{n-1}(j\omega)^{n-1} \\ \gamma_{n-2}(j\omega)^{n-2} \\ \vdots \\ \gamma_1(j\omega) \\ \gamma_0 \end{bmatrix} \neq 0$$

for all $\omega \in \mathbb{R}$ and $|\delta_i| \leq 1$. This is equivalent to the condition that

$$\det \left(I + \frac{1}{P(j\omega, 0)} \begin{bmatrix} \gamma_{n-1}(j\omega)^{n-1} \\ \gamma_{n-2}(j\omega)^{n-2} \\ \vdots \\ \gamma_1(j\omega) \\ \gamma_0 \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots 1 \end{bmatrix} \Delta \right) \neq 0$$

for all $\omega \in \mathbb{R}$ and $\Delta \in \mathbf{\Delta}$ with $\|\Delta\|_\infty \leq 1$. Now using the concept of the structured singular value we arrive at the following condition which is both necessary and sufficient for the Hurwitz stability of the entire family

$$\mu(M(j\omega)) < 1$$

for all $\omega \in \mathbb{R}$, where

$$M(j\omega) = \frac{1}{P(j\omega, 0)} \begin{bmatrix} \gamma_{n-1}(j\omega)^{n-1} \\ \gamma_{n-2}(j\omega)^{n-2} \\ \vdots \\ \gamma_1(j\omega) \\ \gamma_0 \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots 1 \end{bmatrix}.$$

Clearly this is a rank one matrix and by our previous discussion the structured singular value can be computed analytically resulting in the following test

$$\frac{1}{|P(j\omega, 0)|} \sum_{i=1}^{n} |\gamma_{n-i}||\omega|^{n-i} < 1$$

for all $\omega \in \mathbb{R}$.

Exercises

Exercise 21.1 In decentralized control, the plant is assumed to be diagonal and controllers are designed independently for each diagonal element. If however, the real process is not completely decoupled, the interactions between these separate subsystems can drive the system to instability.

Consider the 2×2 plant

$$P(s) = \begin{pmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{pmatrix}.$$

Assume that P_{12} and P_{21} are stable and relatively small in comparison to the diagonal elements, and only a bound on their frequency response is available. Suppose a controller $K = diag(K_1, K_2)$ is designed to stabilize the system $P_0 = diag(P_{11}, P_{22})$.

1. Set-up the problem as a stability robustness problem, i.e., put the problem in the $M - \Delta$ form.
2. Derive a non-conservative condition (necessary and sufficient) that guarantees the stability robustness of the above system. Assume the off-diagonal elements are perturbed independently. Reduce the result to the simplest form (an answer like $\mu(M) < 1$ is not acceptable; this problem has an exact solution which is computable).
3. How does your answer change if the off-diagonal elements are perturbed simultaneously with the same Δ.

Exercise 21.2 Consider the rank 1 μ problem. Suppose $\mathbb{\Delta}$, contains only real perturbations. Compute the exact expression of $\mu(M)$.

Exercise 21.3 Consider the set of plants characterized by the following sets of numerators and denominators of the transfer function:

$$N(s) = N_0(s) + N_\delta(s)\delta, \qquad D(s) = D_0(s) + D_\delta(s)\delta$$

Where both N_0 and D_0 are polynomials in s, $\delta \in \mathbb{R}^n$, and N_δ, D_δ are polynomial row vectors. The set of all plants is then given by:

$$\Omega = \{\frac{N(s)}{D(s)} \,|\, \delta \in \mathbb{R}^n,\ |\delta_i| \leq \gamma\}$$

Let K be a controller that stabilizes $\frac{N_0}{D_0}$. Compute the exact stability margin; i.e., compute the largest γ such that the system is stable.

Chapter 22

Reachability of DT LTI Systems

22.1 Introduction

We now begin a series of lectures to address the question of synthesizing feedback controllers. This objective requires a detailed understanding of how inputs impact the states of a given system, a notion we term *reachability*. Also, this objective requires a detailed understanding of the information the output provides about the rest of the states of the dynamic system, a notion we term observability. These notions together define the minimal set of conditions under which a stabilizing feedback controller exists.

