

Informal Lecture Notes for

M E 4811

M o d e r n C o n t r o l S y s t e m s

Fotis A . Papoulias
Summer 1992
Revised: Summer 1995
Naval Postgraduate School
Monterey, California

Preface

These notes were developed in order to supplement the lectures for the ME 4811, a course on state space analysis and design of control systems. The contents reflect the influence from other prerequisite courses as well as electives in the Dynamics and Control Group of the ME Department. The material of the course is tailored around one academic quarter (11 weeks) with 5 contact hours (4 for lectures and 1 for examples) per week.

References

These notes utilize material from the following references:

1. Bryson, A.E. and Ho, Y.C., Applied Optimal Control, John Wiley & Sons, 1975.
2. Franklin, G.F., Powell, J.D., and Emami-Naeini, A., Feedback Control of Dynamic Systems, Addison-Wesley, 1986.
3. Friedland, B., Control System Design, McGraw-Hill, 1986.
4. Leigh, J.R., Essentials of Nonlinear Control Theory, Peter Peregrinus Ltd., London, UK.
5. Noton, M., Modern Control Engineering, Pergamon Press, 1972.
6. Parsons, M.G., Informal notes for NA 530, Department of Naval Architecture and Marine Engineering, The University of Michigan, Ann Arbor, 1982.

Contents

1 INTRODUCTION

Unlike "classical control" theory (ME 3801) which is based on Laplace transform representations, "modern control" deals directly with systems described in ordinary differential equation form. We assume that given a physical system, we have already developed our equations of motion, in other words the modeling part is complete. The goal here is to affect the dynamic response of the system such that it performs a specific task in a satisfactory way. The first thing we have to do is to rewrite our differential equations of motion in their state space form.

1.1 State Variable System Description

The state is a set of quantities such that given initial conditions $x(t_0)$ and all future inputs $u(t)$, all future response $x(t)$ for $t > t_0$ is uniquely determined. If not enough initial conditions are specified, then more than one responses may be obtained; if too many initial conditions are specified, then a solution may not be possible. Therefore, we can see that for any dynamical system the number of states is unique; the choice, however, is not.

The state equations are a coupled set of first order linear differential equations in the state variables; i.e.,

$$\dot{x} = A x + B u ;$$

where

$$\begin{aligned} x & : \text{state vector; } n \times 1 ; \\ A & : \text{open loop dynamics matrix; } n \times n ; \\ u & : \text{control vector; } m \times 1 ; \\ B & : \text{control distribution matrix; } n \times m ; \end{aligned}$$

along with the output equation

$$y = C x ;$$

where

$$\begin{aligned} y & : \text{output vector; } r \times 1 ; \\ C & : \text{sensor calibration matrix; } r \times n ; \end{aligned}$$

Physically, for mechanical systems, x represents the collection of positions and velocities of the body (so for a complete description this must be twice the number of degrees of freedom), u is the various actuators (such as thrusters, rudders, propulsors), and y the outputs (what is available to us through observation or measurements).

As an example, consider the spring-mass-damper system shown in Figure 1. The equations of motion are

$$\begin{aligned} m_a \ddot{x}_a + k_a x_a + c_a \dot{x}_a + c_1 (x_a - x_b) & = f(t) ; \\ m_b \ddot{x}_b + k_b x_b + c_b \dot{x}_b + c_1 (x_b - x_a) & = 0 ; \end{aligned}$$

If we take as states the position and velocity of each mass

$$\begin{aligned}x_1 &= x_a ; \\x_2 &= \dot{x}_a ; \\x_3 &= x_b ; \\x_4 &= \dot{x}_b ;\end{aligned}$$

we have the equations in state form as

$$\begin{aligned}\dot{x}_1 &= x_2 ; \\ \dot{x}_2 &= -\left[\frac{k_a}{m_a} x_1 + \frac{c_a + c_1}{m_a} x_2 + \frac{c_1}{m_a} x_4 + \frac{1}{m_a} f \right] ; \\ \dot{x}_3 &= x_4 ; \\ \dot{x}_4 &= -\left[\frac{k_b}{m_b} x_3 + \frac{c_b + c_1}{m_b} x_4 + \frac{c_1}{m_b} x_2 \right] ;\end{aligned}$$

and the A , B matrices are

$$A = \begin{matrix} \begin{matrix} \mathbf{2} \\ \mathbf{6} \\ \mathbf{4} \end{matrix} & \begin{matrix} 0 & 1 & 0 & 0 \\ \frac{k_a}{m_a} & -\frac{c_a + c_1}{m_a} & 0 & \frac{c_1}{m_a} \\ 0 & 0 & 0 & 1 \\ 0 & \frac{c_1}{m_b} & \frac{k_b}{m_b} & -\frac{c_b + c_1}{m_b} \end{matrix} & \begin{matrix} \mathbf{3} \\ \mathbf{7} \\ \mathbf{7} \\ \mathbf{7} \\ \mathbf{5} \end{matrix} \end{matrix} ;$$

and

$$B = \begin{matrix} \begin{matrix} \mathbf{2} \\ \mathbf{6} \\ \mathbf{4} \end{matrix} & \begin{matrix} 0 & \mathbf{3} \\ \frac{1}{m_a} & \mathbf{7} \\ 0 & \mathbf{7} \\ 0 & \mathbf{5} \end{matrix} \end{matrix} ;$$

It should be emphasized that here we treat the external force f as our control input, this is of course legitimate if we can and are willing to change f at will so that we can affect the response of the system. This is not always the case of course; there are external forces that affect a given system and they act despite our will or even knowledge. These are called disturbances, and a more general form of the state equations is

$$\dot{x} = A x + B u + j w ;$$

where

$$\begin{aligned}w &: \text{disturbance vector; } d \times 1 ; \\j &: \text{disturbance distribution matrix; } n \times d ;\end{aligned}$$

The above equations are linear; many dynamical systems, however, yield nonlinear equations of motion. The control design problem is significantly simplified when dealing with

linear equations and in such a case we need to linearize the original nonlinear equations about a nominal operating point. This nominal point is physically defined usually by the designer and, roughly speaking, should be the condition where the system is expected to spend most of its life at. Usually, this is some sort of static equilibrium of the system which corresponds to a specified value for the control effort.

To formalize things say we have a nonlinear system of state equations

$$\dot{\underline{x}} = \underline{f}(\underline{x}; \underline{u}) ;$$

Fix the control vector $\underline{u} = \underline{u}_0$, then

$$\dot{\underline{x}} = \underline{f}(\underline{x}; \underline{u}_0) ;$$

Solve the nonlinear coupled algebraic set of equations

$$\underline{f}(\underline{x}; \underline{u}_0) = \underline{0} ;$$

to get the solution $\underline{x} = \underline{x}_0$. This is our nominal point, and solution of this set of equations is the most difficult part of the linearization process. Once \underline{x}_0 has been obtained, we linearize $\dot{\underline{x}} = \underline{f}(\underline{x}; \underline{u})$ around the nominal point $(\underline{x}; \underline{u}) = (\underline{x}_0; \underline{u}_0)$. To do this we expand in Taylor series and keep the first order terms only,

$$\underline{f}(\underline{x}; \underline{u}) = \frac{\partial \underline{f}}{\partial \underline{x}} \bigg|_{(\underline{x}_0; \underline{u}_0)} (\underline{x} - \underline{x}_0) + \frac{\partial \underline{f}}{\partial \underline{u}} \bigg|_{(\underline{x}_0; \underline{u}_0)} (\underline{u} - \underline{u}_0) ;$$

Then by assuming the change in coordinates

$$\begin{aligned} \underline{x} &= \underline{x} - \underline{x}_0 ; \\ \underline{u} &= \underline{u} - \underline{u}_0 ; \end{aligned}$$

the linearized system becomes

$$\dot{\underline{x}} = \underline{A} \underline{x} + \underline{B} \underline{u} ;$$

where \underline{A} and \underline{B} are the constant Jacobian matrices of partial derivatives evaluated at the nominal point $(\underline{x}_0; \underline{u}_0)$

$$\begin{aligned} \underline{A} &= \frac{\partial \underline{f}}{\partial \underline{x}} \bigg|_{(\underline{x}_0; \underline{u}_0)} ; \\ \underline{B} &= \frac{\partial \underline{f}}{\partial \underline{u}} \bigg|_{(\underline{x}_0; \underline{u}_0)} ; \end{aligned}$$

The elements of \underline{A} are given by

$$\underline{A} = [a_{ij}] ; \quad \text{where} \quad a_{ij} = \frac{\partial f_i}{\partial x_j} ;$$

and similarly for \underline{B} .

As an example, consider the simple pendulum shown in Figure 2. The equation of motion is

$$m \ddot{\mu} + m g \sin \mu = T ;$$

or

$$\ddot{\mu} + \frac{g}{l} \sin \mu = \frac{T}{m l} ; \quad \frac{g}{l} = \frac{g}{l} ;$$

Select as state variables

$$\begin{aligned} x_1 &= l \mu ; \\ x_2 &= \dot{\mu} ; \end{aligned}$$

The state equations are then

$$\begin{aligned} \dot{x}_1 &= l x_2 ; \\ \dot{x}_2 &= - \frac{g}{l} \sin \frac{x_1}{l} + \frac{T}{m l} ; \end{aligned}$$

For equilibrium (with no excitation, $T = 0$)

$$\begin{aligned} \sin \frac{x_1}{l} = 0 \quad) \quad (x_1)_0 = 0 \quad \text{or} \quad (x_1)_0 = \pi l ; \\ \dot{x}_2 = 0 \quad) \quad (x_2)_0 = 0 ; \end{aligned}$$

If we choose the down position to linearize we get

$$\sin \frac{x_1}{l} = \frac{x_1}{l}$$

and the linearized equations are

$$\begin{aligned} \dot{x}_1 &= l x_2 ; \\ \dot{x}_2 &= - \frac{g}{l} x_1 + \frac{T}{m l} ; \end{aligned}$$

or

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & l \\ -\frac{g}{l} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{T}{m l} \end{bmatrix} ;$$

A
X
B

Example: Consider the following equations of motion for a submarine in the dive plane (refer to Figure 3)

$$\begin{aligned} (m \dot{z}_w \dot{w} + (Z_q + m x_G) \dot{q} = Z_w U w + (Z_q + m) U q + m Z_G q^2 \\ + (W \dot{z}_B) \cos \mu + Z_{\pm} U^2 \pm ; \\ (I_y \dot{q} + M_q) \dot{q} + (M_w + m x_G) w = M_w U w + (M_q + m x_G) U q \\ + (x_G W \dot{z}_B) \cos \mu + (Z_G W \dot{z}_B) \sin \mu + m Z_G w q + M_{\pm} U^2 \pm ; \\ \dot{\mu} = q ; \\ \dot{z} = \dot{z}_w \sin \mu + w \cos \mu ; \end{aligned}$$

where

- U = forward speed ;
- w = heave velocity ;
- q = pitch rate ;
- μ = pitch angle ;
- \pm = dive plane angle ;
- z = depth ;
- W = weight ;
- B = buoyancy ;
- m = mass ;
- I_y = mass moment of inertia ;
- $(x_G ; z_G)$ = coordinates of center of gravity ;
- $(x_B ; z_B)$ = coordinates of center of buoyancy ;
- Z_w = heave force hydrodynamic coefficient ;
- M_q = pitch moment hydrodynamic coefficient ;

Now say we want to linearize these equations for a level flight path when the dive plane angle is zero, $\pm_0 = 0$. Then by setting all time derivatives to zero (this corresponds to equilibrium) we get

$$\begin{aligned} Z_w U w_0 + (W - B) \cos \mu_0 &= 0 ; \\ M_w U w_0 - (x_G W - x_B B) \cos \mu_0 - (z_G W - z_B B) \sin \mu_0 &= 0 ; \\ q_0 &= 0 ; \\ -U \sin \mu_0 + w_0 \cos \mu_0 &= 0 ; \end{aligned}$$

If we assume that the boat is neutrally buoyant $x_G = x_B$ and $W = B$, we have

$$\begin{aligned} Z_w U w_0 &= 0 ; \\ M_w U w_0 - (z_G - z_B) B \sin \mu_0 &= 0 ; \\ -U \sin \mu_0 + w_0 \cos \mu_0 &= 0 ; \end{aligned}$$

from which we can get the nominal position

$$w_0 = q_0 = 0 ; \quad \text{and} \quad \sin \mu_0 = 0 ;$$

which means

$$\mu_0 = 0 ; \quad \text{or} \quad \mu_0 = \frac{1}{2} \pi ;$$

These correspond to the two possible static equilibrium positions, like a regular or like an inverted pendulum.

If we choose to linearize around the $\mu_0 = 0$ equilibrium we have

$$q^2 = (2q_0)q = 0 ;$$

1. State equations from block diagram

Suppose we have the block diagram shown in Figure 4, and we want to write a set of state equations for this system. We observe that the system is third order (it has three integrators, so its characteristic equation will be third order). Therefore, we need three state equations and three states. One choice is to take as states the outputs of the integrator blocks. This way we get

$$\begin{aligned} \dot{x}_1 &= x_2 ; \\ \dot{x}_2 &= x_3 ; \\ \dot{x}_3 &= -6x_1 - 11x_2 - 6x_3 + 6u ; \end{aligned}$$

and the output equation

$$y = x_1 ;$$

The A, B, and C matrices are

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -6 & -11 & -6 \end{bmatrix} ; \quad B = \begin{bmatrix} 0 \\ 0 \\ 6 \end{bmatrix} ; \quad C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} ;$$

We note that the above choice of states is not unique, we could have selected as states the outputs of the three feedback blocks; this would have produced a different but equivalent (with the same input/output relationship) system of state equations.

2. Block diagram from state equations

Consider the following system of state equations

$$\begin{aligned} \dot{x}_1 &= a_{11}x_1 + a_{12}x_2 + b_1u ; \\ \dot{x}_2 &= a_{21}x_1 + a_{22}x_2 + b_2u ; \\ y &= c_1x_1 + c_2x_2 ; \end{aligned}$$

The A, B, C matrices here are

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} ; \quad B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} ; \quad C = \begin{bmatrix} c_1 & c_2 \end{bmatrix} ;$$

The block diagram is constructed as shown in Figure 5.

3. Block diagram and state equations from differential equation

Consider the transfer function between input u and output y

$$\frac{y}{u} = \frac{b_1s + b_0}{s^3 + a_2s^2 + a_1s + a_0} ;$$

which is equivalent to the differential equation

$$y^{(iii)} + a_2\ddot{y} + a_1\dot{y} + a_0y = b_1\dot{u} + b_0u ;$$

This is a third order system, so we need three states. Let our first state be

$$x_1 = y;$$

so

$$y = \begin{bmatrix} \mathbf{h} \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{1} & \mathbf{6} \\ \mathbf{4} & \mathbf{7} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix};$$

Substitute $x_1 = y$ into the equation,

$$x_1^{(iii)} + a_2 \ddot{x}_1 + a_1 \dot{x}_1 + a_0 y = b_1 \underline{u} + b_0 u;$$

To lower the order let

$$\underline{x}_1 = x_2; \quad \text{this is our first state equation}$$

and substitute again

$$\ddot{x}_2 + a_2 \underline{x}_2 + a_1 x_2 + a_0 x_1 = b_1 \underline{u} + b_0 u;$$

Now if we substitute $x_3 = \underline{x}_2$ we see that the \underline{u} term in the equation will survive, and this goes against our general state space form $\underline{x} = A x + B u$. To eliminate the \underline{u} term we substitute

$$x_3 = \underline{x}_2 + b_1 u \quad \text{or}$$

$$\underline{x}_2 = x_3 - b_1 u \quad \text{this is our second state equation}$$

One more substitution will then result in

$$\underline{x}_3 + b_1 \underline{u} + a_2 x_3 + a_2 b_1 u + a_1 x_2 + a_0 x_1 = b_1 \underline{u} + b_0 u;$$

or

$$\underline{x}_3 = -a_2 x_3 - a_1 x_2 - a_0 x_1 + (b_0 - a_2 b_1) u;$$

which is the third state equation.

The state equations are

$$\begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{1} & \mathbf{6} \\ \mathbf{4} & \mathbf{7} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} \mathbf{2} & \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{1} \end{bmatrix} \begin{bmatrix} \mathbf{3} \mathbf{2} & \mathbf{3} & \mathbf{2} & \mathbf{0} \\ \mathbf{7} \mathbf{6} & \mathbf{7} & \mathbf{6} & \mathbf{0} \\ \mathbf{7} \mathbf{4} & \mathbf{7} & \mathbf{6} & \mathbf{0} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ u \end{bmatrix} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{b}_1 \\ \mathbf{0} & \mathbf{b}_0 - \mathbf{a}_2 \mathbf{b}_1 \end{bmatrix} u;$$

and the output equation

$$y = \begin{bmatrix} \mathbf{h} \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{1} & \mathbf{6} \\ \mathbf{4} & \mathbf{7} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix};$$

The above form of the A matrix is called a companion form (negative coefficients in the last row, and ones in the superdiagonal).

The block diagram appears as shown in Figure 6.

1.3 From State Equations to Transfer Function

Consider the standard state space system

$$\begin{aligned}\underline{x} &= A \underline{x} + B u ; \\ y &= C \underline{x} ;\end{aligned}$$

In the Laplace domain (with zero initial conditions) this becomes

$$\begin{aligned}sX(s) &= A X(s) + B U(s) ; \\ Y(s) &= C X(s) ;\end{aligned}$$

or

$$\begin{aligned}(sI - A)X &= B U \Rightarrow X = (sI - A)^{-1} B U ; \\ Y &= C (sI - A)^{-1} B U ;\end{aligned}$$

If we compare the last expression with

$$Y(s) = G(s)U(s) ; \quad \text{where } G(s) \text{ is the transfer function}$$

we can see that

$$G(s) = C (sI - A)^{-1} B ;$$

is the transfer function of the system. This is of the familiar ME 3801 form only in the case of a single input single output (SISO) system (i.e., both u and y are scalars instead of vectors). In the more general case of a multiple input multiple output system (MIMO), it is a transfer function matrix and its individual elements consist of transfer functions in the usual sense. It can be thought of as a matrix of influence coefficients (the ij element of the matrix depicts the transfer function between the i th output and the j th input).

The above helps in constructing compact generic block diagrams, as shown in Figure 7.

$$\underline{x} = A \underline{x} + B u ; \quad y = C \underline{x}$$

1.4 Poles and Zeros

Recall that for a system in the form

$$\underline{x} = A \underline{x} + B u ; \quad y = C \underline{x}$$

its transfer function is written as

$$G(s) = C (sI - A)^{-1} B ;$$

The poles of the transfer function are defined as those values of s where the denominator goes to zero. This means that

$$\begin{aligned}(sI - A) &\text{ is a singular matrix, or} \\ \det[sI - A] &= 0 \text{ or} \\ s &= \text{eigenvalue of } A ;\end{aligned}$$

The zeros of the transfer function are usually defined for SISO systems. In such a case we have

$$G(s) = \det C (sI - A)^{-1} B ;$$

and using properties of the determinant we get

$$\begin{aligned} \det[C (sI - A)^{-1} B] &= \frac{\det[sI - A] \det[C (sI - A)^{-1} B]}{\det[sI - A]} \\ &= \det \begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} \\ &= \frac{\det[sI - A]}{\det[sI - A]} \end{aligned}$$

where we used the fact that

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det A \det[D - CA^{-1}B] ;$$

Therefore, the zeros of $G(s)$ are solutions of

$$\det \begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} = 0 ;$$

As an example, say we have the system

$$\begin{aligned} \dot{x}_1 &= -3x_1 + x_2 + u ; \\ \dot{x}_2 &= 2u ; \\ y &= x_1 ; \end{aligned}$$

The matrices A , B , C are

$$A = \begin{bmatrix} -3 & 1 \\ 0 & 0 \end{bmatrix} ; B = \begin{bmatrix} 1 \\ 2 \end{bmatrix} ; C = \begin{bmatrix} 1 & 0 \end{bmatrix} ;$$

The poles of the system are

$$\det[sI - A] = \det \begin{bmatrix} s+3 & 1 \\ 0 & s \end{bmatrix} = s(s+3) = 0 \Rightarrow s = 0 ; -3 ;$$

and the zeros

$$\det \begin{bmatrix} sI - A & B \\ C & 0 \end{bmatrix} = \det \begin{bmatrix} s+3 & 1 & 1 \\ 0 & s & 2 \\ 1 & 0 & 0 \end{bmatrix} = 2 + s = 0 \Rightarrow s = -2 ;$$

To verify this, let's get $G(s)$ using classical methods:

$$\begin{aligned} \dot{y} &= -3y + x_2 + u ; \text{ or} \\ \ddot{y} &= -3\dot{y} + 2u + \dot{u} ; \text{ or} \\ \ddot{y} + 3\dot{y} &= \dot{u} + 2u ; \text{ or} \\ Y(s^2 + 3s) &= U(s + 2) ; \text{ or} \\ \frac{Y(s)}{U(s)} &= \frac{s + 2}{s(s + 3)} ; \end{aligned}$$

which agrees with the poles and zeros from state space. These poles and zeros are usually called open loop poles and zeros since no feedback control action has been defined yet.