22.2 The Reachability Problem

In previous lectures we have examined solutions of state-space models, the stability of undriven models, some properties of interconnections, and input-output stability. We now turn to a more detailed examination of how inputs affect states, for the n^{th}-order DT system

$$x(i+1) = Ax(i) + Bu(i) \; . \tag{22.1}$$

(The discussion of reachability in the DT case is generally simpler than in the CT case that we will consider next Chapter, but some structural subtleties that are hidden in the CT case become more apparent in the DT case. For the most part, however, DT results parallel CT results quite closely.) Recall that

$$\begin{aligned} x(k) &= A^k x(0) + \sum_{i=0}^{k-1} A^{k-i-1} B u(i) \\ &= A^k x(0) + \left[A^{k-1}B \mid A^{k-2}B \mid \cdots \mid B \right] \begin{pmatrix} u(0) \\ u(1) \\ \vdots \\ u(k-1) \end{pmatrix} \\ &= A^k x(0) + R_k \, \mathcal{U}_k \end{aligned} \tag{22.2}$$

where the definition of R_k and $\mathcal{U}_k$ should be clear from the equation that precedes them. Now consider whether and how we may choose the input sequence $u(i)$, $i \in [0,\ k-1]$, so as to move the system from $x(0) = 0$ to a desired target state $x(k) = d$ at a given time k. If there is such an input, we say that the state d is **reachable** in k steps. It is evident from (22.2) that — assuming there are no constraints placed on the input — the set $\mathbb{R}_k$ of states reachable from the origin in k steps, or the *k-reachable set*, is precisely the range of R_k, i.e.

$$\mathbb{R}_k = Ra(R_k) \tag{22.3}$$

The k-reachable set is therefore a *subspace*, and may be referred to as the k-reachable subspace. We call the matrix R_k the *k-step reachability matrix*.

Theorem 22.1

For $k \leq n \leq \ell$,

$$Ra(R_k) \subseteq Ra(R_n) = Ra(R_\ell) \tag{22.4}$$

so the set of states reachable from the origin in some (finite) number of steps by appropriate choice of control is precisely the subspace of states reachable in n steps.

Proof.
The fact that $Ra(R_k) \subseteq Ra(R_n)$ for $k \leq n$ follows trivially from the fact that the columns of R_k are included among those of R_n. To show that $Ra(R_n) = Ra(R_\ell)$ for $\ell \geq n$, note from the Cayley-Hamilton theorem that A^i for $i \geq n$ can be written as a linear combination of $A^{n-1}, \cdots, A, I$, so all the columns of R_ℓ for $\ell \geq n$ are linear combinations of the columns of R_n. Thus (22.4) is proved, and the rest of the statement of the theorem follows directly.

In view of Theorem 22.1, the subspace of states reachable in n steps, *i.e.* $Ra(R_n)$, is referred to as *the* reachable subspace, and will be denoted simply by $\mathbb{R}$; any reachable target state, *i.e.* any state in $\mathbb{R}$, is reachable in n steps (or less). The system is termed a *reachable system* if all of $\mathbb{R}^n$ is reachable, *i.e.* if $\text{rank}(R_n) = n$. The matrix

$$R_n = \left[A^{n-1}B \mid A^{n-2}B \mid \cdots \mid B \right], \tag{22.5}$$

is termed the *reachability matrix* (often written with its block entries ordered oppositely to the order that we have used here, but this is not significant).

Example 22.1 Consider the single-input system

$$\begin{bmatrix} x_1(k+1) \\ x_2(k+1) \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1(k) \\ x_2(k) \end{bmatrix} + \begin{bmatrix} 1 \\ 1 \end{bmatrix} u(k).$$

The reachable subspace is evidently (from symmetry) the line $x_1 = x_2$. This system is not reachable.

The following alternative characterization of $\mathbb{R}_k$ is useful, particularly because its CT version will play an important role in our development of the CT reachability story. Let us first define the *k-step reachability Gramian* $\mathcal{P}_k$ by

$$\mathcal{P}_k = R_k R_k^T = \sum_{i=0}^{k-1} A^i B B^T (A^T)^i \tag{22.6}$$

This matrix is therefore symmetric and positive semi-definite. We then have the following result.

Lemma 22.1

$$Ra(\mathcal{P}_k) = Ra(R_k) = \mathbb{R}_k \ . \tag{22.7}$$

Proof.
It is easy to see that $Ra(\mathcal{P}_k) \subset Ra(R_k)$. For the reverse inclusion, we can equivalently show that

$$Ra^{\perp}(\mathcal{P}_k) \subset Ra^{\perp}(R_k)$$

For this, note that

$$\begin{aligned} q^T\mathcal{P}_k = 0 \quad &\Longrightarrow \quad q^T\mathcal{P}_k q = 0 \\ &\Longleftrightarrow \quad \langle R_k^T q, R_k^T q\rangle = 0 \\ &\Longleftrightarrow \quad q^T R_k = 0 \end{aligned}$$

so any vector in $Ra^{\perp}(\mathcal{P}_k)$ is also in $Ra^{\perp}(R_k)$.
Thus the reachable subspace can equivalently be computed as $Ra(P_\ell)$ for any $\ell \geq n$. If the system is *stable*, then $P_\infty := P$ is well defined, and is easily shown to satisfy the Lyapunov equation

$$APA^T - P = -BB^T \tag{22.8}$$

We leave you to show that (22.8) has a (unique) positive definite (and hence full rank) solution P if and only if the system (A, B) is reachable.

Reachability from an Arbitrary Initial State

Note from (22.2) that getting from a nonzero starting state $x(0) = s$ to a target state $x(k) = d$ requires us to find a $\mathcal{U}_k$ for which

$$d - A^k s = R_k \mathcal{U}_k \tag{22.9}$$

For arbitrary d, s, the requisite condition is the same as that for reachability from the origin. Thus we can get from an arbitrary initial state to an arbitrary final state if and only if the system is reachable (from the origin); and we can make the transition in n steps or less, when the transition is possible.

Controllability versus Reachability

Now consider what is called the **controllability** problem, namely that of bringing an arbitrary initial state $x(0)$ to the origin in a finite number of steps. From (22.2) we see that this requires solving

$$-A^k x(0) = R_k \mathcal{U}_k \tag{22.10}$$

If A is invertible and $x(0)$ is arbitrary, then the left side of (22.10) is arbitrary, so the condition for controllability of $x(0)$ to the origin in a finite number of steps is precisely that $\text{rank}(R_k) = n$ for some k, *i.e.* just the reachability condition that $\text{rank}(R_n) = n$.

If, on the other hand, A is singular (*i.e.* has eigenvalues at 0), then the left side of (22.10) will be confined to a subspace of the state space, even when $x(0)$ is unrestricted. The range of A^k for a singular A may decrease initially, but $Ra(A^k) = Ra(A^n)$ for $k \geq n$ (since by stage n the Jordan blocks associated with the zero eigenvalues of A are all guaranteed to have been "zeroed out" in A^n). Meanwhile, as we have seen, the range of R_k may increase initially, but $Ra(R_k) = Ra(R_n)$ for $k \geq n$.

It follows from these facts and (22.10) that an arbitrary initial state is controllable to 0 in finite time, *i.e.* the system is controllable, iff

$$Ra(A^n) \subset Ra(R_n) \tag{22.11}$$

For invertible A, we recover our earlier condition. (The distinction between reachability and controllability is not seen in the CT case, because the state transition matrix there is e^{At} rather than A^k, and is always invertible.)

22.3 Modal Aspects

The following result begins to make the connection of reachability with modal structure.

Corollary 22.1

The reachable subspace $\mathbb{R}$ is A-invariant, *i.e.* $x \in \mathbb{R} \Longrightarrow Ax \in \mathbb{R}$. We write this as $A\mathbb{R} \subset \mathbb{R}$
Proof.
We first show

$$Ra(AR_n) \subset Ra(R_n) \tag{22.12}$$

For this, note that

$$AR_n = \left[\begin{array}{c|c|c|c} A^nB & A^{n-1}B & \cdots & AB \end{array}\right]$$

The last $n-1$ blocks are present in R_n, while the Cayley-Hamilton theorem allows us to write A^nB as a linear combination of blocks in R_n. This establishes (22.12). It follows that $x = R_n\alpha \Longrightarrow Ax = AR_n\alpha = R_n\beta \in \mathbb{R}$.

Some feel for how this result connects to modal structure may be obtained by considering what happens if the subspace $\mathbb{R}$ is one-dimensional. If v ($\neq 0$) is a basis vector for $\mathbb{R}$, then Corollary 22.1 states that

$$Av = \lambda v \tag{22.13}$$

for some λ, *i.e.* $\mathbb{R}$ is the space spanned by an *eigenvector* of A. More generally, it is true that any A-invariant subspace is the span of some eigenvectors and generalized eigenvectors of A. (It turns out that $\mathbb{R}$ is the smallest A-invariant subspace that contains $Ra(B)$, but we shall not pursue this fact.)