Example: Consider the state equations for the submarine example, where the state vector is

$$\mathbf{x} = [\mu; w; q];$$

the output vector is the pitch angle

$$y = \mu;$$

and the control input u is the dive plane angle \pm

$$u = \pm;$$

The state equations are the same as before. Typical values for the coefficients are

$$\begin{aligned} a_{11} &= -j 0.064390823; & a_{12} &= -j 0.1420481; & a_{13} &= 0.1353290; \\ a_{21} &= 0.025208820; & a_{22} &= -j 0.1479027; & a_{23} &= -j 0.3599404; \\ b_1 &= 0.0012883232; & b_2 &= -j 0.0034266096; \\ z_{GB} &= 0.1 \text{ ft}; & U &= 5 \text{ ft} \cdot \text{sec}; \end{aligned}$$

Using MATLAB and the above values we can find the transfer function

$$\frac{\mu}{\pm} = \frac{-j 0.0857s - j 0.0235}{s^3 + 1.0615s^2 + 0.3636s + 0.0099};$$

and we can see that the open loop poles are simply the roots of the denominator polynomial

$$-j 0.5159 \pm 0.2584i; \quad -j 0.0297;$$

These are also given by the eigenvalues of matrix A . Notice that the system is open loop stable. This means that with no control action \pm , if an initial disturbance is introduced in the angle μ , it will go back to zero asymptotically. As the metacentric height z_{GB} gets closer to zero, one open loop pole goes to zero. (Can you see this from the form of the A matrix? What is the physical significance of a zero pole?) The open loop zero is the root of the numerator of the transfer function

$$-j 0.2742;$$

The transfer function can also be computed by starting with the equations of motion

$$\begin{aligned} \ddot{\mu} &= \ddot{q}; \\ \ddot{w} &= a_{13}z_{GB}\mu + a_{11}Uw + a_{12}Uq + b_1U^2\pm; \\ \ddot{q} &= a_{23}z_{GB}\mu + a_{21}Uw + a_{22}Uq + b_2U^2\pm; \end{aligned}$$

constructing the block diagram from \pm to μ , and reducing it, as we did in Section 1.2.

1.5 Time Response Using State Equations

There are two ways to compute the time response of a system using the state equations: numerical and analytical.

1. Numerical

State equations are naturally used in digital computer simulation. For example, if we use Euler's integration: given $x(0)$ and $u(0)$ at $t = 0$, then

$$x(t + \Delta t) = x(t) + \dot{x}(t) \Delta t :$$

Δt is the integration time step which must be selected small enough (with respect to the natural time constant of the system) for results to be valid; and $\dot{x}(t) = Ax(t) + Bu(t)$, in other words we evaluate \dot{x} using the current value of x and u . Continuing the scheme, we get

$$\begin{aligned} x(\Delta t) &= x(0) + [Ax(0) + Bu(0)]\Delta t ; \\ x(2\Delta t) &= x(\Delta t) + [Ax(\Delta t) + Bu(\Delta t)]\Delta t ; \end{aligned}$$

and so on. Although Euler's method is the simplest and most inaccurate numerical integration technique available, it is good enough for naval engineering problems where things do not change very fast in time.

2. Analytical

We want the transient solution for

$$\dot{x} = Ax ; \quad x(t_0) = x(0) ;$$

where x is the $n \times 1$ state vector, A is the $n \times n$ open loop dynamics matrix, and $x(0)$ is the $n \times 1$ vector of initial conditions. Recall that for a n th order system ($n = 1$) we would have

$$\dot{x} = ax ; \quad x(t_0) = x(0) :$$

If we assume

$$x = \alpha e^{st} ;$$

we get

$$\begin{aligned} \alpha j ax &= 0 \text{ or} \\ \alpha e^{st}(s - a) &= 0 \text{ or} \\ s &= a ; \text{ an eigenvalue :} \end{aligned}$$

Therefore, the solution is

$$x = \alpha e^{at} ;$$

The unknown constant α can be computed from the initial condition

$$x(t_0) = \alpha e^{at_0} = x(0) ;$$

giving

$$\otimes = x(0)e^{i a t_0} :$$

The solution is then

$$x(t) = e^{a(t-t_0)}x(0) ;$$

where

$$e^{a(t-t_0)} = 1 + \frac{a(t-t_0)}{1!} + \frac{[a(t-t_0)]^2}{2!} + \frac{[a(t-t_0)]^3}{3!} + \dots$$

When the solution is extended to a matrix system ($n > 1$), the results are completely parallel,

$$\underline{\dot{x}} = A x ;$$

with solution

$$\underset{\text{vector}}{x(t)} = \underset{\text{matrix}}{e^{A(t-t_0)}} \underset{\text{vector}}{x(0)} ;$$

where the matrix exponential is defined through a series expansion analogously to its scalar counterpart

$$e^{A(t-t_0)} = I + \frac{A(t-t_0)}{1!} + \frac{[A(t-t_0)]^2}{2!} + \frac{[A(t-t_0)]^3}{3!} + \dots$$

This is called the state transition matrix denoted by

$$\phi(t, t_0) = e^{A(t-t_0)} :$$

The state transition matrix expresses how the state is changed from its value at t_0 to the state at t by the system with open loop dynamics given by A

$$x(t) = \phi(t, t_0)x(t_0) :$$

We can obtain the complete solution with a control input $u(t)$ as:

$$\frac{d}{dt} e^{i A t} \underset{\text{vector}}{x(t)} = e^{i A t} \underset{\text{matrix}}{A} \underset{\text{vector}}{x(t)} + e^{i A t} \underset{\text{matrix}}{B} u(t) ;$$

$$\underline{\dot{x}}(t) = A x + B u$$

Integrating,

$$e^{i A t} \underset{\text{vector}}{x}(t) = \int_{t_0}^t e^{i A \lambda} \underset{\text{matrix}}{B} u(\lambda) d\lambda + c ;$$

where c is a vector constant of integration. Now at $t = t_0$ we have

$$e^{i A t_0} \underset{\text{vector}}{x}(0) = c ;$$

giving

$$e^{i A t} \underset{\text{vector}}{x}(t) = \int_{t_0}^t e^{i A \lambda} \underset{\text{matrix}}{B} u(\lambda) d\lambda + e^{i A t_0} \underset{\text{vector}}{x}(0) ;$$

Multiplying through by $e^{A t}$

$$\dot{x}(t) = e^{A(t-t_0)} x(0) + \int_{t_0}^t e^{A(t-\zeta)} B u(\zeta) d\zeta ; \quad t \geq t_0 ;$$

or

$$x(t) = \underbrace{e^{A(t-t_0)} x(0)}_{\text{transient}} + \underbrace{\int_{t_0}^t e^{A(t-\zeta)} B u(\zeta) d\zeta}_{\text{steady state}} ;$$

In most cases

transient = response due to initial state

and this will go to zero for a stable system, while

steady state = response due to input

is given by the above convolution integral. For linear systems, the total response is of course the sum of the two responses.

The matrix exponential $e^{A t}$ can be computed using a couple of different ways.

² One way is with the above power series expansion

$$e^{A t} = I + \frac{A t}{1!} + \frac{(A t)^2}{2!} + \frac{(A t)^3}{3!} + \dots ;$$

This is efficient only numerically when the series can be truncated to an arbitrary degree of accuracy. In general, these Taylor series are used to define rather than to compute functions of a matrix (take a 2 x 2 matrix and try to find its cosine using the appropriate series expansion; then check your answer using MATLAB).

² If A can be diagonalized; i.e., if $\alpha = T^{-1} A T$ where T is the matrix of eigenvectors of A and α the diagonal matrix of the eigenvalues of A,

$$\alpha = \text{diag} \{ \lambda_1, \lambda_2, \dots, \lambda_n \} ;$$

then

$$e^{A t} = T^{-1} e^{\alpha t} T ;$$

where

$$e^{\alpha t} = \text{diag} \{ e^{\lambda_1 t}, e^{\lambda_2 t}, \dots, e^{\lambda_n t} \} ;$$

We can easily see from the last expression why if at least one of the eigenvalues λ_i of A is positive, the system will be unstable.

For time varying systems of the form

$$\dot{x} = A(t)x ;$$

the state transition matrix is denoted by

$$\Phi(t; t_0);$$

and the solution is given by

$$x(t) = \Phi(t; t_0)x(0);$$

Notice that the state transition matrix for time varying systems is function of both the current time t and initial time t_0 , unlike the time invariant system case where Φ was a function of one variable only, $t - t_0$, the time interval between t and t_0 . What is more unfortunate is the fact that closed form expression for $\Phi(t; t_0)$ does not exist which makes analysis and control of time varying systems much more difficult than time invariant systems considered here. As a word of caution, in general,

$$\Phi(t; t_0) \neq e^{\int_{t_0}^t A(\zeta) d\zeta};$$

except when the matrices $A(t)$ and $\int_{t_0}^t A(\zeta) d\zeta$ commute; i.e., when

$$A(t) \int_{t_0}^t A(\zeta) d\zeta = \int_{t_0}^t A(\zeta) d\zeta A(t);$$

Some general properties of the state transition matrix $\Phi(t; t_0)$ are

1. It satisfies the differential equation with identity initial conditions,

$$\begin{aligned} \dot{\Phi}(t; t_0) &= -A(t)\Phi(t; t_0); \\ \Phi(t_0; t_0) &= I; \end{aligned}$$

2. It satisfies the semigroup property,

$$\Phi(t; t_0) = \Phi(t; t_1)\Phi(t_1; t_0);$$

3. It is always nonsingular,

$$\Phi^{-1}(t; t_0) = \Phi(t_0; t);$$

4. It has a computable determinant,

$$\det \Phi(t; t_0) = e^{-\int_{t_0}^t \text{trace} A(\zeta) d\zeta};$$

The main advantages of using the state transition matrix in system dynamics are two:

1. Helps in proving other theorems.
2. Once it has been determined, it makes calculation of the particular solution in response to some initial conditions and input, much faster.

In general, the analytic method of solution is employed only for theoretic purposes or in special circumstances; in almost all cases we obtain the solutions numerically. This has the added advantage that it is not restricted to linear systems, nonlinear systems can be simulated numerically in much the same way.

Example: Consider the submarine linear equations of motion

$$\begin{aligned}\dot{\mu} &= q; \\ \dot{w} &= a_{13}Z_{GB} \mu + a_{11}U w + a_{12}U q + b_1 U^2 \pm; \\ \dot{q} &= a_{23}Z_{GB} \mu + a_{21}U w + a_{22}U q + b_2 U^2 \pm;\end{aligned}$$

where we assume a dive plane deflection $\pm = \pm 0.2$ radians (± 11.5 degrees). A simulation algorithm using Euler's integration is as follows:

- ² Step 1: Choose integration time step Δt and initial conditions μ_0, w_0, q_0 . Set $i = 0$.
- ² Step 2: Using the values of μ_i, w_i, q_i , compute $\dot{\mu}_i, \dot{w}_i, \dot{q}_i$ from the equations of motion.
- ² Step 3: Compute

$$\begin{aligned}\mu_{i+1} &= \mu_i + \dot{\mu}_i \Delta t; \\ w_{i+1} &= w_i + \dot{w}_i \Delta t; \\ q_{i+1} &= q_i + \dot{q}_i \Delta t;\end{aligned}$$

- ² Step 4: Set $i = i + 1$ and go back to Step 2.

Typical results of the simulation in terms of the pitch angle μ are shown in Figure 8. As with any numerical results, however, the real question is: are they correct? The answer to this borders between art and science, and in the context of system simulations here is a set of a few checks:

1. In this particular simulation we used a time step $\Delta t = 0.01$ seconds. Is this small enough? The easiest way to check this is to reduce (or increase) Δt , say by a factor of 10, and re-run the program. If the results do not change, the above choice for Δt was good. A more rational way to do the same thing would be to look at the natural time constant of the dynamics of the system. The system poles were found in page 18. It seems that the fastest pole of the system has real part -0.5159 , and the time constant that corresponds to this is about $1/0.5$ or 2 seconds. This means that it takes a couple of seconds for the boat to "listen" to its dive planes, so $\Delta t = 0.01$ should give very accurate results. In fact in this case we could go as far as $\Delta t = 0.5$ and we would still be reasonably accurate.
2. Look again at the system eigenvalues: one of them is certainly dominant, -0.0297 , so the response should approximate that of a first order system with a time constant $1/0.0297$, or about 33.5 seconds. Now look at the response of the figure: does it take approximately 33.5 seconds to go up to 60% of its final value?

3. By now we are convinced that the transient response we see in the figure agrees with our engineering intuition. How about the final or steady state value of the response? This is something we can compute exactly. At steady state we should have, $\dot{\mu} = \dot{w} = \dot{q} = 0$, so that our equations become at steady state:

$$\begin{aligned} \dot{q} &= 0 ; \\ a_{13}Z_G B \mu + a_{11}U \dot{w} + a_{12}U \dot{q} + b_1 U^2 \dot{\mu} &; \\ a_{23}Z_G B \mu + a_{21}U \dot{w} + a_{22}U \dot{q} + b_2 U^2 \dot{\mu} &: \end{aligned}$$

Using $\dot{q} = 0$, the second and third equations give

$$\begin{aligned} a_{13}Z_G B \mu + a_{11}U \dot{w} &= -b_1 U^2 \dot{\mu} ; \\ a_{23}Z_G B \mu + a_{21}U \dot{w} &= -b_2 U^2 \dot{\mu} : \end{aligned}$$

Substituting $\dot{\mu} = -\dot{w}/0.2$ and using the values from page 17 we find

$$\mu = 0.476 \text{ radians or } 27.3 \text{ degrees} ;$$

a result which agrees with the figure.

Simulation of a nonlinear set of equations proceeds in a similar manner. Let's assume that the only important nonlinearities in our example come from the trigonometric functions and not the hydrodynamic forces and moments; in other words the nonlinear equations of motion are

$$\begin{aligned} \dot{\mu} &= -q ; \\ \dot{w} &= a_{13}Z_G B \sin \mu + a_{11}U \dot{w} + a_{12}U \dot{q} + b_1 U^2 \dot{\mu} ; \\ \dot{q} &= a_{23}Z_G B \sin \mu + a_{21}U \dot{w} + a_{22}U \dot{q} + b_2 U^2 \dot{\mu} : \end{aligned}$$

The numerical integration proceeds in exactly the same way as before; the only difference is that here the values for \dot{w} and \dot{q} are computed from the new equations. Typical results are shown in the previous figure where the difference between linear and nonlinear simulations is also shown. Naturally, whenever possible, simulations must be performed for the nonlinear systems since these model the underlying physics more accurately. The steady state value for μ can be computed from the nonlinear equations in the same way as before, the algebra is easy in the example case but keep in mind that for general nonlinear equations it may be very difficult. Here we can find

$$\sin \mu = 0.476 \quad \text{or} \quad \mu = 28.5 \text{ degrees} ;$$

1.6 Canonical Forms

Consider the general state equations

$$\begin{aligned} \dot{x} &= A x + B u ; \\ y &= C x : \end{aligned}$$

and

$$y = \begin{bmatrix} \mathbf{h} \\ b_0 \end{bmatrix} \begin{bmatrix} b_1 & b_2 \\ \mathbf{i} & \mathbf{6} \\ \mathbf{4} & \mathbf{7} \\ \mathbf{X}_3 \end{bmatrix} \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix};$$

with the block diagram shown in Figure 11.

The observer canonical form for the same system is

$$\begin{bmatrix} \mathbf{2} & \mathbf{3} & \mathbf{2} \\ \mathbf{6} & \mathbf{7} & \mathbf{4} \\ \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{i} & \mathbf{a}_2 & \mathbf{1} & \mathbf{0} \\ \mathbf{i} & \mathbf{a}_1 & \mathbf{0} & \mathbf{1} \\ \mathbf{i} & \mathbf{a}_0 & \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{3} & \mathbf{2} & \mathbf{3} \\ \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{bmatrix} + \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{6} & \mathbf{7} \\ \mathbf{4} & \mathbf{5} \end{bmatrix} \begin{bmatrix} b_2 \\ b_1 \\ b_0 \end{bmatrix} \mathbf{z}_u;$$

and

$$y = \begin{bmatrix} \mathbf{h} \\ 1 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ \mathbf{i} & \mathbf{6} \\ \mathbf{4} & \mathbf{7} \\ \mathbf{X}_3 \end{bmatrix} \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix};$$

with the block diagram shown in Figure 12.

You should, of course, verify the above forms! The main difference between the two forms is that in the control canonical form the B matrix is "clean", whereas in the observer canonical form it is the C matrix that appears to be "clean" instead. In both cases, observe that the characteristic equation of the A matrix can be obtained easily without any algebra. This is a very nice property of matrices in companion form and is true regardless of the order of the matrix. Finally, it should be emphasized that both forms represent exactly the same physical system; the definitions for the state are different in the two forms. In practice, one definition may make more sense than the other physically, and this is the one that should be chosen. Although defining convenient states may make the algebra simpler, it is much more preferable to choose as states variables that make sense physically; using MATLAB makes all linear algebra calculations relatively straightforward.

1.7 Controllability and Observability

Consider the system

$$\begin{bmatrix} \mathbf{2} & \mathbf{3} & \mathbf{2} \\ \mathbf{6} & \mathbf{7} & \mathbf{4} \\ \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{2} & \mathbf{3} & \mathbf{2} & \mathbf{1} \\ \mathbf{i} & \mathbf{2} & \mathbf{i} & \mathbf{3} & \mathbf{0} & \mathbf{0} \\ \mathbf{i} & \mathbf{2} & \mathbf{i} & \mathbf{2} & \mathbf{i} & \mathbf{4} & \mathbf{0} \\ \mathbf{i} & \mathbf{2} & \mathbf{i} & \mathbf{2} & \mathbf{i} & \mathbf{2} & \mathbf{i} & \mathbf{5} \end{bmatrix} \begin{bmatrix} \mathbf{3} & \mathbf{2} & \mathbf{3} \\ \mathbf{X}_1 & \mathbf{X}_2 & \mathbf{X}_3 \\ \mathbf{X}_4 \end{bmatrix} + \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{6} & \mathbf{7} \\ \mathbf{4} & \mathbf{5} \end{bmatrix} \begin{bmatrix} \mathbf{1} \\ \mathbf{2} \\ \mathbf{2} \\ \mathbf{5} \end{bmatrix} \mathbf{z}_u;$$

and

$$y = \begin{bmatrix} \mathbf{h} \\ 7 \end{bmatrix} \begin{bmatrix} \mathbf{6} & \mathbf{4} & \mathbf{2} \\ \mathbf{i} & \mathbf{6} & \mathbf{4} \\ \mathbf{6} & \mathbf{7} & \mathbf{4} \\ \mathbf{X}_3 & \mathbf{X}_4 \end{bmatrix} \begin{bmatrix} \mathbf{2} & \mathbf{3} \\ \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix};$$

So far, the system looks nice. Let's find the transfer function:

$$\begin{aligned}
 G(s) &= \frac{Y(s)}{U(s)} \\
 &= C(sI - A)^{-1}B \\
 &= \frac{(s+2)(s+3)(s+4)}{(s+1)(s+2)(s+3)(s+4)} \\
 &= \frac{1}{s+1};
 \end{aligned}$$

which is first order instead of fourth as the original system, due to the multiple zero/pole cancellation. To see what went wrong, let's transform the system to its normal coordinate form by diagonalizing A. The matrix of eigenvectors of A is

$$T = \begin{bmatrix} 0.7071 & 0.4082 & 0.0000 & 0.0000 \\ 0.7071 & 0.8165 & 0.4082 & 0.0000 \\ 0.0000 & 0.4082 & 0.8165 & 0.4472 \\ 0.0000 & 0.0000 & 0.4082 & 0.8944 \end{bmatrix}$$

Then using our familiar transformation

$$x = Tx^0 \text{ or } x^0 = T^{-1}x;$$

the system is transformed into

$$\begin{aligned}
 \dot{x}^0 &= Ax^0 + Bu; \\
 y &= Cx^0;
 \end{aligned}$$

where

$$\begin{aligned}
 A^0 &= T^{-1}AT = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}; \\
 B^0 &= T^{-1}B = \begin{bmatrix} 1.4142 \\ 0 \\ 2.4495 \\ 0 \end{bmatrix}; \\
 C^0 &= CT = \begin{bmatrix} 0.7071 & 0.4082 & 0 & 0 \end{bmatrix};
 \end{aligned}$$

The state equations are then

$$\begin{aligned}
 \dot{x}_1^0 &= x_1^0 + 1.4142u; \\
 \dot{x}_2^0 &= 2x_2^0; \\
 \dot{x}_3^0 &= 3x_3^0 + 2.4495u; \\
 \dot{x}_4^0 &= 4x_4^0;
 \end{aligned}$$

and the output equation

$$y = 0.7071x_1^O + 0.4082x_2^O;$$

In block diagram the system in normal coordinates appears as shown in Figure 13. Looking at this block diagram we can see the following

1. x_1^O : affected by the input; visible in the output;
2. x_2^O : unaffected by the input; visible in the output;
3. x_3^O : affected by the input; invisible in the output;
4. x_4^O : unaffected by the input; invisible in the output.