Standard Form for Unreachable Systems

If a system of the form (22.1) is unreachable, it is convenient to choose coordinates that highlight this fact. Specifically, we shall show how to change coordinates (using a similarity transformation) from $x = Tz$ to

$$z = T^{-1}x = \begin{array}{c} z_1 \\ z_2 \end{array}$$

where z_1 is an r-vector and z_2 is an $(n-r)$-vector, with r denoting the dimension of the reachable subspace, $r = \dim \mathbb{R}$. In these new coordinates, the system (22.1) will take the form

$$\begin{array}{c} z_1(k+1) \\ z_2(k+1) \end{array} = \left[\begin{array}{c|c} A_1 & A_{12} \\ \hline 0 & A_2 \end{array}\right] \begin{array}{c} z_1(k) \\ z_2(k) \end{array} + \left[\begin{array}{c} B_1 \\ \hline 0 \end{array}\right] u(k) \tag{22.14}$$

with the reachable subspace being the subspace with $z_2 = 0$. We shall refer to a system in the form (22.14) as being in the *standard form* for an unreachable system.

The matrix T is constructed as follows. Let $T_1^{n\times r}$ be a matrix whose columns form a basis for the reachable subspace, *i.e.*

$$\mathcal{R}a(T_1) = \mathcal{R}a(R_n)\ ,$$

and let $T_2^{n\times(n-r)}$ be a matrix whose columns are independent of each other and of those in T_1. Then choose

$$T = [\,T_1\,|\,T_2\,]\ .$$

This matrix is invertible, since its columns are independent by construction. We now claim that

$$\begin{aligned} A\,[\,T_1\,|\,T_2\,] &= T\bar{A} = [\,T_1\,|\,T_2\,] \begin{matrix} A_1^{r\times r} & A_{12} \\ 0 & A_2 \end{matrix} \\ B &= T\bar{B} = [\,T_1\,|\,T_2\,] \begin{bmatrix} B_1^{r\times m} \\ --- \\ 0 \end{bmatrix}. \end{aligned} \tag{22.15}$$

Our reasoning is as follows. Since the reachable subspace is A-invariant, the columns of AT_1 must remain in $\mathcal{R}a(T_1)$, which forces the 0 block in the indicated position in $\bar{A}$. Similarly, the presence of the zero block in $\bar{B}$ is a consequence of the fact that the columns of B are in the reachable subspace.

The above standard form is not uniquely defined, but it can be shown (we leave you to show it!) that any two such standard forms are related by a block upper triangular similarity transformation. As a result, A_1 and A_2 are *unique up to similarity transformations* (so, in particular, their Jordan forms are uniquely determined).

From (22.14) it is evident that if $z_2(0) = 0$ then the motion of $z_1(k)$ is described by the r^{th}-order *reachable* state-space model

$$z_1(k+1) = A_1 z_1(k) + B_1 u(k). \tag{22.16}$$

This is also called the *reachable subsystem* of (22.1) or (22.14). The eigenvalues of A_1, which we may refer to as the *reachable eigenvalues*, govern the ZIR in the reachable subspace. Also, the behavior of $z_2(k)$ is described by the *undriven* state-space model

$$z_2(k+1) = A_2 z_2(k) \tag{22.17}$$

and is governed by the eigenvalues of A_2, which we may call the *unreachable* eigenvalues.

There is no loss of generality in assuming a given unreachable system has been put in the standard form for unreachable systems; proofs of statements about unreachable systems are often much more transparent if done in these coordinates.

Modal Reachability Tests

An immediate application of the standard form is to prove the following *modal test* for (un)reachability.

Theorem 22.2

The system (22.1) is unreachable if and only if $w^T B = 0$ for some left eigenvector w^T of A. We say that the corresponding eigenvalue λ is an unreachable eigenvalue.

Proof.

If $w^T B = 0$ and $w^T A = \lambda w^T$ with $w^T \neq 0$, then $w^T AB = \lambda w^T B = 0$ and similarly $w^T A^k B = 0$, so $w^T R_n = 0$, *i.e.* the system is unreachable.

Conversely, if the system is unreachable, transform it to the standard form (22.14). Now let w_2^T denote a left eigenvector of A_2, with eigenvalue λ. Then $w^T = [\ 0 \quad w_2^T\ ]$ is a left eigenvector of the transformed A matrix, namely $\bar{A}$, and is orthogonal to the (columns of the) transformed B, namely $\bar{B}$.