Therefore, it is fair to say that as far as the state variables go:

1. x_1^O : we can control it and we can observe it;
2. x_2^O : we can not control it but we can observe it;
3. x_3^O : we can control it but we can not observe it;
4. x_4^O : we can not control it and we can not observe it.

The normal transfer function, $G(s)$, shows the first subsystem, x_1^O , only.

In general, every system

$$\begin{aligned}\dot{\underline{x}} &= \underline{A} \underline{x} + \underline{B} u; \\ y &= \underline{C} \underline{x};\end{aligned}$$

can be divided through a series of transformations into four subsystems:

1. A controllable and observable part.
2. An uncontrollable and observable part.
3. A controllable and unobservable part.
4. An uncontrollable and unobservable part.

This is known as Kalman's decomposition theorem. The thing to remember is that the transfer function of any system is determined only by the controllable and observable subsystem. That is, the transfer function may contain less information than what is actually needed to model the complete system.

The precise definition of controllability is:

² A system is said to be state controllable if any initial state $x(t_0)$ can be driven to any final state $x(t_f)$ using possibly unbounded control $u(t)$ in finite time $t_0 < t < t_f$.

From the state equations

$$\dot{x} = \begin{matrix} \mathbf{A} \\ n \times n \end{matrix} x + B u ;$$

this should depend only on A and B . The test for controllability is as follows: Compute the

$$\text{controllability matrix } C = \begin{matrix} \mathbf{h} \\ \mathbf{B} ; \mathbf{A} \mathbf{B} ; \mathbf{A}^2 \mathbf{B} ; \dots ; \mathbf{A}^{n-1} \mathbf{B} \end{matrix} \mathbf{i} ;$$

and the system is controllable if and only if the rank of C (the number of linearly independent rows or columns) is n . Roughly speaking, C shows how possible it is to change the state of a system using the input. For a single input system B is $n \times 1$ and C is a square matrix. The test is then that C be nonsingular

$$\det C \neq 0 ;$$

We can also test controllability by transforming to the normal coordinate form (with distinct eigenvalues). The system is then controllable if $B^0 = T^{-1}B$ has no zero row.

Example: Consider the submarine equations of motion

$$\begin{matrix} \mathbf{2} & \mathbf{3} & \mathbf{2} & & & & \mathbf{3} \mathbf{2} & \mathbf{3} & \mathbf{2} & \mathbf{0} & \mathbf{3} \\ \mathbf{4} \mu & \mathbf{3} \mathbf{z} & \mathbf{4} & 0 & 0 & 1 & \mathbf{3} \mathbf{4} \mu & \mathbf{3} \mathbf{z} & \mathbf{4} & 0 & \mathbf{3} \mathbf{z} \\ \mathbf{q} & & a_{13} Z_{GB} & a_{11} U & a_{12} U & & \mathbf{q} & & b_1 U^2 & \mathbf{z} \pm & \mathbf{z} \pm \\ & & a_{23} Z_{GB} & a_{21} U & a_{22} U & & & & b_2 U^2 & & \end{matrix} ;$$

and substituting the values for the coefficients

$$\begin{matrix} \mathbf{2} & \mathbf{3} & \mathbf{2} & & & & \mathbf{3} \mathbf{2} & \mathbf{3} & \mathbf{2} & \mathbf{0} & \mathbf{3} \\ \mathbf{4} \mu & \mathbf{3} \mathbf{z} & \mathbf{4} & 0 & 0 & 1 & \mathbf{3} \mathbf{4} \mu & \mathbf{3} \mathbf{z} & \mathbf{4} & 0 & \mathbf{3} \mathbf{z} \\ \mathbf{q} & & 0:0135 & 0:3220 & 0:7102 & & \mathbf{q} & & 0:0322 & \mathbf{z} \pm & \mathbf{z} \pm \\ & & 0:0360 & 0:1260 & 0:7395 & & & & 0:0857 & & \end{matrix} ;$$

The controllability matrix is

$$C = \begin{matrix} \mathbf{2} & & & \mathbf{3} \\ & 0 & 0:0857 & 0:0674 \\ \mathbf{4} & 0:0322 & 0:0505 & 0:0653 \\ & 0:0857 & 0:0674 & 0:0404 \end{matrix} \mathbf{z} ;$$

which is full rank, 3. Therefore, the system is controllable and we can change any state μ , w , or q using the dive planes at will. Note, however, that some changes may be impractical or even impossible in practice; for example, even if the system is controllable it is not feasible to change the pitch angle to, say, 90 degrees! This would require an enormous dive plane strength which is not available in practice.

The definition for observability is

² A system is observable if any value of the state $x(t_0)$ can be exactly determined using a set of measurements over a finite period $t_0 < t < t_f$.

Observability depends on A and C only, and the test is: Compute the

$$\text{observability matrix } O = \begin{bmatrix} C \\ C A \\ C A^2 \\ \vdots \\ C A^{n-1} \end{bmatrix};$$

and the system is observable if and only if the rank of O is n. Roughly speaking, O shows how possible it is to reconstruct the state, x, of a system using a limited set of measurements, y. For a single output case C is 1 x n and O is a square matrix. The test is then that O be nonsingular

$$\det O \neq 0:$$

We can also test observability by transforming the system to the normal coordinate form (with distinct eigenvalues). The system will then be observable if C^o = C T has no zero column.

Example: Consider the previous submarine equations of motion, and assume that the only sensor aboard measures the pitch angle, μ. The measurement equation is

$$y = \begin{bmatrix} \mathbf{h} \\ 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ w \\ q \end{bmatrix} \mu$$

Using A and C, the observability matrix is

$$O = \begin{bmatrix} \mathbf{h} & 1 & 0 & 0 \\ \mathbf{h} A & 0 & 0 & 1 \\ \mathbf{h} A^2 & 0 & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ \mathbf{h} A^{n-1} & 0 & 0 & 1 \end{bmatrix};$$

j 0:0360 0:1260 j 0:7395

and this has rank 3. Therefore the system is observable: using μ measurements only we can get an estimate of both heave velocity w and pitch rate q (how to do this we will see later).

Now let's say we are interested in depth as well. The linear equation for the rate of change of submarine depth, z, is

$$\dot{z} = j U \mu + w$$

If we incorporate this as our fourth state equation, the new A matrix is now 4 x 4 and B is 4 x 1. Keeping the same measurement, μ only, we have

$$C = \begin{bmatrix} \mathbf{h} & 1 & 0 & 0 & 0 \end{bmatrix} \mu$$

If we compute the observability matrix O, its rank is 3 instead of 4. Therefore, the system is unobservable and one state (4 - 3 = 1) can not be estimated by looking at the angle μ only. This is, of course, z. If we assume that we have measurements of z only,

$$C = \begin{bmatrix} \mathbf{h} & 0 & 0 & 0 & 1 \end{bmatrix}$$

The new observability matrix has now full rank (4) which means that using a depth sensor only we should, in principle, be able to guess all the rest: μ , w , and q . The formalization of this "guess" constitutes the observer or estimator problem we discuss in Section 3.

2 CONTROLLER DESIGN

The control design problem can be stated as follows: Given the system

$$\dot{x} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} x + \begin{bmatrix} E \\ F \end{bmatrix} u ;$$

how do we find u such that x behaves nicely? We consider for now single input systems (u is scalar and B is a vector), the multiple input case is studied later. We are particularly interested in closed loop control, where u is a function of the state x . The case where u is an explicit function of time only and not x is called open loop control and is studied under system dynamics. Since we are using the state x to determine the control effort $u(x)$ we call it feedback control.

2.1 Pole Placement

The simplest case of feedback control $u(x)$ is when u is linear in x ,

$$u = -Kx ;$$

where K is the feedback gain vector to be determined. Substituting $u = -Kx$ into $\dot{x} = Ax + Bu$ we get

$$\begin{aligned} \dot{x} &= (A - BK)x ; \quad \text{or} \\ \dot{x} &= (A - BK)x ; \end{aligned}$$

The actual characteristic equation of this closed loop system is given by

$$\det [A - BK - sI] = 0 ;$$

We can now pick K such that the actual characteristic equation assumes any desired set of eigenvalues. If we choose the desired locations of the closed loop poles at $s = s_i$ for $i = 1, \dots, n$, the desired characteristic equation is

$$(s - s_1)(s - s_2) \dots (s - s_n) = 0 ;$$

The required values of K are obtained then by matching coefficients in the two polynomials of the actual and desired characteristic equations.

Consider the example:

$$A = \begin{bmatrix} 1 & 5 \\ 5 & 1 \end{bmatrix}; \quad B = \begin{bmatrix} 1 \\ 0 \end{bmatrix};$$

The open loop eigenvalues are

$$\det[sI - A] = \begin{vmatrix} s-1 & -5 \\ -5 & s-1 \end{vmatrix} = 0 \Rightarrow (s-1)^2 - 5 = 0 \Rightarrow (s-6)(s+4) = 0;$$

so we have an unstable system with no control. If the pair $(A; B)$ is controllable we are guaranteed that we can pick the elements of K to produce an arbitrary characteristic equation. In this case we have

$$AB = \begin{bmatrix} 1 \\ 5 \end{bmatrix}; \quad C = \begin{bmatrix} 1 & 1 \\ 0 & 5 \end{bmatrix}; \quad \det C = 5 \neq 0;$$

so the system is controllable. Now suppose we want closed loop eigenvalues at $-10 \pm 10i$ so that we get a damping ratio $\zeta = 0.707$. The desired closed loop characteristic equation is

$$(s + 10 - 10i)(s + 10 + 10i) = s^2 + 20s + 200 = 0;$$

Form the matrix

$$A - BK = \begin{bmatrix} 1 & 5 \\ 5 & 1 \end{bmatrix} - \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} k_1 & k_2 \end{bmatrix} = \begin{bmatrix} 1-k_1 & 5-k_2 \\ 5 & 1 \end{bmatrix};$$

and the actual closed loop characteristic equation is

$$\det[A - BK - sI] = \begin{vmatrix} 1-k_1-s & 5-k_2 \\ 5 & 1-s \end{vmatrix} = 0; \quad \text{or}$$

$$(1-k_1-s)(1-s) + 5(5-k_2) = 0; \quad \text{or}$$

$$s^2 + (k_1-2)s + (k_1-5k_2+24) = 0;$$

requiring

$$\begin{aligned} k_1 - 2 &= -20; \\ k_1 - 5k_2 + 24 &= 200; \end{aligned}$$

Solving this we get

$$\begin{aligned} k_1 &= -18; \\ k_2 &= \frac{246}{5}; \end{aligned}$$

and the control law is

$$u = -k_1 x_1 - k_2 x_2 = 18x_1 + \frac{246}{5}x_2;$$

Note that these gains may be impossible or impractical to build for this system. This would require some compromise in the specification which led to the desired closed loop eigenvalues. In general, the above approach yields a system of n linear equations to be solved for the elements of K provided $(A; B)$ is controllable. This method is known as pole placement.

Example: Consider the submarine equations

$$\begin{aligned} \mu &= q; \\ \underline{w} &= a_{13}z_{GB}\mu + a_{11}Uw + a_{12}Uq + b_1U^2\pm; \\ \underline{q} &= a_{23}z_{GB}\mu + a_{21}Uw + a_{22}Uq + b_2U^2\pm \end{aligned}$$

Let the control law be

$$\pm = -k_1\mu - k_2w - k_3q;$$

Substituting into the equations we get the closed loop system

$$\begin{aligned} \mu &= q; \\ \underline{w} &= (a_{13}z_{GB} - b_1U^2k_1)\mu + (a_{11}U - b_1U^2k_2)w + (a_{12}U - b_1U^2k_3)q; \\ \underline{q} &= (a_{23}z_{GB} - b_2U^2k_1)\mu + (a_{21}U - b_2U^2k_2)w + (a_{22}U - b_2U^2k_3)q; \end{aligned}$$

or, in matrix form,

$$\begin{bmatrix} \mu \\ \underline{w} \\ \underline{q} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 \\ a_{13}z_{GB} - b_1U^2k_1 & a_{11}U - b_1U^2k_2 & a_{12}U - b_1U^2k_3 \\ a_{23}z_{GB} - b_2U^2k_1 & a_{21}U - b_2U^2k_2 & a_{22}U - b_2U^2k_3 \end{bmatrix} \begin{bmatrix} \mu \\ \underline{w} \\ \underline{q} \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

The characteristic equation of the closed loop system is

$$\det \begin{bmatrix} s & 0 & 1 \\ a_{13}z_{GB} - b_1U^2k_1 & a_{11}U - b_1U^2k_2 - s & a_{12}U - b_1U^2k_3 \\ a_{23}z_{GB} - b_2U^2k_1 & a_{21}U - b_2U^2k_2 & a_{22}U - b_2U^2k_3 - s \end{bmatrix} = 0;$$

and after some algebra this reduces to

$$\begin{aligned} s^3 + (D_1^0 + A_2k_2 + A_3k_3)s^2 + (B_1k_1 - B_2k_2 - B_3k_3 + D_2^0)s \\ + (C_1k_1 - C_2k_2 - D_3^0) = 0; \end{aligned}$$

where we have denoted

$$\begin{aligned} A_2 &= b_1U^2; \quad A_3 = -B_1 = -b_2U^2; \\ B_2 &= (b_1a_{22} - b_2a_{12})U^3; \quad B_3 = C_1 = (b_2a_{11} - b_1a_{21})U^3; \\ C_2 &= (a_{23}b_1 - a_{13}b_2)U^2z_{GB}; \quad D_1^0 = (a_{11} + a_{22})U; \\ D_2^0 &= a_{23}z_{GB} + (a_{12}a_{21} - a_{11}a_{22})U^2; \quad D_3^0 = (a_{13}a_{21} - a_{11}a_{23})z_{GB}U; \end{aligned}$$

Now assume that we wish to place the closed loop poles at $-p_1, -p_2, -p_3$. This means that the desired characteristic equation is

$$\begin{aligned} (s + p_1)(s + p_2)(s + p_3) = 0; \quad \text{or} \\ s^3 + \alpha_1s^2 + \alpha_2s + \alpha_3 = 0; \end{aligned}$$

with

$$\begin{aligned} \textcircled{1} &= p_1 + p_2 + p_3 ; \\ \textcircled{2} &= p_1 p_2 + p_2 p_3 + p_3 p_1 ; \\ \textcircled{3} &= p_1 p_2 p_3 ; \end{aligned}$$

Then, the control gains can be computed by equating coefficients of the actual and the desired characteristic equations

$$\begin{aligned} A_2 k_2 + A_3 k_3 &= \textcircled{1} - D_1^0 ; \\ B_1 k_1 + B_2 k_2 + B_3 k_3 &= \textcircled{2} + D_2^0 ; \\ C_1 k_1 + C_2 k_2 &= \textcircled{3} + D_3^0 ; \end{aligned}$$

This method of equating coefficients is feasible only for small systems and it always produces a linear system in the unknown gains k_i .

The above approach can be simplified if the system is written in its control canonical form

$$\dot{\mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{B} u ; \quad y = \mathbf{C} \mathbf{x} ;$$

and we are seeking a control law of the form

$$u = -\mathbf{K} \mathbf{x} ;$$

As an example say the open loop characteristic equation is

$$s^4 + a_3 s^3 + a_2 s^2 + a_1 s + a_0 = 0 ;$$

and the state space form of the system is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -a_0 & -a_1 & -a_2 & -a_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} u ;$$

and

$$y = \begin{bmatrix} b_0 & b_1 & b_2 & b_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} ;$$

with the control law

$$u = -\begin{bmatrix} k_1 & k_2 & k_3 & k_4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} ;$$

The transfer function is

$$\frac{Y(s)}{U(s)} = \frac{b_3 s^3 + b_2 s^2 + b_1 s + b_0}{s^4 + a_3 s^3 + a_2 s^2 + a_1 s + a_0} ;$$

If the system is not in the control canonical form we have to transform it. Suppose that the original state x is transformed into x^0 through the transformation

$$x^0 = T x ;$$

and

$$\dot{x} = A x + B u ;$$

becomes

$$\dot{x}^0 = T A T^{-1} x^0 + T B u ;$$

For the transformed system, which is in the control canonical form,

$$u = - \sum_i K_i x_i^0 ;$$

where

$$K_i^0 = - \sum_j a_{j+}^0 \delta_{ij} = - \sum_j a_{j+}^0 \delta_{ij} ;$$

since the characteristic equation is invariant under a change of state variables. The control law is

$$\begin{aligned} u &= - \sum_i K_i x_i^0 ; \\ &= - \sum_i K_i T x_i ; \\ &= - \sum_i K_i x_i ; \end{aligned}$$

where

$$K_i = \sum_{j=1}^n K_{ij}^0 T_{ij} ;$$

is the gain in the original system. This can also be written as

$$K_i^T = T_{transpose}^T \left(\sum_{j=1}^n a_{j+}^0 \delta_{ij} \right) ;$$

We only need to find the transformation matrix T which will transform any system into its control canonical form. The desired matrix T is the product of two matrices

$$T = V U ;$$

where U is the inverse of the controllability matrix C

$$U = C^{-1} ;$$

Notice that if the system is uncontrollable, U does not exist. Matrix V is given by

$$V = W^{-1} ;$$

where

$$W = \begin{matrix} & \mathbf{2} & & & & & \mathbf{3} \\ & 1 & a_{n_i-1} & a_{n_i-2} & \zeta & \zeta & \zeta & a_1 \\ \mathbf{1} & 0 & 1 & a_{n_i-1} & \zeta & \zeta & \zeta & a_2 \\ & 0 & 0 & 1 & \zeta & \zeta & \zeta & a_3 \\ & \zeta & \zeta & \zeta & \zeta & & & \zeta \\ & \zeta & \zeta & \zeta & & \zeta & & \zeta \\ & \zeta & \zeta & \zeta & & \zeta & \zeta & \zeta \\ \mathbf{4} & 0 & 0 & 0 & \zeta & \zeta & \zeta & 1 \end{matrix};$$

the first row is formed by the coefficients of the characteristic polynomial of A

$$\det[A - sI] = s^n + a_{n_i-1}s^{n_i-1} + \dots + a_1s + a_0 = 0;$$

and the other rows are pushed left by one at a time. Therefore, the desired control law is

$$K^T = \mathbf{h} (CW)^T \mathbf{i}_{i-1} (j a + \otimes):$$

Now that we have a formula for the gains of a controllable single input system that will place the poles at any desired location, several questions arise:

1. If the closed loop poles can be placed anywhere, where should they be placed?
2. How can the technique be extended to multiple input systems?
3. What if not all states are available for feedback and we have to use output measurements only?
4. What do we do if we have external disturbances and we want to track a reference input?
5. How do we handle effects of sensor noise?
6. Can we optimize the performance of a control system?

The above questions are the subject of the remaining of these notes.

2.2 Pole Location Selection

For a second order system we may have some transient response specifications, such as rise time, percent overshoot, or settling time. These result in an allowable region in the s-plane from which we can easily get the desired locations of the poles. For higher order systems we can employ the concept of dominant roots, select two roots as dominant which means that we want to place the remaining roots more negative so that the transient response is not affected significantly. In selecting poles for a physical system we need to look at the physics; we can not specify poles that are too negative, for example. This would demand a very small time constant for the control system and the physical system may not be able to react that fast.

The control law $u = -Kx$ implies that for a given state x the larger the gain, the larger the control input. In practice, however, there are limits on u : actuator size and saturation. Occasional control saturation is not serious and may be even desirable; a system which never saturates is probably overdesigned.

Example: Control design by pole placement is very easy using MATLAB, the appropriate command is **place** which accepts as inputs the A , B matrices and a vector of the desired closed loop poles, and returns the gain vector K . For example, consider the submarine equations

$$\begin{matrix} \dot{\mu} \\ \dot{w} \\ \dot{q} \end{matrix} = \begin{matrix} 0 & 0 & 1 \\ 0.0135 & 0.3220 & 0.7102 \\ 0.0360 & 0.1260 & 0.7395 \end{matrix} \begin{matrix} \mu \\ w \\ q \end{matrix} + \begin{matrix} 0 \\ 0.0322 \\ 0.0857 \end{matrix} \begin{matrix} \mu \\ w \\ q \end{matrix}$$

A
B

Say we want to design a control law to stabilize the submarine to a level flight path at $\mu = 0$. We want to be able to return to level after an initial small disturbance in μ within the time it takes to travel one ship length, this is reasonable. Since the boat is about 17 feet long and it travels at 5 ft/sec, that time is about 3.5 seconds; so we want the control law to have a time constant of 3 seconds. This means we want to place the closed loop poles at approximately -0.3 . Using **place** we specify poles at -0.3 , -0.31 , -0.32 (**place** does not like poles that are exactly the same) and we find the gains in the control law

$$K = [-0.8451 \quad -1.4733 \quad 0.9807] :$$

Using a simulation program we plot the response starting from 30 degrees positive (bow up) pitch angle. We also set a limit in the dive plane angle between ± 0.4 radians. We can see from the results that initially the planes saturate at the upper limit and they come out as μ approaches zero. For comparison, we show the response with no control (planes fixed at zero). If we specify more negative poles, at -0.9 , -0.91 , -0.92 , the control law becomes

$$K = [-31.6147 \quad -1.2581 \quad 24.6634] :$$

Observe how unrealistically high these gains are: for a unit change in the pitch angle μ our controller demands 32 degrees of plane action! The response is also shown in the figure; there is more plane activity than in the previous case. However, since we hit the saturation limit, the response is not any faster and it overshoots the desired value. If we specify less negative poles at -0.1 , -0.11 , -0.12 , we end up with a control law

$$K = [0.3640 \quad -1.2581 \quad 8.0657] :$$

This is a very soft control law, it takes considerably longer for μ to reach zero and there is very limited plane activity.

From the above results, that are plotted in Figures 14 and 15, we can see that:

² Poles that are specified too negative will not necessarily result in faster response for a physical system; we may reach the hardware limitations of the system.

- ² Poles that are specified too negative will result in a high gain tight control law which will exhibit continuous control action; the system will over{respond to everything, including measurement noise.
- ² Poles that are specified not negative enough will result in soft response with a very quite control system that hardly works at all.
- ² Proper pole selection can be achieved by knowing the physics of the system we are trying to control, and by a trial{and{error simulation process.

The effect of control system gain on pole locations can be appreciated by considering the formula

$$K^T = \frac{\mathbf{h}}{(C\mathbf{W})^T \mathbf{i}_i^{-1}} (j\omega + \sigma) :$$

The gains are proportional to the amounts that the poles are to be moved: the less the poles are moved the smaller the gain matrix and therefore the control effort. It is also seen that the control system gains are inversely proportional to the controllability test matrix C. The less controllable the system, the larger the gains that are needed to make a change in the system poles.

Some broad guidelines for pole selection are:

- ² Select a bandwidth high enough to achieve desired speed of response.
- ² Keep the bandwidth low enough to avoid exciting unmodeled high frequency effects and undesired response to noise.
- ² Place the poles at approximately uniform distances from the origin for efficient use of the control effort.

We can also use standard characteristic polynomials such as minimizing the ITAE criterion, Bessel transfer functions, or Butterworth pole configurations. A typical sketch of the Butterworth poles is shown in Figure 16.

The closed loop poles tend to radiate out from the origin along the spokes of a wheel in the left half plane as given by the roots of

$$\frac{s^k}{\omega_0^k} = (j-1)^{k+1} ;$$

where k is the number of roots in the left half plane and ω_0 the natural frequency. In the absence of any other consideration, a Butterworth configuration is often suitable. Note, however, that as the order of the system k becomes high, one pair of poles comes very close to the imaginary axis. It might be desirable then to move these poles further into the left half plane.

Optimal control strategies can also be used to optimize some performance index. One common choice here is

$$\min J = \int_0^T (x^T Q x + u^T R u) dt ;$$

where

$$\begin{aligned} Q &= \text{weighting matrix of the error } x ; \\ R &= \text{weighting matrix of the control effort } u ; \end{aligned}$$

This is the Linear Quadratic Regulator problem which is studied later in these notes.

2.3 Multiple Input Systems

If the dynamic system under consideration

$$\dot{x} = Ax + Bu ;$$

has more than one inputs, that is B has more than one columns, then the gain matrix K in the control law

$$u = -Kx ;$$

has more than one rows. Since each row of K furnishes n adjustable gains, it is clear that in a controllable system there will be more gains available than needed to place all of the closed loop poles. If we have m inputs, then the equation

$$\det(j\omega I - A - BK) = \text{specified characteristic polynomial}$$

gives n equations with $n \times m$ unknowns. More than one solutions exist in general. This gives the designer more flexibility: it is possible to specify all the closed loop poles and still be able to satisfy other requirements. There are several possibilities here, some of them are briefly discussed below.

1. We can make one control proportional (or related) to the other. For example if we have a two input system

$$\dot{x} = Ax + \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} ;$$

we can choose

$$u_2 = \gamma u_1 ;$$

with γ some selected constant of proportionality, and the system becomes

$$\dot{x} = Ax + \begin{bmatrix} h_1 \\ h_1 + \gamma h_2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} ;$$

which is single input. The underlying physics should be the guidance for selection in this method. For example, say that our submarine is equipped with two inputs for depth control: independent stern and bow planes, call them \pm_s and \pm_b . If rapid depth change is what we want at a regular cruising speed then it makes sense to assume that $\pm_b = \gamma \pm_s$. This deflects the bow planes differentially than stern planes and produces maximum control authority through maximizing the vehicle pitching moment. If on the other hand, the vehicle is equipped with vertical stern and bow thrusters and is operating near hover, it is natural to command the same instead of opposite values for the two control inputs in order to achieve depth control.

2. Another possible method of selecting a particular structure for the gain matrix is to make each control variable depend on a different group of state variables that are physically more closely related to that control variable than to the other control variables. For example, suppose that our submarine is equipped with stern planes and sail planes at about amidships. Then it makes sense to use the stern planes to directly control pitch angle and the sail planes for direct depth control. Formally, what we are doing in this case is to specify not just the eigenvalues of the closed loop matrix but also (some of) its eigenvectors. This achieves a more precise shaping of the response.
3. Another possibility might be to set some of the gains to zero. For example, it is possible (sometimes) to place the closed loop poles at the desired locations with a gain matrix which has a column of zeros. This means that the state variable corresponding to that column is not needed in the generation of any of the control signals in the vector u , and hence there is no need to measure (or estimate) that state variable. This simplifies the resulting control system structure. If all the state variables, except those corresponding to columns of zeros in the gain matrix, are accessible for measurement then there is no need for an observer to estimate the state variables that cannot be measured. A very simple and robust control system is the result.

Hand calculation of the system of equations to be solved for the gains is possible for the multiple input case just like the single input. The only difference here is that unlike the single input where we always end up with a system of linear simultaneous equations in k_i , for multiple inputs it is possible to come up with a nonlinear system for k_{ij} .

3 OBSERVER DESIGN

So far we have developed the means to establish a control law; i.e., software which commands a certain action from the system actuators. What is needed is the state x . In reality, however, what is available to us from hardware is the output y through a set of sensors. In order to complete the picture, therefore, we need to estimate x given y .

3.1 State Estimators

Say we have the system

$$\dot{x} = Ax + Bu;$$

and we want to use a control

$$u = -Kx;$$

Suppose, however, that we only have the measurements (output)

$$y = Cx; \quad \begin{matrix} p \times 1 \\ p \times n \end{matrix}; \quad \begin{matrix} n \times 1 \\ n \times n \end{matrix}; \quad p < n;$$

instead of x . Note that if p were equal to n then we could use $x = C^{-1}y$ and our troubles would be over; the interesting case is when we have less sensors available than the number of states, $p < n$. It may be undesirable, expensive, or impossible to directly measure all of the states. What we can do is to dynamically use the p measurements to estimate all the states in x . If we denote the estimate of the state x as \hat{x} , the error in that estimate will be

$$e = x - \hat{x} :$$

error actual estimate

Then we could feed back this estimate \hat{x} in place of the actual state; i.e.,

$$u = -K\hat{x} :$$

What we need now is to construct a state estimator or observer. Consider feeding back the difference between the estimated and measured outputs and correcting the model continuously with this error signal

$$\dot{\hat{x}} = A\hat{x} + Bu + L(y - C\hat{x}) ;$$

where

- $A\hat{x} + Bu$: system model, \hat{x} should behave like x ;
- L : observer gain matrix, to be determined ;
- y : actual measurement ;
- $C\hat{x}$: measurement if x were \hat{x} :

In order to establish L we can consider the dynamics of the error in the estimate,

$$\begin{aligned} e &= x - \hat{x} \\ \dot{e} &= Ax + Bu - A\hat{x} - Bu - L(y - C\hat{x}) \\ \dot{e} &= A(x - \hat{x}) - L(Cx - C\hat{x}) \\ \dot{e} &= (A - LC)e : \end{aligned}$$

The error in the estimate will be determined by the eigenvalues of $[A - LC]$ which we can obtain from $\det[A - LC - sI] = 0$. If $(A;C)$ is observable, we can pick the elements of L to give the error arbitrary dynamics, similarly to the control design. We should choose the eigenvalues of $[A - LC]$ to be further to the left in the s {plane than the eigenvalues of $[A - BK]$. Then the error in the estimate will die quickly compared to the dynamics of the system.

The combined controller and observer equations are

$$\begin{aligned} \dot{x} &= Ax - BK\hat{x} ; \\ \dot{\hat{x}} &= LCx + (A - LC - BK)\hat{x} ; \\ y &= Cx ; \end{aligned}$$

or

$$\begin{bmatrix} \dot{x} \\ \dot{\hat{x}} \end{bmatrix} = \begin{bmatrix} A & -BK \\ LC & A - LC - BK \end{bmatrix} \begin{bmatrix} x \\ \hat{x} \end{bmatrix} ;$$

and

$$y = \begin{bmatrix} \mathbf{h} & \mathbf{0} \\ \mathbf{C} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{e} \end{bmatrix} :$$

In block diagram form this appears as shown in Figure 17.

If we use

$$u = -K \mathbf{x} = -K (\mathbf{x} + \mathbf{e}) ;$$

we get

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A} + \mathbf{B} \mathbf{K} & \mathbf{B} \mathbf{K} \\ \mathbf{0} & \mathbf{A} + \mathbf{L} \mathbf{C} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{e} \end{bmatrix} ;$$

which has the following characteristic equation

$$\det[\mathbf{A} + \mathbf{B} \mathbf{K} - s\mathbf{I}] \det[\mathbf{A} + \mathbf{L} \mathbf{C} - s\mathbf{I}] = 0 :$$

This indicates that the dynamics of the observer are completely independent of the dynamics (eigenvalues) of the controller. Thus, K and L can be designed separately.

3.2 Duality

Remember the controller design for $\dot{\mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u}$, $y = \mathbf{C} \mathbf{x}$ by placing the eigenvalues of $[\mathbf{A} \quad \mathbf{B} \mathbf{K}]$. For the observer design we want to place the eigenvalues of $[\mathbf{A} \quad \mathbf{L} \mathbf{C}]$. But the eigenvalues of $[\mathbf{A} \quad \mathbf{L} \mathbf{C}]$ are the same as the eigenvalues of $[\mathbf{A} \quad \mathbf{L} \mathbf{C}]^T$ and these are the same as the eigenvalues of $[\mathbf{A}^T \quad \mathbf{C}^T \mathbf{L}^T]$. Therefore, instead of designing an observer for the system $\dot{\mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u}$, $y = \mathbf{C} \mathbf{x}$ we can design a controller for $\dot{\mathbf{x}} = \mathbf{A}^T \mathbf{x} + \mathbf{C}^T \mathbf{u}$. This is the duality principle between controller and observer,

controller	$\tilde{\mathbf{A}}$	observer
\mathbf{A}	$\tilde{\mathbf{A}}$	\mathbf{A}^T
\mathbf{B}	$\tilde{\mathbf{A}}$	\mathbf{C}^T
\mathbf{C}	$\tilde{\mathbf{A}}$	\mathbf{B}^T

For any system

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u} ; \\ y &= \mathbf{C} \mathbf{x} ; \end{aligned}$$

its dual system is

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}^T \mathbf{x} + \mathbf{C}^T \mathbf{u} ; \\ y &= \mathbf{B}^T \mathbf{x} ; \end{aligned}$$

The controllability matrix of a system is the observability matrix of its dual and vice versa. If in the observer canonical form, starting from the output, all signal flows are reversed | summers are changed to nodes and nodes are changed to summers | we obtain the control canonical form.

3.3 Pole Placement for Single Output Systems

When there is only one output variable, the output equation is

$$y = \begin{matrix} \mathbf{h} \\ c_1 & c_2 & \dots & c_n \end{matrix} \begin{matrix} \mathbf{i} \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{matrix} :$$

Thus, C is a row vector

$$C = \begin{matrix} \mathbf{h} \\ c_1 & c_2 & \dots & c_n \end{matrix} \mathbf{i} ;$$

and the observer gain matrix L is a column vector

$$L = \begin{matrix} \mathbf{z} \\ l_1 \\ l_2 \\ \vdots \\ l_n \end{matrix} :$$

Now recall the expression we had for the controller gain matrix

$$K^T = \begin{matrix} \mathbf{h} \\ (CW)^T \end{matrix} \mathbf{i}_i^{-1} (j a + \textcircled{R}) :$$

By duality, the observer gain matrix must be

$$L = \begin{matrix} \mathbf{h} \\ (OW)^T \end{matrix} \mathbf{i}_i^{-1} (j a + \textcircled{R}) ;$$

where

- O = observability matrix ;
- a = coefficients of original characteristic equation ;
- \textcircled{R} = coefficients of desired characteristic equation ;

The presence of more than one outputs provides more flexibility; it is possible to place all the eigenvalues and do other things too. Or, alternatively, some of the observer gains can be set to zero to simplify the resulting observer structure.

3.4 Compensator Design

Recall that the eigenvalues of the controller were not affected by the eigenvalues of the observer, this allows us to design the controller and observer separately which is known as the separation principle. The combination is called a compensator,

$$(\text{controller}) + (\text{estimator}) = (\text{compensator}) :$$

For the system

$$\begin{aligned}\underline{x} &= A \underline{x} + B u ; \\ y &= C \underline{x} ;\end{aligned}$$

we have the controller

$$u = -K \underline{x} ;$$

the observer

$$\dot{\hat{\mathbf{x}}} = A \hat{\mathbf{x}} + B u + L (y - C \hat{\mathbf{x}}) ;$$

and, using the separation principle, we can write

$$u = -K \hat{\mathbf{x}} ;$$

The block diagram of the compensator is shown in Figure 17.

Using the above equations we get

$$\begin{aligned}\underline{x} &= A \underline{x} - B K \hat{\mathbf{x}} \\ &= A \underline{x} - B K (\underline{x} - \mathbf{e}) \\ &= (A - B K) \underline{x} + B K \mathbf{e} \\ &= A_c \underline{x} + B K \mathbf{e} ;\end{aligned}$$

and

$$\dot{\hat{\mathbf{x}}} = A \hat{\mathbf{x}} - B K \hat{\mathbf{x}} + L (C \underline{x} - C \hat{\mathbf{x}}) ;$$

Therefore,

$$\dot{\mathbf{e}} = \underline{\dot{x}} - \dot{\hat{\mathbf{x}}} = (A - B K - LC) \mathbf{e} = A_e \mathbf{e} ;$$

Taking Laplace transforms,

$$\begin{aligned}(sI - A_c) \underline{x}(s) &= B K \mathbf{e}(s) + \underline{x}(t_0) ; \\ (sI - A_e) \mathbf{e}(s) &= \mathbf{e}(t_0) \Rightarrow \mathbf{e}(s) = (sI - A_e)^{-1} \mathbf{e}(t_0) ;\end{aligned}$$

Therefore,

$$\begin{aligned}\underline{x}(s) &= (sI - A_c)^{-1} B K \mathbf{e}(s) + (sI - A_c)^{-1} \underline{x}(t_0) ; \\ &= (sI - A_c)^{-1} B K (sI - A_e)^{-1} \mathbf{e}(t_0) + (sI - A_c)^{-1} \underline{x}(t_0) ;\end{aligned}$$

and we can see that the transient response of the state is the sum of two part: one part due to the initial estimation error $\mathbf{e}(t_0)$, and one part due to the initial state $\underline{x}(t_0)$.

In order to obtain the transfer function of the compensator, we have

$$\hat{\mathbf{x}} = (A - B K - LC) \hat{\mathbf{x}} + L y ;$$

or

$$\hat{\mathbf{x}}(s) = (sI - A + B K + LC)^{-1} L y(s) ;$$

Then

$$u(s) = -K \mathbf{b}(s) = -K (sI - A + BK + LC)^{-1} L y(s) :$$

The transfer function of the compensator, $D(s)$, is defined between plant output and plant input by

$$u(s) = -D(s)y(s) ;$$

so

$$\begin{aligned} D(s) &= K (sI - A + BK + LC)^{-1} L \\ &= K (sI - A_c)^{-1} L ; \end{aligned}$$

where

$$A_c = A - BK - LC = A_c - LC = A_c - BK :$$

We can define the following:

- ² compensator poles = zeros of $sI - A_c$,
- ² open loop plant poles = zeros of $sI - A$,
- ² controller poles = zeros of $sI - A_c$,
- ² observer poles = zeros of $sI - A_o$.

All of the above are, in general, different. If A_c and A_o are chosen independently, it may even happen that A_c has roots in the right half s -plane, which means that even though the complete system is still stable, we can get an "unstable" compensator. This is not catastrophic, the main serious consequence of an unstable compensator is that the closed loop system will only be conditionally stable and, therefore, may not be very robust with respect to unmodeled dynamics and parameter variations.

In summary, the compensator design proceeds as follows:

1. Design a control law assuming that all states are available.
2. Design an observer to estimate the (missing) states.
3. Combine the full state control law with the observer to obtain the compensator design.