An alternative form of this test appears in the following result.

Corollary 22.2

The system (22.1) is unreachable if and only if $[\ zI - A \mid B\ ]$ loses rank for some $z = \lambda$. This λ is then an unreachable eigenvalue.

Proof.
The matrix $[\ zI - A \mid B\ ]$ has less than full rank at $z = \lambda$ iff $w^T[\ sI - A \mid B\ ] = 0$ for some $w^T \neq 0$. But this is equivalent to having a left eigenvector of A being orthogonal to (the columns of) B.

Example 22.2

Consider the system

$$x(k+1) = \underbrace{\begin{matrix} 3 & 0 \\ 0 & 3 \end{matrix}}_{A} x(k) + \underbrace{\begin{matrix} 1 \\ 1 \end{matrix}}_{B} u(k)$$

Left eigenvectors of A associated with its eigenvalue at $\lambda = 3$ are $w_1^T = [\ 1 \quad 0\ ]$ and $w_=^T[\ 0 \quad 1\ ]$, neither of which is orthogonal to B. However, $w_0^T = [\ 1 \quad -1\ ]$ is *also* a left eigenvector associated with $\lambda = 3$, and *is* orthogonal to B. This example drives home the fact that the modal unreachability test only asks for *some* left eigenvector to be orthogonal to B.

Jordan Chain Interpretation

Recall that the system (22.1) may be thought of as having a collection of "Jordan chains" at its core. Reachability, which we first introduced in terms of reaching target states, turns out to also describe our ability to independently "excite" or drive the Jordan chains. This is the implication of the reachable subspace being an A-invariant subspace, and is the reason why the preceding modal tests for reachability exist.

The critical thing for reachability is to be able to excite the *beginning* of each chain; this excitation can then propagate down the chain. An additional condition is needed if several chains have the same eigenvalue; in this case, we need to be able to *independently* excite the beginning of each of these chains. (Example 22.2 illustrates that reachability is lost otherwise; with just a single input, we are unable to excite the two identical chains independently.) With distinct eigenvalues, we do not need to impose this independence condition; the distinctness of the eigenvalues permits independent motions.

Some additional insight is obtained by considering the distinct eigenvalue case in more detail. In this case, A in (22.1) is diagonalizable, and $A = V\Lambda W$, where the columns of V are the right eigenvectors of A and the rows of W are the left eigenvectors of A. For $x(0) = 0$ we have

$$x(k) = \sum_{\ell=1}^{n} v_\ell w_\ell^T B g_\ell(k) \tag{22.18}$$

where

$$g_\ell(k) = \sum_{i=0}^{k-1} \lambda_\ell^{k-i-1} u(i) \tag{22.19}$$

If $w_j^T B = 0$ for some j, then (22.18) shows that $x(k)$ is confined to the span of $\{v_\ell\}_{\ell \neq j}$, *i.e.* the system is not reachable. For example, suppose we have a second-order system ($n = 2$), and suppose $w_1^T B = 0$. Then if $x(0) = 0$, the response to *any* input must lie along v_2. This means that v_2 spans the reachable space, and that any state which has a component along v_1 is not reachable.

Exercises

Exercise 22.1 Suppose you are given the single-input, nth-order system $x(k+1) = Ax(k) + bu(k)$, and assume the control u at every time step is confined to lie in the interval $[0, 1]$. Assume also that an eigenvalue of A, say λ_1, is real and nonnegative. Show that the set of states reachable from the origin is confined to one side of a hyperplane through the origin in $\mathcal{R}^n$. (Hint: An eigenvector associated with λ_1 will help you make the argument.)

[A hyperplane through the origin is an $(n-1)$-dimensional subspace defined as the set of vectors x in $\mathcal{R}^n$ for which $a'x = 0$, where a is some fixed nonzero vector in $\mathcal{R}^n$. Evidently a is normal to the hyperplane. The two "sides" of the hyperplane, or the two "half-spaces" defined by it, are the sets of x for which $a'x \leq 0$ and $a'x \geq 0$.]

Exercise 22.2 Given the system

$$x(k+1) = \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} x(k) + \begin{pmatrix} d \\ e \end{pmatrix} u(k)$$

where a, b, c, d, e are scalars, deduce precisely what condition these coefficients satisfy when the system is *not* reachable. Draw a block diagram corresponding to the above system and use it to interpret the following special cases in which reachability is lost: (a) $e = 0$; (b) $b = 0$ and $d = 0$; (c) $b = 0$ and $c = a$.