Example: Consider the submarine pitch angle control developed in the previous section. With poles at -0.3 , and if not all states μ, w, q are directly measurable, we have to use

$$\pm = -(0.8451p \pm 1.4733w + 0.9807q) :$$

Assume, however, that the only sensor we have is a rate gyro that measures the pitch rate q . We have to design an observer to estimate μ, w, q , using the q measurements. First, is this

possible? To do this we have to check the observability of the system. The output equation is

$$y = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ w \\ q \end{bmatrix};$$

and the observability matrix is

$$O = \begin{bmatrix} 0 & 0 & 1 \\ 0.0360 & 0.1260 & 0.7395 \\ 0.0283 & 0.1338 & 0.4214 \end{bmatrix};$$

which has rank 3; i.e., the system is observable. In order to design the observer gains we use the duality principle and we issue the MATLAB command **place** which we already used for the controller: here we use A^0 instead of A and C^0 instead of B (the prime in MATLAB signifies a transpose). The observer poles are selected, say at -0.6 , -0.61 , -0.62 ; these are twice as negative as the controller poles so the error in the estimate should die out faster than the system dynamics. The observer gains are

$$L = \begin{bmatrix} 21.9614 \\ 2.2636 \\ 0.7685 \end{bmatrix};$$

and the observer equations

$$\dot{\mathbf{b}} = A \mathbf{b} + B u + L (y - C \mathbf{b});$$

or, if we substitute the values for A , B , C , and L ,

$$\begin{aligned} \dot{\mu} &= -0.135\mu - 0.322w - 0.7102q; \\ \dot{w} &= 0.0135\mu + 0.322w + 0.7102q - 0.0322\mu - 2.2636(q - \mu); \\ \dot{q} &= 0.036\mu + 0.126w + 0.7395q - 0.0857\mu + 0.7685(q - \mu); \end{aligned}$$

The observer produces estimates of the states and these are used in the control law we established previously (with poles at -0.3),

$$u = 0.8451\hat{\mu} + 1.4733\hat{w} + 0.9807\hat{q};$$

The system is subjected to an initial disturbance $\mu = 30$ degrees, while for the observer we use $\hat{\mathbf{b}} = 0$ since the observer does not know the true value of μ . The results of the simulation are presented in Figure 18 where it can be seen that μ approaches zero in much the same way as for the complete state measurement case of the previous section. The estimate $\hat{\mu}$ approaches the true value of μ quickly.

If we were to reduce the absolute value of the observer poles, say to -0.1 , -0.11 , -0.12 we are faced with the following pathological situation: In order for the control law to return the system to its equilibrium, it needs an accurate estimate of the states as quickly as possible. Since the observer poles, however, are less negative than the controller poles this estimate

will be slow which means that it will take longer for the control law to stabilize the system to its equilibrium point. Indeed, in such a case the observer gains are

$$L = \begin{bmatrix} 0.8664 & 0 \\ 0.4817 & 0 \\ 0.7315 & 0 \end{bmatrix};$$

and the results of the simulation are shown in Figure 19. It can be seen that the response of the system is slow; even though the control poles were specified at ± 0.3 the response looks more like the ± 0.1 controller poles of the perfect state knowledge case of the previous section (why?).

It appears that we need to have the observer poles as negative as possible, compared to the closed loop control poles. A good rule of thumb practice is twice as negative. Beyond that we do not gain much and we run into problems with sensor noise, more about this later. For now, it is enough to recognize the fact that as the observer poles become more negative, the elements of L become larger in absolute value (verify this using MATLAB) and any kind of sensor noise that gets into our measurements will be magnified. There is a limit on how large the elements of L can be and this depends on the quality of our sensors. This is the optimal observer design or Kalman filter problem which we discuss later.

3.5 Reduced Order Observers

The previously developed observer is usually called a full order observer: its order is the same as that of the system. A full order observer estimates all the states in a system, regardless whether they are measured or not. This does not seem to be too bad, except imagine we have a system with ten states and we can measure eight of them; wouldn't it be better to estimate two instead of all ten states? The formalization of this procedure leads to the reduced order estimator.

Suppose we can measure some of the state variables contained in x . We partition the state vector x into two sets,

$$\begin{aligned} x_1 &: \text{variables that can be measured directly;} \\ x_2 &: \text{variables that cannot be measured directly;} \end{aligned}$$

The state equations are broken down to

$$\begin{aligned} \dot{x}_1 &= A_{11}x_1 + A_{12}x_2 + B_1u; \\ \dot{x}_2 &= A_{21}x_1 + A_{22}x_2 + B_2u; \end{aligned}$$

and the observation equation is

$$y = C_1x_1;$$

where C_1 is square and nonsingular matrix. The full order observer for the states is then

$$\begin{aligned} \dot{\hat{x}}_1 &= A_{11}\hat{x}_1 + A_{12}\hat{x}_2 + B_1u + L_1(y - C_1\hat{x}_1); \\ \dot{\hat{x}}_2 &= A_{21}\hat{x}_1 + A_{22}\hat{x}_2 + B_2u + L_2(y - C_1\hat{x}_1); \end{aligned}$$

But why take the trouble to implement the first observer equation for \mathbf{b}_1 when we can solve for x_1 directly?

$$\mathbf{b}_1 = x_1 = C_1^{-1}y :$$

In this case the observer for those states that cannot be measured directly becomes

$$\dot{\mathbf{b}}_2 = A_{21}C_1^{-1}y + A_{22}\mathbf{b}_2 + B_2u ;$$

which is a dynamic system of the same order as the number of state variables that cannot be measured directly. The dynamic behavior of this reduced order observer is governed by the eigenvalues of A_{22} , a matrix over which the designer has no control. Since there is no assurance that the eigenvalues of A_{22} are suitable, we need a more general system for the reconstruction of \mathbf{b}_2 . We take

$$\mathbf{b}_2 = Ly + z ;$$

where

$$\dot{z} = Fz + Gy + Hu :$$

Define the estimation error

$$e = \begin{bmatrix} x_1 - \mathbf{b}_1 \\ x_2 - \mathbf{b}_2 \end{bmatrix} = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} 0 \\ e_2 \end{bmatrix} ;$$

and we get

$$\begin{aligned} \dot{e}_2 &= \dot{x}_2 - \dot{\mathbf{b}}_2 \\ &= A_{21}x_1 + A_{22}x_2 + B_2u - Ly - \dot{z} \\ &= A_{21}x_1 + A_{22}x_2 + B_2u - LC_1x_1 - Fz - Gy - Hu \\ &= A_{21}x_1 + A_{22}x_2 + B_2u - LC_1(A_{11}x_1 + A_{12}x_2 + B_1u) \\ &\quad - F(\mathbf{b}_2 - Ly) - Gy - Hu : \end{aligned}$$

Since

$$\mathbf{b}_2 - Ly = x_2 - e_2 - Ly = x_2 - e_2 - LC_1x_1 ;$$

we get

$$\begin{aligned} \dot{e}_2 &= Fe_2 + (A_{21} - LC_1A_{11} - GC_1 + FLC_1)x_1 \\ &\quad + (A_{22} - LC_1A_{12} - F)x_2 + (B_2 - LC_1B_1 - H)u : \end{aligned}$$

In order for the error to be independent of x_1 , x_2 , and u , the matrices multiplying x_1 , x_2 , and u must vanish

$$\begin{aligned} F &= A_{22} - LC_1A_{12} ; \\ H &= B_2 - LC_1B_1 ; \\ G &= (A_{21} - LC_1A_{11})C_1^{-1} + FL : \end{aligned}$$

Then

$$\dot{e}_2 = Fe_2 ;$$

and for stability the eigenvalues of F must lie in the left half s -plane. Therefore, we see that the problem of reduced order observer is similar to the full order observer with $(A_{22} \quad LC_1A_{12})$ playing the role of $(A \quad LC)$. The block diagram schematic appears as shown in Figure 20.

Example: Consider the submarine problem, and assume that both the pitch angle μ and pitch rate q are available through measurements. What we need is to estimate the vertical translational (heave) velocity w . Let's design a reduced order observer to do the job. We start with our equations of motion and we re-write them so that the variables that are measurable go first

$$\begin{aligned} \dot{\mu} &= q; \\ \dot{q} &= a_{21}Uw + a_{22}Uq + a_{23}Z_{GB}\mu + b_2U^2\pm; \\ \dot{w} &= a_{11}Uw + a_{12}Uq + a_{13}Z_{GB}\mu + b_1U^2\pm; \end{aligned}$$

In matrix form we have

$$\begin{bmatrix} \dot{\mu} \\ \dot{q} \\ \dot{w} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ a_{23}Z_{GB} & a_{22}U & a_{21}U \\ a_{13}Z_{GB} & a_{12}U & a_{11}U \end{bmatrix} \begin{bmatrix} \mu \\ q \\ w \end{bmatrix} + \begin{bmatrix} 0 \\ b_2U^2 \\ b_1U^2 \end{bmatrix} \pm;$$

and the measurement equation is

$$y = \begin{bmatrix} \mu \\ q \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ q \\ w \end{bmatrix};$$

Therefore, the matrices are

$$\begin{aligned} x_1 &= \begin{bmatrix} \mu \\ q \end{bmatrix}; x_2 = w; \\ A_{11} &= \begin{bmatrix} 0 & 1 \\ a_{23}Z_{GB} & a_{22}U \end{bmatrix}; A_{12} = \begin{bmatrix} 0 \\ a_{21}U \end{bmatrix}; A_{21} = \begin{bmatrix} a_{13}Z_{GB} & a_{12}U \end{bmatrix}; A_{22} = \begin{bmatrix} a_{11}U \end{bmatrix}; \\ B_1 &= \begin{bmatrix} 0 \\ b_2U^2 \end{bmatrix}; B_2 = \begin{bmatrix} b_1U^2 \end{bmatrix}; C_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; L = \begin{bmatrix} \lambda_1 & \lambda_2 \end{bmatrix}; \end{aligned}$$

The reduced order observer equations are

$$\begin{aligned} \dot{z} &= \lambda_1\mu + \lambda_2q + z; \\ \dot{w} &= Fz + Gy + H\pm; \end{aligned}$$

Following the design procedure we have

$$F = a_{11}U - \lambda_1 - \lambda_2 = a_{11}U - \lambda_1 - \lambda_2 a_{21}U = p;$$

where p is the desired observer pole (F here is a scalar since there is only one state variable to be estimated). We see that λ_1 plays no role in determining F and, therefore, we can choose

$$\lambda_1 = 0;$$

for simplicity. The other observer gain \hat{p}_2 is computed from

$$\hat{p}_2 = \frac{a_{11}U + p}{a_{21}U}.$$

Then we get

$$\begin{aligned} H &= b_1U^2 + \mathbf{h}^T \mathbf{0} + \hat{p}_2 \mathbf{i}^T \mathbf{0} \\ &= b_1U^2 + \hat{p}_2 b_2U^2; \\ G &= A_{21} + LA_{11} + FL \\ &= a_{13}Z_{GB} + a_{12}U + \mathbf{h}^T \mathbf{0} + \hat{p}_2 \mathbf{i}^T \mathbf{0} \\ &= a_{13}Z_{GB} + a_{12}U + \hat{p}_2 a_{23}Z_{GB} + \hat{p}_2 a_{22}U + \mathbf{h}^T \mathbf{0} + \hat{p}_2 \mathbf{i}^T \mathbf{0} \\ &= a_{13}Z_{GB} + \hat{p}_2 a_{23}Z_{GB} + a_{12}U + \hat{p}_2 a_{22}U + \mathbf{h}^T \mathbf{0} + \hat{p}_2 \mathbf{i}^T \mathbf{0} \end{aligned}$$

The observer equations are then

$$\begin{aligned} \dot{\mathbf{z}} &= p\mathbf{z} + (a_{13}Z_{GB} + \hat{p}_2 a_{23}Z_{GB})\mu + (a_{12}U + \hat{p}_2 a_{22}U + p)\mathbf{q} \\ &\quad + (b_1U^2 + \hat{p}_2 b_2U^2)\pm; \\ \dot{\mathbf{w}} &= \hat{p}_2 \mathbf{q} + \mathbf{z}; \end{aligned}$$

Simulation results for control poles at $s = -0.3$ and observer pole at $s = -0.6$ are shown in Figure 21, in terms of w and \mathbf{w} versus time. In this simulation the initial conditions were changed to $\mu = \mathbf{q} = 0$, $w = 0.5$ ft/sec, and $\mathbf{w} = 0$. This was done to better show the convergence of \mathbf{w} to the true value w . The same remarks concerning selection of observer poles apply for the reduced order observer as for the full order observer design.

4 DISTURBANCES AND TRACKING SYSTEMS

The best way to start with the introduction of the reference input is via our submarine example:

Example: Once more, consider the submarine equations of motion

$$\begin{aligned} \dot{\mu} &= \mathbf{q}; \\ \dot{\mathbf{w}} &= a_{11}U\mathbf{w} + a_{12}U\mathbf{q} + a_{13}Z_{GB}\mu + b_1U^2\pm; \\ \dot{\mathbf{q}} &= a_{21}U\mathbf{w} + a_{22}U\mathbf{q} + a_{23}Z_{GB}\mu + b_2U^2\pm; \end{aligned}$$

We have a feedback control law which will guarantee stability, of the form

$$\pm = -k_1\mu - k_2\mathbf{w} - k_3\mathbf{q};$$

where the gains k_1, k_2, k_3 correspond, say, to the ± 0.3 poles. What if we wanted the boat to stabilize to, say, $\mu = \xi$ where $\xi \neq 0$? The first reaction might be to use

$$\pm = -k_1(\mu - \xi) - k_2w - k_3q :$$

To see if this is enough let's simulate the system with $\xi = 20$ degrees, and starting with zero initial conditions. The results are shown in Figure 22, in terms of $\mu = \xi$ versus t (solid curve) where it is clear that the system missed its final value, it stabilized but to the wrong angle. To see what went wrong, consider the above equations. At steady state all time derivatives go to zero, which means $\dot{\mu} = \dot{q} = 0$, $\dot{w} = 0$, and $\dot{q} = 0$. From the equations of motion this means that

$$\begin{aligned} a_{11}Uw + a_{13}Z_{GB}\mu + b_1U^2\pm &= 0 ; \\ a_{21}Uw + a_{23}Z_{GB}\mu + b_2U^2\pm &= 0 ; \end{aligned}$$

and if we use the steady state control law

$$\pm = -k_1\mu + k_1\xi - k_2w ;$$

we get

$$\begin{aligned} (a_{11}U - b_1U^2k_2)w + (a_{13}Z_{GB} - b_1U^2k_1)\mu &= -b_1U^2k_1\xi ; \\ (a_{21}U - b_2U^2k_2)w + (a_{23}Z_{GB} - b_2U^2k_1)\mu &= -b_2U^2k_1\xi ; \end{aligned}$$

This system of linear equations can be solved for the steady state values of w and μ . Using the gains that correspond to the ± 0.3 poles design, we find

$$\mu = 0.6679\xi ;$$

which agrees with the simulation results exactly. It seems, therefore, that the above control law can guarantee stability but it needs something extra to ensure steady state accuracy, in other words we need to add (or subtract) a little more plane action to bring μ up to ξ . We might be motivated then to use a control law of the form

$$\pm = -k_1(\mu - \xi) - k_2w - k_3q - k_0 ;$$

where the feedback gains k_1, k_2, k_3 remain the same as before, and k_0 is an unknown gain which is computed such that at steady state we get the desired result $\mu = \xi$. Therefore, at steady state we have

$$\begin{aligned} a_{11}Uw + b_1U^2\pm &= -a_{13}Z_{GB}\xi ; \\ a_{21}Uw + b_2U^2\pm &= -a_{23}Z_{GB}\xi ; \end{aligned}$$

The solution is

$$w \approx 0 ; \quad \text{and} \quad \pm = -0.4202\xi ;$$

Substituting into the steady state dive plane angle we get

$$\pm = -k_1(\mu - \xi) - k_2w - k_0 ;$$

or

$$k_0 = 0.4202 \ell :$$

This extra gain 0.4202 which multiplies the desired value ℓ is called a feedforward gain. By incorporating this in the previous control law, we achieve the desired steady state accuracy as shown in the results of Figure 22 with the dotted curve. It seems then that when a non-zero set point is commanded we can still use the same control law we developed before but augmented with an extra term to ensure that the commanded set point is achieved. The formalism of this result, along with the disturbance rejection, occupies the rest of this section.

4.1 Feedforward Control

So far we have considered the design of regulators in which the performance objective has been to achieve a specified closed loop dynamic behavior (i.e., pole locations) of the system in response to arbitrary initial disturbances. A more general design objective is to control the system error not only for initial disturbances, but also for persistent disturbances, and also to track reference inputs.

Say our system is

$$\dot{\underline{x}} = A \underline{x} + B \underline{u} + F \underline{x}_d ;$$

where \underline{x} is the $n \times 1$ state vector, \underline{u} is the $m \times 1$ control vector, and \underline{x}_d is a $d \times 1$ disturbance vector. To make things even more interesting suppose that we want to track a reference input \underline{x}_r in the presence of the disturbances \underline{x}_d , where the reference input has its own dynamics

$$\dot{\underline{x}}_r = A_r \underline{x}_r ;$$

We are concerned here with the error

$$\underline{e} = \underline{x} - \underline{x}_r ;$$

between the actual state \underline{x} and the reference state \underline{x}_r . What we need then is a differential equation in \underline{e} ,

$$\begin{aligned} \dot{\underline{e}} &= \dot{\underline{x}} - \dot{\underline{x}}_r \\ &= A(\underline{e} + \underline{x}_r) + B \underline{u} + F \underline{x}_d - A_r \underline{x}_r \\ &= A \underline{e} + (A - A_r) \underline{x}_r + F \underline{x}_d + B \underline{u} \\ &= A \underline{e} + B \underline{u} + E \underline{x}_0 ; \end{aligned}$$

where we have denoted

$$\underline{x}_0 = \begin{bmatrix} \underline{x}_r \\ \underline{x}_d \end{bmatrix} ;$$

a $(n + d) \times 1$ vector containing both the reference inputs and the disturbances, and

$$E = \begin{bmatrix} \mathbf{h} \\ A - A_r & F \end{bmatrix} \mathbf{i} ;$$

a $(n + d) \times n$ augmented matrix.

Consider a control law of the form

$$u = -K e - K_0 x_0 ;$$

Then the error dynamics becomes

$$\dot{e} = (A - B K) e + (B K_0 - E) x_0 ;$$

If it were possible it would be desirable to choose the gains K and K_0 to keep the system error e at zero. As we will see shortly though, this is not always possible. More reasonable performance objectives would be the following:

1. The closed loop system should be asymptotically stable.
2. A linear combination of the error state variables (rather than the entire state vector) is to be zero at steady state.

The first objective is met by placing the poles of $(A - B K)$ in the left half s -plane. At steady state we have

$$\dot{e} = 0 ;$$

which gives

$$(A - B K) e = (B K_0 - E) x_0 ;$$

and the steady state error is

$$e = (A - B K)^{-1} (B K_0 - E) x_0 ;$$

Now B is $n \times m$, K_0 is $m \times (n + d)$, and E is $(n + d) \times n$. We see, therefore, that only if we have as many inputs as there are states $n = m$ we can choose $K_0 = B^{-1} E$ to make e zero at steady state. In practice we have $m < n$ which means that we cannot make $e = 0$. Therefore, in general we can only require that some desired output y_d is zero at steady state,

$$y_d = C_d e = 0 ;$$

where C_d is a $m \times n$ matrix, so the number of inputs m is the same as the dimension of y_d . Then we can require

$$C_d (A - B K)^{-1} (B K_0 - E) x_0 = 0 ;$$

for all x_0 , or

$$C_d (A - B K)^{-1} (B K_0 - E) = 0 ;$$

or

$$C_d (A - B K)^{-1} B K_0 = C_d (A - B K)^{-1} E ;$$

Now we see that

$$\begin{array}{c} \begin{array}{c} C_d \\ m \times n \end{array} \begin{array}{c} \xrightarrow{\quad} \\ \underbrace{\quad} \\ n \times n \end{array} \begin{array}{c} (A - B K)^{-1} \\ n \times n \end{array} \begin{array}{c} \xrightarrow{\quad} \\ \underbrace{\quad} \\ n \times m \end{array} \begin{array}{c} B \\ n \times m \end{array} \end{array}$$

is $m \times m$ and can be inverted. Therefore, we can choose

$$K_0 = C_d(A - BK)^{-1} B^{-1} C_d(A - BK)^{-1} E;$$

and the steady state requirement $y_d = 0$ has been achieved.

Example: Let's illustrate the procedure with the submarine example. Suppose our objective is to keep constant depth z in the presence of two external disturbances f_1, f_2 (arising, say, from near surface effects at periscope depth). The linearized equations of motion, including the disturbance effects, are

$$\begin{aligned} \dot{\mu} &= q; \\ \dot{w} &= a_{11}Uw + a_{12}Uq + a_{13}Z_{GB}\mu + b_1U^2\pm + f_1; \\ \dot{q} &= a_{21}Uw + a_{22}Uq + a_{23}Z_{GB}\mu + b_2U^2\pm + f_2; \\ \dot{z} &= jU\mu + w; \end{aligned}$$

or, in matrix form,

$$\begin{bmatrix} \dot{\mu} \\ \dot{w} \\ \dot{q} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ a_{13}Z_{GB} & a_{11}U & a_{12}U & 0 \\ a_{23}Z_{GB} & a_{21}U & a_{22}U & 0 \\ jU & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ w \\ q \\ z \end{bmatrix} + \begin{bmatrix} 0 \\ b_1U^2 \\ b_2U^2 \\ 0 \end{bmatrix} \pm + \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} z_d$$

The objective is to keep depth $z = 0$ in the presence of f_1, f_2 . The first thing we have to do is to stabilize the system by placing the poles of $(A - BK)$. We do this by using a control law of the form

$$\pm = -j k_1 \mu - k_2 w - k_3 q - k_4 z;$$

Selection of poles at $-0.3, -0.31, -0.32, -0.33$ produces a stable system whose response in the absence of external disturbances is shown in Figure 23 (curve 1).

Use of the above feedback control law when $f_1 \neq 0, f_2 \neq 0$ produces stable response but with a nonzero steady state error, as expected; see curve 2 in the figure, with $f_1 = 0.005$ and $f_2 = -0.01$. In order to achieve the desired depth we introduce a general feedforward term in the control law

$$\pm = -j k_1 \mu - k_2 w - k_3 q - k_4 z - k_0;$$

where the feedback gains k_1, k_2, k_3, k_4 remain the same as before, and the feedforward gain k_0 will be determined such that $z = 0$ at steady state. At steady state we get $q = 0$ from the μ equation and $w = U\mu$ from $\dot{z} = 0$. The steady state control law becomes

$$\pm = -j k_1 \mu - k_2 U \mu - k_0;$$

where we have imposed the requirement $z = 0$. The w and q equations yield

$$\begin{aligned} a_{11}U^2\mu + a_{13}Z_{GB}\mu + b_1U^2\pm + f_1 &= 0; \\ a_{21}U^2\mu + a_{23}Z_{GB}\mu + b_2U^2\pm + f_2 &= 0; \end{aligned}$$

or, if we substitute in the expression for \pm ,

$$\begin{aligned} (a_{11}U^2 + a_{13}z_{GB} + b_1U^2k_1 + b_1U^3k_2)\mu + b_1U^2k_0 &= \pm f_1; \\ (a_{21}U^2 + a_{23}z_{GB} + b_2U^2k_1 + b_2U^3k_2)\mu + b_2U^2k_0 &= \pm f_2; \end{aligned}$$

Substituting in numerical values we can find

$$k_0 = 1:4312f_1 + 11:1353f_2;$$

and we can write then the complete control law as

$$\pm = 3:0673\mu + 1:1668w + 2:7562q + 0:0835z + 1:4312f_1 + 11:1353f_2;$$

where the feedback gains correspond to the $\pm 0:3$ pole selection as we mentioned before. If we simulate the system using this control law we see that the response gets to its desired value in the presence of nonzero f_1 and f_2 (curve 3). We should comment here that from the above two equations which were used to compute k_0 we can see that, in general, we get a nonzero pitch angle μ at steady state. This, similar to the set{and}{drift in currents, demonstrates that in the presence of disturbances it is in general impossible to keep all the state variables of a system to their desirable values.

We can get the same result by applying the general formula derived in this section. We have

$$\begin{aligned} \pm &= \pm K e + K_0 x_0 \\ &= \pm k_1(\mu + \mu_r) + k_2(w + w_r) + k_3(q + q_r) + k_4(z + z_r) \\ &\quad + k_{01}\mu_r + k_{02}w_r + k_{03}q_r + k_{04}z_r + k_{05}f_1 + k_{06}f_2; \end{aligned}$$

where the subscript r indicates the reference input states, which are zero in our case. The general equation for K_0 is

$$K_0 = \begin{matrix} \mathbf{h} \\ \mathbf{C}_d \end{matrix} (\mathbf{A} + \mathbf{B} \mathbf{K})^{-1} \mathbf{B}^{-1} \begin{matrix} \mathbf{i} \\ \mathbf{C}_d \end{matrix} (\mathbf{A} + \mathbf{B} \mathbf{K})^{-1} \mathbf{E};$$

The above matrices are (verify the calculations)

$$\begin{aligned} \mathbf{A} &= \begin{matrix} \mathbf{2} & & & & \mathbf{3} \\ & 0 & 0 & 1 & 0 \\ \mathbf{6} & 0:0135 & + & 0:3220 & + & 0:7102 & 0 \\ \mathbf{4} & + & 0:0360 & + & 0:1260 & + & 0:7395 & 0 \\ & + & 5 & 1 & 0 & 0 & & \end{matrix}; \quad \mathbf{B} = \begin{matrix} \mathbf{2} & & \mathbf{3} \\ & 0 & \\ \mathbf{6} & 0:0322 & \\ \mathbf{4} & + & 0:0857 \\ & & 0 & \end{matrix}; \\ \mathbf{E} &= \begin{matrix} \mathbf{2} & & & & \mathbf{3} \\ & 0 & 0 & 1 & 0 & 0 & 0 \\ \mathbf{6} & 0:0135 & + & 0:3220 & + & 0:7102 & 0 & 1 & 0 \\ \mathbf{4} & + & 0:0360 & + & 0:1260 & + & 0:7395 & 0 & 0 & 1 \\ & + & 5 & 1 & 0 & 0 & 0 & 0 & 0 & \end{matrix}; \quad \mathbf{C}_d = \begin{matrix} \mathbf{h} \\ 0 & 0 & 0 & 1 \end{matrix}; \\ \mathbf{K} &= \begin{matrix} \mathbf{h} \\ k_1 & k_2 & k_3 & k_4 \end{matrix} = \begin{matrix} \mathbf{h} \\ + & 3:0673 & + & 1:1668 & + & 2:7562 & 0:0835 \end{matrix}; \end{aligned}$$

Using these we find

$$\begin{aligned} K_0 &= \begin{matrix} \mathbf{h} \\ k_{01} & k_{02} & k_{03} & k_{04} & k_{05} & k_{06} \end{matrix} \\ &= \begin{matrix} \mathbf{h} \\ + & 3:0673 & + & 1:1668 & + & 2:7562 & 0 & 1:4312 & + & 11:1353 \end{matrix}; \end{aligned}$$

and substituting into the expression for \pm we get

$$\pm = 3:0673\mu + 1:1668w + 2:7562q \quad ; \quad 0:0835(z \quad ; \quad z_r) \quad ; \quad 1:4312f_1 + 11:1353f_2 \quad ;$$

the same control law as before.

4.2 Disturbance Estimation

Recall that the previous procedure was given a system with reference input x_r and disturbance x_d ,

$$\begin{aligned} \underline{x} &= A x + B u + F x_d ; \\ \underline{x}_r &= A_r x_r ; \end{aligned}$$

we form the error

$$e = x \quad ; \quad x_r ;$$

and the equation for the error dynamics

$$\dot{e} = A e + B u + E x_0 ;$$

with

$$x_0 = \begin{bmatrix} x_r \\ x_d \end{bmatrix} ; \quad E = \begin{bmatrix} \mathbf{h} \\ A \quad ; \quad A_r \quad F \quad \mathbf{i} \end{bmatrix} ;$$

The control law was

$$u = \quad ; \quad K e \quad ; \quad K_0 x_0 ;$$

where K is computed from stability requirements by pole{placing $(A \quad ; \quad B K)$, and K_0 is computed from the steady state accuracy requirement

$$y_d = C_d e = 0 \quad \text{at steady state} ;$$

by computing

$$K_0 = \begin{bmatrix} \mathbf{h} \\ C_d(A \quad ; \quad B K) \end{bmatrix}^{-1} B^{-1} \begin{bmatrix} \mathbf{i} \\ C_d(A \quad ; \quad B K) \end{bmatrix}^{-1} E ;$$

The above process requires knowledge of x_0 , which contains both the reference input x_r and the disturbances x_d . If direct measurement of x_0 is not possible (usually we know what the reference input x_r is but we cannot measure the disturbance x_d), estimation of x_0 is necessary. In order to estimate x_0 we need to assume a "model" for the disturbance, $\dot{x}_d = A_d x_d$; i.e., whether the disturbances are fairly constant, oscillatory, and so on. The complete system is then

$$\begin{aligned} \underline{x} &= A x + B u + F x_d ; \\ \dot{x}_d &= A_d x_d ; \\ \underline{x}_r &= A_r x_r ; \end{aligned}$$

Define a new augmented state vector

$$\underline{x} = \begin{bmatrix} e \\ x_0 \end{bmatrix} :$$

The new system is then written as

$$\dot{\underline{x}} = \underline{A} \underline{c} \underline{x} + \underline{B} u ;$$

where

$$\underline{A} = \begin{bmatrix} A & E \\ 0 & A_0 \end{bmatrix} ; \quad A_0 = \begin{bmatrix} A_r & 0 \\ 0 & A_d \end{bmatrix} ; \quad \underline{B} = \begin{bmatrix} B \\ 0 \end{bmatrix} :$$

We assume that the observation (measurement) vector y depends on both the error e and the vector x_0 ,

$$y = C e + D x_0 = \underline{C} \underline{c} \underline{x} ; \quad \underline{C} = \begin{bmatrix} \mathbf{h} & \mathbf{i} \\ C & D \end{bmatrix} :$$

We can apply now the general observer equation to the new augmented system \underline{x} ,

$$\dot{\mathbf{b}} = \underline{A} \underline{c} \underline{x} + \underline{B} u + \underline{L} (y - \underline{C} \underline{c} \underline{x}) ;$$

where \underline{L} is computed by pole placement of $(\underline{A} - \underline{L} \underline{C})$ as before. This procedure will produce a full order estimator for the augmented system, assuming of course that the augmented system is observable. In the same way we can design a reduced order estimator for the augmented system to estimate those states and disturbances that are not directly measurable. The key for the above procedure is to treat the disturbances as extra states; although we cannot control a disturbance we can estimate it by observing its effects on the system.

Separating the above full order observer equation into equations for the system error estimate \mathbf{b} and the error in estimating x_0 , we get

$$\begin{aligned} \dot{\mathbf{b}} &= A \mathbf{b} + B u + E \mathbf{b}_0 + L_e (y - C \mathbf{b} - D \mathbf{b}_0) ; \\ \dot{\mathbf{b}}_0 &= A_0 \mathbf{b}_0 + L_0 (y - C \mathbf{b} - D \mathbf{b}_0) ; \end{aligned}$$

The block diagram is presented in Figure 24. We can see from this block diagram that if x_0 is constant ($A_0 = 0$) and $D = 0$, then there are integrators in parallel to the path through L_e . This means that in the determination of \mathbf{b} there exists a path proportional to the integral of the residual $r = y - C \mathbf{b}$ in addition to the path through L_e which is proportional to the residual itself. Because of this integral path it is possible for r to become zero without \mathbf{b}_0 going to zero. Therefore, we can produce a nonzero control signal u , even when the system error is zero. In classical control system design this is achieved by means of control action; here it is achieved automatically by using an observer to estimate the unmeasurable x_0 .

With the above estimates \mathbf{b} and \mathbf{b}_0 , the control law for the compensator is

$$u = -K \mathbf{b} - K_0 \mathbf{b}_0 :$$

We refer to this technique as the disturbance estimation and compensation method.

Example: Consider the control law of the previous example. In general, the disturbance forces f_1, f_2 are unknown, so we have to use

$$\dot{\mathbf{x}} = 3.0673\mathbf{p} + 1.1668\mathbf{w} + 2.7562\mathbf{q} + 0.0835\mathbf{b}_1 + 1.4312\mathbf{b}_2 + 11.1353\mathbf{f}_2 :$$

In order to estimate f_1 and f_2 we first have to assume their dynamics. This is based on fairly general physical considerations. In our case, since both f_1 and f_2 are assumed to model free surface suction effects we can assume them to be relatively constant; i.e., $\dot{f}_1 = \dot{f}_2 = 0$. The equations of motion then are

$$\begin{aligned} \dot{\mu} &= q ; \\ \dot{w} &= a_{11}Uw + a_{12}Uq + a_{13}Z_{GB}\mu + b_1U^2\pm + f_1 ; \\ \dot{q} &= a_{21}Uw + a_{22}Uq + a_{23}Z_{GB}\mu + b_2U^2\pm + f_2 ; \\ \dot{z} &= j U\mu + w ; \end{aligned}$$

together with

$$\begin{aligned} f_1 &= 0 ; \\ f_2 &= 0 ; \end{aligned}$$

In matrix form the augmented system becomes

$$\begin{array}{c} \begin{array}{c} \mu \\ w \\ q \\ z \\ f_1 \\ f_2 \end{array} \\ \underbrace{\begin{array}{c} \mu \\ w \\ q \\ z \\ f_1 \\ f_2 \end{array}}_{\mathbf{x}} \end{array} = \begin{array}{c} \begin{array}{ccccccc} 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ a_{13}Z_{GB} & a_{11}U & a_{12}U & 0 & 1 & 0 & 0 \\ a_{23}Z_{GB} & a_{21}U & a_{22}U & 0 & 0 & 1 & 0 \\ j U & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \\ \underbrace{\begin{array}{ccccccc} \mu & w & q & z & f_1 & f_2 \end{array}}_{\mathbf{z}} \end{array} \begin{array}{c} \begin{array}{c} \mu \\ w \\ q \\ z \\ f_1 \\ f_2 \end{array} \\ \underbrace{\begin{array}{c} \mu \\ w \\ q \\ z \\ f_1 \\ f_2 \end{array}}_{\mathbf{x}} \end{array} + \begin{array}{c} \begin{array}{c} 0 \\ b_1U^2 \\ b_2U^2 \\ 0 \\ 0 \\ 0 \end{array} \\ \underbrace{\begin{array}{c} 0 \\ b_1U^2 \\ b_2U^2 \\ 0 \\ 0 \\ 0 \end{array}}_{\mathbf{z}} \end{array} :$$

Let's assume that z, μ, q are measurable (remember from Section 1.7 that we have to measure z); can we estimate $w, f_1,$ and f_2 ? The measurement equation is

$$\begin{array}{c} \begin{array}{c} \mu \\ q \\ z \end{array} \\ \underbrace{\begin{array}{c} \mu \\ q \\ z \end{array}}_{\mathbf{y}} \end{array} = \begin{array}{c} \begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{array} \\ \underbrace{\begin{array}{ccccccc} \mu & w & q & z & f_1 & f_2 \end{array}}_{\mathbf{z}} \end{array} \begin{array}{c} \begin{array}{c} \mu \\ w \\ q \\ z \\ f_1 \\ f_2 \end{array} \\ \underbrace{\begin{array}{c} \mu \\ w \\ q \\ z \\ f_1 \\ f_2 \end{array}}_{\mathbf{x}} \end{array} :$$

Using MATLAB we can see that the system is observable (the rank of the $(A;C)$ observability matrix is 6), so we should be able to estimate all states. Selecting observer poles at $j 0.6, j 0.61, j 0.62, j 0.63, j 0.64,$ and $j 0.65,$ we can get the (full) observer matrix

$$L = [l_{ij}] = \begin{array}{c} \begin{array}{ccc} 0.6234 & 1.0000 & 0.0000 \\ 13.3973 & j 0.6743 & 0.6681 \\ 1.6255 & 0.5153 & 0.0378 \\ 9.4327 & 0.0377 & 1.5498 \\ 5.6251 & 0.0153 & 0.2424 \\ j 0.0990 & 0.3905 & j 0.0492 \end{array} \\ \underbrace{\begin{array}{ccc} \mu & w & q \\ z & f_1 & f_2 \end{array}}_{\mathbf{z}} \end{array} :$$

The observer equations are then

$$\dot{\mathbf{b}} = \mathbf{A} \mathbf{b} + \mathbf{B} u + \mathbf{L} (\mathbf{y} - \mathbf{C} \mathbf{b}) ;$$

or

$$\begin{aligned} \dot{p} &= p + \lambda_{11}(\mu - p) + \lambda_{12}(q - \phi) + \lambda_{13}(z - b) ; \\ \dot{v} &= a_{11}U v + a_{12}U p + a_{13}Z_{GB} p + f_1 + b_1 U^2 \pm + \lambda_{21}(\mu - p) + \lambda_{22}(q - \phi) + \lambda_{23}(z - b) ; \\ \dot{\phi} &= a_{21}U v + a_{22}U p + a_{23}Z_{GB} p + f_2 + b_2 U^2 \pm + \lambda_{31}(\mu - p) + \lambda_{32}(q - \phi) + \lambda_{33}(z - b) ; \\ \dot{b} &= j U p + v + \lambda_{41}(\mu - p) + \lambda_{42}(q - \phi) + \lambda_{43}(z - b) ; \\ \dot{f}_1 &= \lambda_{51}(\mu - p) + \lambda_{52}(q - \phi) + \lambda_{53}(z - b) ; \\ \dot{f}_2 &= \lambda_{61}(\mu - p) + \lambda_{62}(q - \phi) + \lambda_{63}(z - b) ; \end{aligned}$$

Simulation results in terms of z , $f_1=f_1$, and $f_2=f_2$ versus t are presented in Figures 25 and 26.

We can see that the response goes to zero, as it should. The initial condition for \mathbf{b} was the same as for z , this is fair since z is measurable. The initial conditions for f_1 and f_2 were both zero, we have no knowledge of free surface effect forces and moments, and we can see that they converge to the actual values of f_1 , f_2 quickly.

4.3 Integral Control

The disturbance estimation and compensation technique will work well if we can have a fairly good idea of what kind of disturbances will affect the system. In our submarine example it may not be very hard to guess some kind of external forces and moments, but this is not always so easy. In order to produce good performance with a nonzero set point (reference input) and steady disturbances we need to introduce some sort of integral control behavior. It should be pointed out that integral control is an alternative to the disturbance estimation and compensation technique of the previous section, in fact the two methods are very closely related. Both techniques achieve the same thing, zero steady state error, and both have their advantages and disadvantages.

A typical state variable feedback control law feeds back the coordinates and their derivatives. From Newton's law we obtain second order ordinary differential equations for our systems and we often use the positions and velocities as the states. The state variable feedback thus produces a proportional+derivative (PD) type of feedback. Suppose that we are primarily interested in some desired output

$$z = \mathbf{D} \mathbf{x} ;$$

where z is $m \times 1$. It is for this output z that we want to maintain a desirable value in the presence of disturbances. If the desired value of z is z_d one way to introduce integral control characteristics is to introduce new state variables; i.e., augment the state vector,

$$\underline{v} = \mathbf{D} \mathbf{x} + z_d ;$$

Feedback of v then will produce an integral of the error z \dot{z}_d .

More specifically in the $z_d = 0$ case,

$$\dot{x} = Ax + Bu :$$

The new state variable is

$$v = z = Dx ;$$

or

$$\dot{v} = -z dt :$$

The augmented system is

$$\begin{bmatrix} \dot{x} \\ \dot{v} \end{bmatrix} = \begin{bmatrix} A & 0 \\ D & 0 \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} + \begin{bmatrix} B \\ 0 \end{bmatrix} u :$$

The control law is obtained by pole{placement of this system

$$u = - \begin{bmatrix} h & i \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} ;$$

or

$$\begin{aligned} u &= - \begin{bmatrix} K_0 & K_I \end{bmatrix} \begin{bmatrix} x \\ v \end{bmatrix} \\ &= \underbrace{-K_0 x}_{\text{PD action}} - \underbrace{K_I \int z dt}_{\text{Integral action}} ; \end{aligned}$$

so this is a generalized PID {control.