Exercise 22.3 **(a)** Given m-input system $x(k+1) = Ax(k) + Bu(k)$, where A is the Jordan-form matrix

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 0 & 3 \end{pmatrix}$$

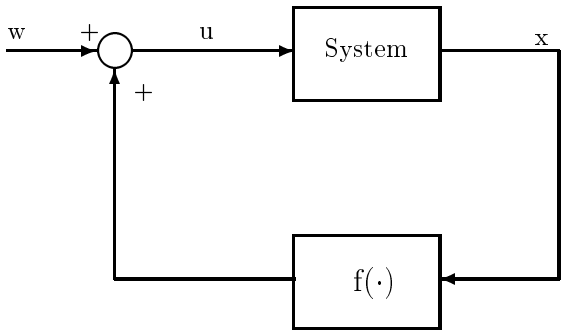
obtain conditions that are necessary and sufficent for the system to be reachable. (Hint: Your conditions should involve the *rows* b_i of B. Some form of the modal reachability test will — not surprisingly! — lead to the simplest solution.)

(b) Generalize this reachability result to the case where A is a general $n \times n$ Jordan-form matrix.

(c) Given the *single-input, reachable* system $x(k+1) = Ax(k) + bu(k)$, show that there can be only *one* Jordan block associated with each distinct eigenvalue of A.

Exercise 22.4 Given the n-dimensional reachable system $x(k+1) = Ax(k) + Bu(k)$, suppose that $u(k)$ is generated according to the nonlinear feedback scheme shown in the figure, where $u(k) = w(k) + f(x(k))$, with $f(.)$ being an arbitrary but known function, and $w(k)$ being the new control input for the closed-loop system.

Show that $w(k)$ can always be chosen to take the system state from the origin to any specified target state in no more than n steps. You will thereby have proved that *reachability is preserved under (even nonlinear) state feedback.*



$$x_{k+1} = Ax_k + B(w_k + f(x_k))$$

Exercise 22.5 Consider the following linear SISO System, Σ:

$$\begin{aligned} x(k+1) &= A(k)x(k) + B(k)u(k) \\ y(k) &= C(k)x(k) + D(k)u(k) \end{aligned}$$

where $A(k) = A(k+N)\,\forall k \geq 0$, similarly for $B(k), C(k)$, and $D(k)$.

(a) Show that Σ is N-Periodic, i.e., for zero initial conditions, show that if y is the output response for some input u, then $y(k-N)$ is the output response for $u(k-N)$. Assume for simplicity that $u(k) = 0$ for $k < 0$.

We want to get a different representation of this system that is easier to work with. To achieve this, we will group together every N successive inputs starting from $k = 0$. We will also do the same for the output. To be more precise, we will define a mapping L, called a *lifting*, such that

$$L : (u(0),\, u(1),\, u(2),\, \ldots,\, u(k),\, \ldots) \to \tilde{u}$$

where

$$\tilde{u} = \left(\begin{pmatrix} u(0) \\ u(1) \\ \vdots \\ u(N-1) \end{pmatrix}, \begin{pmatrix} u(N) \\ u(N+1) \\ \vdots \\ u(2N-1) \end{pmatrix}, \ldots, \begin{pmatrix} u(kN) \\ u(kN+1) \\ \vdots \\ u((k+1)N-1) \end{pmatrix}, \ldots \right).$$

Similarly, $L : y \to \tilde{y}$.

(b) Show that the system mapping $\tilde{u}$ to $\tilde{y}$ is linear time invariant. We will denote this by $\tilde{\Sigma}$, the lifted system. What are the dimensions of the inputs and outputs. (In other words, by lifting the inputs and outputs, we got rid of the periodicity of the system and obtained a Multi-Input Multi-Output System).

(c) Give a state-space description of the lifted system. (Hint: Choose as a state variable $\tilde{x}(k) = x(kN)$, i.e., samples of the original state vector. Justify this choice).

(d) Show that the reachable subspace of the lifted system $\tilde{\Sigma}$ is included in the reachable subspace of the periodic system Σ. Show that the converse is true if the periodic system is reachable in T steps with $T = rN$ (a multiple of the period).

(e) Is it true that reachability of the periodic system Σ implies reachability of the lifted system $\tilde{\Sigma}$. Prove or show a counter example.

6.241J Dynamic Controls
Spring 2011