Example: Consider the submarine equations of motion

$$\begin{aligned} \dot{\mu} &= q ; \\ \dot{w} &= a_{11}Uw + a_{12}Uq + a_{13}Z_{GB}\mu + b_1U^2z + f_1 ; \\ \dot{q} &= a_{21}Uw + a_{22}Uq + a_{23}Z_{GB}\mu + b_2U^2z + f_2 ; \\ \dot{z} &= -U\mu + w ; \end{aligned}$$

If we want to maintain depth z at its desired value $z = 0$ we introduce a new state equation

$$\dot{z}_I = z ;$$

where z_I denotes the integral of z . We can see that steady state accuracy ($z = 0$) is automatically ensured. The augmented system is now

$$\begin{bmatrix} \dot{\mu} \\ \dot{w} \\ \dot{q} \\ \dot{z} \\ \dot{z}_I \end{bmatrix} = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ a_{13}Z_{GB} & a_{11}U & a_{12}U & 0 & 0 \\ a_{23}Z_{GB} & a_{21}U & a_{22}U & 0 & 0 \\ -U & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ w \\ q \\ z \\ z_I \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \\ 0 \\ 0 \\ 0 \end{bmatrix} u :$$

We select the closed loop controller poles at $j 0:30$, $j 0:31$, $j 0:32$, $j 0:33$ | same as before | with the i -th pole corresponding to z_I at $j 0:10$. The reason for this is that we want the integrator to correct the error only at steady state, while we would like to maintain the same transient response. As a result, the integrator must be relatively slow compared to the other poles of the system. The control law is

$$u = 4:0647\mu + 1:0237w + 3:8698q + 0:1533z + 0:0084z_I :$$

Results are shown in Figure 27 for the same values of the disturbances f_1, f_2 as before. We can see that z approaches zero, as it should. The main advantage of the integral control technique is that the desired response will approach its commanded value regardless of the exact type of disturbances. Also, no disturbance estimation is necessary. The main disadvantage is that the integral control response tends to be oscillatory especially if no disturbances are acting. In contrast, the response using the disturbance estimation and compensation technique is, in general, much smoother.

5 LYAPUNOV STABILITY

The concept of stability according to Lyapunov has found many applications in control systems; in fact the whole theory of dynamical systems is based, to a great extent, on Lyapunov's methods.

5.1 Lyapunov Functions

Consider the nonlinear system

$$\dot{x} = f(x) :$$

Let an equilibrium point of the system be \bar{x} ,

$$f(\bar{x}) = 0 :$$

We say that \bar{x} is stable in the sense of Lyapunov if there exists a positive quantity ϵ such that for every $\delta = \delta(\epsilon)$ we have

$$|x(t_0) - \bar{x}| < \delta \Rightarrow |x(t) - \bar{x}| < \epsilon ;$$

for all $t > t_0$. We say that \bar{x} is asymptotically stable if it is stable and,

$$|x(t) - \bar{x}| \rightarrow 0 \text{ as } t \rightarrow \infty :$$

We call \bar{x} unstable if it is not stable.

The question, of course, is: How do we determine stability or instability of \bar{x} ? Lyapunov introduced two main methods:

The first is called Lyapunov's first or indirect method: we have already seen it as the linearization technique. Start with a nonlinear system

$$\dot{x} = f(x) :$$

Expand in Taylor series around \bar{x} (we also redefine $x = x - \bar{x}$),

$$\dot{x} = Ax + g(x) ;$$

where

$$A = \frac{\partial f}{\partial x} \Big|_{\bar{x}} ;$$

is the Jacobian matrix of $f(x)$ evaluated at \bar{x} , and $g(x)$ contains the higher order terms; i.e.,

$$\lim_{|x| \rightarrow 0} \frac{|g(x)|}{|x|} = 0 :$$

Then, the nonlinear system $\dot{x} = f(x)$ is asymptotically stable if and only if the linear system $\dot{x} = Ax$ is; i.e., if all eigenvalues of A have negative real parts. This method is very popular because it is easy to apply and it works well for most systems, all we need to do is to be able to evaluate partial derivatives. One disadvantage of the method is that if some eigenvalues of A are zero and the rest have negative real parts, then we cannot draw any conclusions on the nonlinear system, the equilibrium \bar{x} can be either stable or unstable. The major drawback of the method, however, is that since it involves linearization it is applied for situations when the initial conditions are "close" to the equilibrium \bar{x} . The method provides no indication as to how close is "close", this is something which may be extremely important in practical applications.

The second method is Lyapunov's second or direct method: this is a generalization of Lagrange's concept of stability of minimum potential energy. Consider the nonlinear system $\dot{x} = f(x)$. Suppose that there exists a function, called Lyapunov function, $V(x)$ with the following properties:

1. $V(\bar{x}) = 0$.
2. $V(x) > 0$, for $x \neq \bar{x}$.
3. $\dot{V}(x) < 0$ along trajectories of $\dot{x} = f(x)$.

Then, \bar{x} is asymptotically stable. We can see that the method hinges on the existence of a Lyapunov function, which is an energy-like function, zero at equilibrium, positive definite everywhere else, and continuously decreasing as we approach the equilibrium. It should be noted that the derivative $\dot{V}(x)$ is understood as the total differential along solution curves of $\dot{x} = f(x)$; i.e.,

$$\begin{aligned} \dot{V}(x) &= \frac{\partial V}{\partial x} \cdot \frac{dx}{dt} \\ &= \frac{\partial V}{\partial x} f(x) \\ &= \frac{\partial V}{\partial x_1} f_1 + \frac{\partial V}{\partial x_2} f_2 + \dots + \frac{\partial V}{\partial x_n} f_n : \end{aligned}$$

The method is very powerful and it has several advantages:

- 2 answers questions of stability of nonlinear systems,
- 2 can easily handle time varying systems $\dot{x} = f(x; t)$,
- 2 can determine asymptotic stability as well as plain stability,
- 2 can determine the region of asymptotic stability or the domain of attraction of an equilibrium.

As an example, consider an oscillator with a nonlinear spring:

$$\ddot{y} + 3y + y^3 = 0 :$$

If we were to linearize this system we would get $\ddot{y} + 3y = 0$, which has characteristic equation $s(s + 3) = 0$. The -3 characteristic root corresponds to the damping term but notice the existence of a 0 root from the lack of a linear term in the spring restoring force. The linearized version of the system cannot recognize the existence of a nonlinear spring term and it fails to produce a non-zero characteristic root related to the restoring force. To see if this nonlinear spring produces a stable or unstable system we have to resort to Lyapunov functions. The state space form of the system is

$$\begin{aligned} \dot{x}_1 &= x_2 ; \\ \dot{x}_2 &= -3x_2 - x_1^3 ; \end{aligned}$$

with equilibrium $\bar{x}_1 = \bar{x}_2 = 0$. Let's try for a Lyapunov function

$$V(x) = \frac{1}{2}x_2^2 + \frac{1}{4}x_1^4 :$$

We can see that $V(x) > 0$ for all x_1, x_2 . The time derivative of V is

$$\begin{aligned} \dot{V}(x) &= \frac{\partial V}{\partial x_1}x_1 + \frac{\partial V}{\partial x_2}x_2 \\ &= x_1^3x_2 + x_2(-3x_2 - x_1^3) \\ &= -3x_2^2 \\ &< 0 : \end{aligned}$$

It follows then that \bar{x} is asymptotically stable.

The main drawback of the method is that there is no systematic way of obtaining Lyapunov functions, this is more of an art than science. For simple second order systems (like the one above) a good selection for a Lyapunov function is the total energy of the system (kinetic plus potential energy). Also, it is always possible to find a Lyapunov function for a linear system in the form

$$V(x) = Ax :$$

Choose as Lyapunov function the quadratic form

$$V(x) = x^T P x ;$$

where P is a symmetric positive definite matrix. Then we have

$$\begin{aligned} \dot{V} &= \dot{x}^T P x + x^T P \dot{x} \\ &= (Ax)^T P x + x^T P Ax \\ &= x^T A^T P x + x^T P Ax \\ &= x^T (A^T P + P A)x \\ &= -\lambda x^T Q x ; \end{aligned}$$

where we have denoted

$$A^T P + P A = -\lambda Q ;$$

If the matrix Q is positive definite, then the system is asymptotically stable. Therefore, we could pick $Q = I$, the identity matrix, and solve

$$A^T P + P A = -\lambda I ;$$

for P and see if P is positive definite (we can do this by looking at the n principal minors of P | Sylvester's criterion). The equation

$$A^T P + P A = -\lambda Q ;$$

is called Lyapunov's matrix equation and its solution is easy through MATLAB by using the command **lyap**. Of course one could argue that having an equation to determine a Lyapunov function for linear systems is useless; after all for a linear system we can always look at the eigenvalues of A to determine stability or instability. This is true, the usefulness of Lyapunov's matrix equation for linear systems is that it can provide an initial estimate for a Lyapunov function for a nonlinear system in cases where this is done computationally. Furthermore, it can be used to show stability, as we will see in the next section, of the linear quadratic regulator design.

5.2 Examples

We present three examples here that demonstrate three important applications of Lyapunov's method, namely

1. How to assess the importance of nonlinear terms in stability or instability.
2. How to estimate the domain of attraction of an equilibrium point.
3. How to design a control law that guarantees global asymptotic stability; i.e., with infinitely large domain of attraction, for a nonlinear system.

All of the above problems are very difficult, in general, and we shouldn't think that we can easily generalize the relatively simple examples we present here.

As our first example, suppose we have the system

$$\begin{aligned} \dot{x}_1 &= -x_2 + ax_1x_2^2; \\ \dot{x}_2 &= x_1 - bx_1^2x_2; \end{aligned}$$

with $a \neq b$. To find the equilibrium of the system we have to solve

$$\begin{aligned} -x_2 + ax_1x_2^2 &= 0; \\ x_1 - bx_1^2x_2 &= 0. \end{aligned}$$

Multiplying the first equation by x_1 , the second by x_2 and adding we get

$$x_1^2x_2^2(a - b) = 0;$$

from which $x_1 = 0$ or $x_2 = 0$. If $x_1 = 0$ then we see from the first equation that $x_2 = 0$ as well, and similarly if we assume that $x_2 = 0$. Therefore, the unique equilibrium of the system is $x_1 = x_2 = 0$. The linearized system is

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

The characteristic equation is

$$\det \begin{pmatrix} s & -1 \\ 1 & s \end{pmatrix} = 0 \Rightarrow s^2 + 1 = 0 \Rightarrow s = \pm i;$$

Since the characteristic roots are purely imaginary, we cannot draw any conclusion on the stability of the nonlinear system. We have to resort to Lyapunov functions. Let's try for $V(x)$ the sum of the "kinetic" and "potential" energy of the linear system (this doesn't always work of course), we get

$$V(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2;$$

We see that $V(x) > 0$ for all x_1, x_2 . Then

$$\begin{aligned} \dot{V}(x) &= x_1(-x_2 + ax_1x_2^2) + x_2(x_1 - bx_1^2x_2) \\ &= -x_1x_2 + ax_1^2x_2^2 + x_1x_2 - bx_1^2x_2^2 \\ &= (a - b)x_1^2x_2^2. \end{aligned}$$

Therefore, we see that

- if $a < b$ \Rightarrow the system is asymptotically stable;
- if $a > b$ \Rightarrow the system is unstable;

a result which could not have been obtained by linearization.

As our **second example**, suppose we want to determine the stability of the origin (0;0) of the nonlinear system (show that this is the equilibrium of the system),

$$\begin{aligned}\dot{x}_1 &= -x_1 + x_2 + x_1(x_1^2 + x_2^2); \\ \dot{x}_2 &= -x_1 - x_2 + x_2(x_1^2 + x_2^2).\end{aligned}$$

The easiest way to show stability is by linearization. The linearized form of the system is

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} :$$

The characteristic equation of the system is

$$s^2 + 2s + 2 = 0 ;$$

and we can see that the system is stable, the roots of the characteristic equation have negative real parts. Now since this result is based on linearization, it says that if the initial condition is "close" to the equilibrium point (0;0) then the solution will tend to the equilibrium as $t \rightarrow \infty$. To find how close is "close" we need to get an estimate of the domain of attraction. We can do this by using Lyapunov theory. Let's try a Lyapunov function candidate

$$V(x) = \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 :$$

Form

$$\begin{aligned}V(x) &= x_1\dot{x}_1 + x_2\dot{x}_2 \\ &= x_1(-x_1 + x_2 + x_1^3 + x_1x_2^2) + x_2(-x_1 - x_2 + x_2x_1^2 + x_2^3) \\ &= -x_1^2 + x_1x_2 + x_1^4 + x_1^2x_2^2 - x_1x_2 - x_2^2 + x_2^2x_1^2 + x_2^4 \\ &= x_1^4 + x_2^4 + 2x_1^2x_2^2 - x_1^2 - x_2^2 \\ &= (x_1^2 + x_2^2)^2 - (x_1^2 + x_2^2) \\ &= (x_1^2 + x_2^2)(x_1^2 + x_2^2 - 1) :\end{aligned}$$

We can see, therefore, that stability is guaranteed if

$$V(x) < 0 \quad \text{or} \quad x_1^2 + x_2^2 < 1 ;$$

which means that the domain of attraction of the equilibrium is a circular disk of radius 1. As long as the initial conditions are inside this disk, it is guaranteed that the solution will end up at the stable equilibrium. In case where the initial conditions lie outside the disk then convergence is not guaranteed. It should be mentioned that the above disk is an estimate of the domain of attraction based on the particular Lyapunov function we selected. A different Lyapunov function could have produced a different estimate of the domain of attraction.

As our **third example**, consider the motion of a space vehicle about the principal axes of inertia. The Euler equations are

$$\begin{aligned}A\dot{\omega}_x - (B - C)\omega_y\omega_z &= T_x ; \\ B\dot{\omega}_y - (C - A)\omega_z\omega_x &= T_y ; \\ C\dot{\omega}_z - (A - B)\omega_x\omega_y &= T_z ;\end{aligned}$$

where A , B , and C denote the moments of inertia about the principal axes, $\dot{\theta}_x$, $\dot{\theta}_y$, and $\dot{\theta}_z$ denote the angular velocities about the principal axes; and T_x , T_y , T_z are the control torques. Assume that the space vehicle is tumbling in orbit. It is desired to stop the tumbling by applying control torques which are assumed to be

$$\begin{aligned} T_x &= k_1 A \dot{\theta}_x ; \\ T_y &= k_2 B \dot{\theta}_y ; \\ T_z &= k_3 C \dot{\theta}_z ; \end{aligned}$$

where k_1 , k_2 , k_3 are the feedback gains. The unique equilibrium of the system is $\dot{\theta}_x = \dot{\theta}_y = \dot{\theta}_z = 0$. If we substitute the equations for the control torques we get the closed loop system

$$\begin{aligned} \ddot{\theta}_x &= \frac{B}{A} \dot{\theta}_y \dot{\theta}_z + k_1 \dot{\theta}_x ; \\ \ddot{\theta}_y &= \frac{C}{B} \dot{\theta}_z \dot{\theta}_x + k_2 \dot{\theta}_y ; \\ \ddot{\theta}_z &= \frac{A}{C} \dot{\theta}_x \dot{\theta}_y + k_3 \dot{\theta}_z ; \end{aligned}$$

If we linearize the system around its equilibrium we have

$$\begin{bmatrix} \ddot{\theta}_x \\ \ddot{\theta}_y \\ \ddot{\theta}_z \end{bmatrix} = \begin{bmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{bmatrix} \begin{bmatrix} \dot{\theta}_x \\ \dot{\theta}_y \\ \dot{\theta}_z \end{bmatrix} :$$

We can see that the eigenvalues of the closed loop matrix are the same as the feedback gains k_1 , k_2 , k_3 . Therefore, for stability we want negative poles and, as a result, we select negative gains k_1 , k_2 , k_3 for the three control torques. So far we have used linear methods. What we are really interested though is the following: will the above gain selection guarantee globally stable operation of the system? In other words, will our control law be able to stop the vehicle tumbling for any set of initial conditions? To see this we have to resort to Lyapunov methods. Choose as our Lyapunov function

$$V(\dot{\theta}) = \frac{1}{2} A \dot{\theta}_x^2 + \frac{1}{2} B \dot{\theta}_y^2 + \frac{1}{2} C \dot{\theta}_z^2 ;$$

which is the total kinetic energy of the vehicle. We see that V is positive definite, and its time derivative is

$$\dot{V}(\dot{\theta}) = k_1 A \dot{\theta}_x^2 + k_2 B \dot{\theta}_y^2 + k_3 C \dot{\theta}_z^2 ;$$

which is always negative if the gains are selected negative. Therefore, the above gain selection guarantees stability of the nonlinear system regardless of the initial conditions.

5.3 Sliding Mode Control

As an application of Lyapunov method, consider a single input system linear in the control effort

$$\dot{x} = f(x) + g(x)u ;$$

where $f(x)$, $g(x)$ are, in general, nonlinear functions in x . We want to design u such that we guarantee stability of $x = 0$.

Choose the Lyapunov function

$$V(x) = \frac{1}{2} [\varphi(x)]^2 ;$$

where

$$\varphi(x) = s^T x ;$$

The scalar function $\varphi(x)$ can be viewed as a weighted sum of the errors in the states x . For stability, we want the time derivative of $V(x)$ to be negative,

$$\dot{V}(x) = \dot{\varphi} \varphi < 0 ;$$

which can be achieved if

$$\dot{\varphi} \varphi = -\gamma \varphi ;$$

which means that

$$\dot{\varphi} = -\gamma \text{sign}(\varphi) ;$$

where

$$\text{sign}(\varphi) = \begin{cases} 1 & \text{if } \varphi > 0 ; \\ -1 & \text{if } \varphi < 0 ; \end{cases}$$

Using $\varphi(x) = s^T x$, we get

$$\dot{\varphi} = s^T \dot{x} = s^T f(x) + s^T g(x)u = -\gamma \text{sign}(\varphi) ;$$

and solving for u we get the control law

$$u = -\gamma \frac{s^T g(x)}{g^T g(x)} \text{sign}(\varphi) - \frac{s^T f(x)}{g^T g(x)} ;$$

We can see that this control law is the sum of two terms. The first term is a nonlinear state feedback, and the second term is a switching control law. The term γ is an arbitrary positive quantity, we usually select it such that \dot{V} is negative even in the presence of modeling errors and disturbances.

The above control law guarantees stability of $\varphi(x) = 0$, or $s^T x = 0$. We need to find s such that stability of $x = 0$ is guaranteed. If $\varphi(x) = 0$, the system becomes

$$\dot{x} = f(x) - \frac{s^T f(x)}{s^T g(x)} g(x) ;$$

and

$$\dot{x} = f(x) - \frac{s^T f(x)}{s^T g(x)} g(x) ;$$

If we linearize this system,

$$A = \frac{\partial f}{\partial x} \bigg|_0 ; \quad b = g(0) ;$$

we get a linear system

$$\dot{\underline{x}} = A \underline{x} + b u :$$

Then, on $\frac{3}{4}(x) = 0$ we have

$$\begin{aligned} \underline{x} &= \frac{A \underline{x} + b(s^T b)^{-1} s^T A \underline{x}}{\mathbf{h}^T A \underline{x} + b(s^T b)^{-1} s^T A \underline{x}} \\ &= \frac{A \underline{x} + b(s^T b)^{-1} s^T A \underline{x}}{\mathbf{h}^T A \underline{x} + b(s^T b)^{-1} s^T A \underline{x}} : \end{aligned}$$

The closed loop dynamics matrix is

$$A_c = A + b \underbrace{(s^T b)^{-1} s^T A}_{k} = A + b k :$$

Then

$$k = (s^T b)^{-1} s^T A \Rightarrow s^T b k = s^T A \Rightarrow s^T A + s^T b k = 0 ;$$

or

$$s^T (A + b k) = 0 \Rightarrow (A + b k)^T s = 0 \Rightarrow A_c^T s = 0 \Rightarrow (A_c^T + 0 \cdot I) = 0 :$$

We see then that s is the eigenvector of A_c^T that corresponds to the zero eigenvalue. The design procedure, therefore, can be summarized as follows:

- ² Pole placement of $A + b k$, specify one eigenvalue to be zero and the rest negative. Find k and therefore, find $A_c = A + b k$.
- ² Find s from $A_c^T s = 0$. Set $\frac{3}{4} = s^T x$.
- ² Implement the control law

$$u = - \frac{\mathbf{h}^T s^T g(x) + \mathbf{i}^T s^T f(x)}{\mathbf{h}^T s^T g(x) + \mathbf{i}^T s^T f(x)} \text{sign}(\frac{3}{4}) ;$$

if we have a nonlinear system, or

$$u = - \frac{(s^T b)^{-1} s^T A x + (s^T b)^{-1} s^T f(x)}{(s^T b)^{-1} s^T A x + (s^T b)^{-1} s^T f(x)} \text{sign}(\frac{3}{4}) ;$$

if we have a linear system.

6 OPTIMAL CONTROL

So far we were concerned with control design where the objective was either to stabilize a system (the regulator problem), or to track a reference input (the servomechanism problem). We can do better than this though! In particular, what if we wanted to design the "best" controller, where the word "best" is understood with respect to some measure of merit or performance index? In classical control design we have already seen the use of integral performance criteria (such as ITAE) in order to obtain desirable characteristic equations for use in pole placement. Other criteria could lead to minimizing the travel time (minimal time control), fuel consumption (minimal fuel control), miss distance (optimal rendezvous), and so on. These requirements lead to the design of optimal controllers.

6.1 Optimal Control Problems

In general terms, the problem is to find a control law u for the system $\dot{x} = f(x; u)$ such that a certain index J is minimized. Therefore, the basic problem of optimal control is

$$\text{minimize } J = K(x_0; x_f) + \int_{t_0}^{t_f} L(x; u) dt;$$

under the constraint

$$\dot{x} = f(x; u);$$

K, L are specified functions, and

$$\begin{array}{l} x_0(t_0) : \text{initial state (time } t_0) \\ x_f(t_f) : \text{final state (time } t_f) \end{array} \quad \left. \begin{array}{l} \\ \\ \end{array} \right\} \begin{array}{l} \text{given} \\ \text{or} \\ \text{free} \end{array} :$$

This formulation is general enough to allow for several interesting cases, for instance,

$\begin{cases} K = 0, L = 1 \end{cases} \Rightarrow$ minimal time problem,

$\begin{cases} K = 0, L = |u| \end{cases} \Rightarrow$ minimal fuel problem,

and so on.

Specifically, we have the following problem statement:

1. System equations $\dot{x} = f(x; u; t)$ where $x \in \mathbb{R}^n$ is the state vector, and $u \in \mathbb{R}^m$ is the controls vector.
2. Boundary conditions on the starting time, t_0 , initial state $x_0 = x(t_0)$, final time t_f , and final state $x_f = x(t_f)$. These may or may not be given, therefore we can have a number of combinations fixed/free, free/free, free/fixed problems.
3. Performance index

$$J = K(x_f; t_f) + \int_{t_0}^{t_f} L(x(t); u(t); t) dt;$$

A few special cases for this are:

$\begin{cases} \\ \\ \end{cases} \Rightarrow$ The Mayer problem,

$$J = K(x_f; t_f);$$

$\begin{cases} \\ \\ \end{cases} \Rightarrow$ The Lagrange problem,

$$J = \int_{t_0}^{t_f} L(x(t); u(t); t) dt;$$

$\begin{cases} \\ \\ \end{cases} \Rightarrow$ The Bolza problem, both K and L are non-zero.

4. Constraints can be either on control; i.e., $|u_j| \leq 1$ (very common), or on the state; i.e., $G(x_f; t_f) = 0$ (target sets), $|x_i| \leq X_i$ (inequality constraints, very hard to handle in general). These constraints determine a set of admissible control histories, U , and a set of admissible state trajectories, X .

The general problem of optimal control can then be stated as:

Find $u(t) \in U$ which takes the system from x_0 at t_0 to x_f at t_f by $\dot{x} = f(x; u; t)$ in such a way as to minimize J while $x(t) \in X$.

6.2 Examples

Some examples of optimal control problems are:

1. Time Optimal Control:

Consider $J = \int_{t_0}^{t_f} dt$ where t_0 is fixed and t_f is free. We can have fixed end points or belonging on target sets. Usually, we also need constraints on u to make the problem well-posed. As a particular example consider $\dot{x} = u$, where $|u| \leq 1$. Say we start from initial conditions $x_0; \dot{x}_0$ both positive and we want to get to the origin $x_f = \dot{x}_f = 0$, as quickly as possible. We can see that since we initially have positive x and positive \dot{x} we must apply full negative control $u = -1$ in order to get negative \dot{x} (i.e., towards the origin) while x remains positive. Then at some instant we should switch to full positive control $u = +1$ to stop at $x = 0$ with zero speed. The precise instant of switching from $u = -1$ to $u = +1$ is, of course, not known for now. This is an example of a bang-bang control problem, which most time optimal control problems lead to.

2. Fuel Optimal Control:

A typical example is,

$$J = \int_{t_0}^{t_f} \sum_{i=1}^n |u_i| dt :$$

Typically, such problems lead to bang-bang controls and with t_f free, the problem may be ill-posed for certain initial conditions | i.e., if no restrictions on t_f are placed minimum fuel could mean coast to the destination with very small speed.

3. Minimum Integral Square Error:

Here,

$$J = \int_{t_0}^{t_f} x^T x dt \quad \text{or} \quad J = \int_{t_0}^{t_f} x^T Q x dt ;$$

where Q is a symmetric and positive definite matrix. Typically we need constraints on u to prevent it from becoming infinitely large. In the special case of linear state feedback, we get the familiar ISE criterion.

4. Minimum Energy Problems:

Here,

$$J = \int_{t_0}^{t_f} \mathbf{u}^T \mathbf{R} \mathbf{u} dt;$$

where \mathbf{R} is symmetric and positive definite.

5. Final Value Optimal Control:

Here, $J = K(\mathbf{x}_f; t_f)$, for example

$$J = \sum_{i=1}^n (\mathbf{x}_i(t_f) - \mathbf{x}_i^*)^2;$$

Combinations of the above are, of course, also possible examples.

6.3 Calculus of Variations

A real function of a real variable is a map between a real number to another real number. A map between a function to a real number is called a functional. The performance index J is an example of a functional. Minimization of a functional is the subject of a branch of mathematics, called calculus of variations. The simplest problem of the calculus of variations is,

$$\min J = \int_{t_0}^{t_f} L(\mathbf{x}; \dot{\mathbf{x}}; t) dt;$$

where \mathbf{x} is a scalar function, t_0 , $\mathbf{x}(t_0)$, t_f , $\mathbf{x}(t_f)$ are given, and all functions are smooth. It should be mentioned here that t in the above equation is not necessarily time (although in control problems it most likely is); t simply denotes the dependent variable. The function \mathbf{x} then which minimizes J satisfies the so-called Euler-Lagrange equations,

$$\frac{\partial L}{\partial \mathbf{x}} - \frac{d}{dt} \frac{\partial L}{\partial \dot{\mathbf{x}}} = 0;$$

together with the boundary conditions $\mathbf{x}(t_0) = \mathbf{x}_0$, $\mathbf{x}(t_f) = \mathbf{x}_f$.

The solutions to these equations are called the extremals. The equations are usually referred to as Euler's equations in calculus of variations textbooks and Lagrange's equations in dynamics, where L is called the Lagrangian and is the kinetic minus the potential energy of a conservative system. Again in dynamics, the fact that the Lagrangian L is a stationary value for J is called Hamilton's principle.

The Euler-Lagrange (E-L) equations are in general 2nd order nonlinear differential equations, which means that we need two boundary conditions $\mathbf{x}(t_0) = \mathbf{x}_0$ and $\mathbf{x}(t_f) = \mathbf{x}_f$ to solve them. Existence, however, is not guaranteed here. This is not a Cauchy initial value problem, it is called a two-point boundary value problem (more later) and can be rather difficult to solve numerically.

Some particular cases of E-L are:

1. Suppose that $L(x; \underline{x}; t)$ is independent of x (this is called an ignorable coordinate in dynamics). Then, $E\{L$ results in

$$\frac{d}{dt} \frac{\partial L}{\partial \underline{x}} = 0 \Rightarrow \frac{\partial L}{\partial \underline{x}} = \text{const:}$$

which is the principle of conservation of conjugate momentum in dynamics.

2. Suppose we have a time invariant system and $L(x; \underline{x}; t)$ is independent of t . Then,

$$\frac{\partial L}{\partial x} + \frac{d}{dt} \frac{\partial L}{\partial \underline{x}} = \frac{\partial L}{\partial x} + \frac{\partial^2 L}{\partial x \partial \underline{x}} \underline{x} + \frac{\partial^2 L}{\partial \underline{x}^2} \ddot{\underline{x}} = 0;$$

or

$$\underline{x} \frac{\partial L}{\partial x} + \frac{\partial^2 L}{\partial x \partial \underline{x}} \underline{x}^2 + \frac{\partial^2 L}{\partial \underline{x}^2} \ddot{\underline{x}} = \frac{d}{dt} \left(L + \underline{x} \frac{\partial L}{\partial \underline{x}} \right) = 0;$$

This of course means that

$$L + \underline{x} \frac{\partial L}{\partial \underline{x}} = \text{const:};$$

which is the conservation of Hamiltonian.

3. If $L(x; \underline{x}; t)$ is independent of \underline{x} , then $E\{L$ becomes simply $\frac{\partial L}{\partial x} = 0$.

6.4 Example: The Brachistochrone Problem

The brachistochrone problem is one of the oldest problems that in fact initiated efforts towards calculus of variations. It can be simply stated as follows: Given a point O in a vertical plane with coordinates $(t_0; x_0)$ and another point also in the same vertical plane with coordinates $(t_f; x_f)$ find the shape of a curve connecting the two points such that a frictionless mass can start at O with zero speed and slide down in minimal time. The geometry is shown in Figure 28. We should exercise caution here in that t is not time; x and t are the two spatial coordinates of the problem.

To formulate the problem we use the kinetic energy $\frac{mv^2}{2}$ and the potential energy $-mgx$. Conservation of energy requires that $mv^2 + 2mgx = 0$ from which $v = \sqrt{2gx}$. The elapsed time is,

$$d\tau = \frac{ds}{v} = \frac{\sqrt{dx^2 + dt^2}}{\sqrt{2gx}} = \sqrt{\frac{1+x^2}{2gx}} dt;$$

The total elapsed time to minimize is then given by,

$$T = \int_{t_0}^{t_f} \sqrt{\frac{1+x^2}{2gx}} dt;$$

Since the Lagrangian

$$L(x; \underline{x}; t) = \sqrt{\frac{1+x^2}{x}};$$

is independent of t , the Euler-Lagrange equations become

$$\begin{aligned} L(x, \dot{x}) = \text{const} & \Rightarrow \frac{d}{dt} \left(\frac{\partial L}{\partial \dot{x}} \right) - \frac{\partial L}{\partial x} = 0 \\ \frac{d}{dt} \left(\frac{2x\dot{x}}{1+x^2} \right) - \frac{2x}{1+x^2} & = 0 \\ \frac{2x\ddot{x} + 2\dot{x}^2}{1+x^2} - \frac{2x}{1+x^2} & = 0 \\ \ddot{x} + \frac{\dot{x}^2}{x} - \frac{1}{x} & = 0 \end{aligned}$$

If we let

$$x = C_1 \sin^2 \mu ;$$

we get

$$dx = 2C_1 \sin \mu \cos \mu d\mu ;$$

and

$$dt = 2C_1 \sin^2 \mu d\mu = C_1 (1 - \cos 2\mu) d\mu ;$$

Integrating,

$$t = C_1 \left(\mu - \frac{\sin 2\mu}{2} \right) + C_2 ;$$

Since $x(\mu = 0) = 0$ and $t(\mu = 0) = 0$ we get,

$$\begin{aligned} x &= \frac{C_1}{2} (1 - \cos 2\mu) ; \\ t &= \frac{C_1}{2} (2\mu - \sin 2\mu) ; \end{aligned}$$

Geometrically, these equations represent (parametrically) an arc of a cycloid generated by rotating a circle of radius $C_1/2$ by an angle 2μ . The two constants C_1 and μ can be determined by enforcing the remaining two boundary conditions,

$$x(\mu_f) = x_f \quad \text{and} \quad t(\mu_f) = t_f ;$$

Some comments on the brachistochrone are:

1. Every sub-arc of a brachistochrone with appropriate boundary velocities is by itself a brachistochrone. With regards to Figure 28, if $A \rightarrow B$ is a brachistochrone with $v_A = 0$ and $v_B = \sqrt{2gh_B}$, then the brachistochrone between points C and D with velocities $v_C = \sqrt{2gh_C}$ and $v_D = \sqrt{2gh_D}$ is precisely the arc $C \rightarrow D$. This is called the Principle of Optimality.
2. A brachistochrone remains optimal after time reversal.
3. The brachistochrone helps make "strange" results in optimal control look more plausible, see Figure 28 for a couple of possible examples.

6.5 Optimality Conditions

We can use calculus of variations to derive the optimal control. We seek a function of time $u(t)$ to minimize J subject to the state equations $\dot{x} = f(x; u)$. Ordinary calculus can be used to solve for a parameter to minimize a scalar. Calculus of variations is used to solve for a function to minimize a scalar J . This is similar to the previous E {L equations, except that here we need to satisfy the state equations as well. The approach is directly parallel to the Lagrange multiplier method for parameter optimization subject to a constraint.

The final result is as follows: In order to solve

$$\begin{aligned} \min J &= K(x_0; x_f) + \int_{t_0}^{t_f} L(x; u) dt; \\ \text{such that } \dot{x} &= f(x; u); \end{aligned}$$

we define the Hamiltonian

$$H(x; p; u) = p^T f(x; u) + L(x; u);$$

where x is the state vector, and p is an unknown vector (called the co{state vector). The necessary conditions for optimality are the following sets of equations:

1. The state equations,

$$\dot{x} = \frac{\partial H}{\partial p} = f(x; u);$$

2. The adjoint equations,

$$\dot{p} = - \frac{\partial H}{\partial x};$$

3. Maximization of Hamiltonian,

$$\frac{\partial H}{\partial u} = 0;$$

which is known as Pontryagin's maximum principle.

4. Boundary conditions,

$$\pm K + p^T \pm x + \int_{t_0}^{t_f} H \pm t = 0;$$

Solution of these formidable equations yields the optimal control law u . This is a very difficult task, and even when it is possible, usually the procedure yields an open loop control; i.e., u is obtained as a function of time rather than state. A special case where solution can be obtained in closed loop form is the Linear Quadratic Regulator (LQR) problem.

6.6 The Linear Quadratic Regulator

Suppose we have a linear system,

$$\dot{\mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u} ;$$

and a quadratic cost function,

$$J = \frac{1}{2} \mathbf{x}_f^T \mathbf{F} \mathbf{x}_f + \frac{1}{2} \int_{t_0}^{t_f} \mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u} \, dt ;$$

where \mathbf{x}_0 , t_0 , t_f are given (fixed) and \mathbf{x}_f is free to vary. This is the LQR problem: we seek a control law \mathbf{u} to minimize J . It should be emphasized that the above matrices \mathbf{A} , \mathbf{B} , \mathbf{Q} , \mathbf{R} are assumed, in general, to be functions of time. This is our first attempt, so far, to design a control law for a linear, time-varying system.

The weighting matrices \mathbf{F} , \mathbf{Q} , \mathbf{R} are symmetric and positive definite and are at the discretion of the designer. \mathbf{Q} is the state weighting matrix, \mathbf{R} penalizes the control effort, and \mathbf{F} penalizes the final state (or miss distance). Relatively small elements of \mathbf{Q} compared to \mathbf{R} will result in a control law which will tolerate errors in \mathbf{x} with low control effort. On the other hand, if \mathbf{Q} is made large compared to \mathbf{R} this will result in tight control; small errors in the state with considerable control effort. We can also use different values of the entries of \mathbf{Q} (or \mathbf{R}). For example, say the (2;2) element of \mathbf{Q} is large compared to the rest. This will result in improved control of the state x_2 at the expense of control accuracy of the other states and more control effort.

In order to solve the LQR problem we apply the general equations of optimal control. The Hamiltonian is

$$H(\mathbf{x}; \mathbf{p}; \mathbf{u}) = \mathbf{p}^T (\mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u}) + \frac{1}{2} (\mathbf{x}^T \mathbf{Q} \mathbf{x} + \mathbf{u}^T \mathbf{R} \mathbf{u}) ;$$

and the necessary conditions for optimality are

$$\begin{aligned} \dot{\mathbf{x}} &= \frac{\partial H}{\partial \mathbf{p}} = \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u} ; \\ \dot{\mathbf{p}} &= - \frac{\partial H}{\partial \mathbf{x}} = - \mathbf{A}^T \mathbf{p} - \mathbf{Q} \mathbf{x} ; \\ \frac{\partial H}{\partial \mathbf{u}} &= 0 \Rightarrow \mathbf{B}^T \mathbf{p} + \mathbf{R} \mathbf{u} = 0 \Rightarrow \mathbf{u} = - \mathbf{R}^{-1} \mathbf{B}^T \mathbf{p} ; \end{aligned}$$

The boundary conditions are

$$\mathbf{p}^T(t_f) + \mathbf{H}(t_f) + \frac{1}{2} \mathbf{x}_f^T \mathbf{F} \mathbf{x}_f = 0 ;$$

or

$$\mathbf{p}^T(t_f) + \mathbf{H}(t_f) + \mathbf{x}_f^T \mathbf{F} \mathbf{x}_f = 0 ;$$

Since \mathbf{x}_0 , t_0 , and t_f are fixed we have

$$\mathbf{x}(t_0) = \mathbf{x}(t_f) = 0 ;$$

and the boundary condition becomes

$$\begin{aligned} \mathbf{p}^T(t_f) \pm \mathbf{x}_f^T \mathbf{F} \pm \mathbf{x}_f &= 0 \\ \mathbf{p}^T(t_f) + \mathbf{x}_f^T \mathbf{F} \pm \mathbf{x}_f &= 0 : \end{aligned}$$

Since \mathbf{x}_f is free, its variation $\pm \mathbf{x}_f$ is arbitrary. Therefore, the quantity inside the square brackets must vanish, and this produces the desired boundary condition in the form

$$\mathbf{p}(t_f) = -\mathbf{F}^T \mathbf{x}(t_f) :$$

In summary, the problem we have to solve is

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{p} ; \\ \dot{\mathbf{p}} &= -\mathbf{Q} \mathbf{x} - \mathbf{A}^T \mathbf{p} ; \\ \mathbf{x}(t_0) &= \mathbf{x}_0 ; \\ \mathbf{p}(t_f) &= -\mathbf{F}^T \mathbf{x}(t_f) : \end{aligned}$$

Solution of these ordinary differential equations will provide $\mathbf{p}(t)$ and this will allow calculation of \mathbf{u} as a function of time from $\mathbf{u} = \mathbf{R}^{-1} \mathbf{B}^T \mathbf{p}(t)$. However, solving these equations is not as easy as it may seem. Notice that for a numerical integration of $\dot{\mathbf{x}}$ and $\dot{\mathbf{p}}$ we need to know the initial conditions at t_0 ; i.e., $\mathbf{x}(t_0)$ and $\mathbf{p}(t_0)$. But we know $\mathbf{p}(t_f) = -\mathbf{F}^T \mathbf{x}(t_f)$ instead of $\mathbf{p}(t_0)$. This is called a two-point boundary value problem with half of the boundary conditions at t_0 and the other half at t_f . Solution of two-point boundary value problems requires iterative (shooting) techniques: assume an initial condition $\mathbf{p}(t_0)$, integrate numerically the system and at the end check whether the condition $\mathbf{p}(t_f) = -\mathbf{F}^T \mathbf{x}(t_f)$ is satisfied, if it is not change the initial condition $\mathbf{p}(t_0)$ and iterate until convergence. To make things worse, even if we could easily solve this problem, still the optimal control \mathbf{u} would be open loop, $\mathbf{u}(t)$ instead of $\mathbf{u}(\mathbf{x})$.

Kalman's idea comes here to the rescue: Let

$$\mathbf{p}(t) = -\mathbf{S}(t) \mathbf{x}(t) ;$$

where $\mathbf{S}(t)$ is a symmetric positive definite matrix to be determined. Then we have

$$\begin{aligned} \dot{\mathbf{p}} &= -\dot{\mathbf{S}} \mathbf{x} - \mathbf{S} \dot{\mathbf{x}} \\ &= -\dot{\mathbf{S}} \mathbf{x} - \mathbf{S} (\mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{p}) ; \end{aligned}$$

or

$$\begin{aligned} \mathbf{Q} \mathbf{x} - \mathbf{A}^T \mathbf{p} &= -\dot{\mathbf{S}} \mathbf{x} - \mathbf{S} (\mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{p}) \\ \mathbf{Q} \mathbf{x} + \mathbf{A}^T \mathbf{S} \mathbf{x} &= -\dot{\mathbf{S}} \mathbf{x} - \mathbf{S} \mathbf{A} \mathbf{x} - \mathbf{S} \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{S} \mathbf{x} \\ -\dot{\mathbf{S}} \mathbf{x} &= \mathbf{A}^T \mathbf{S} + \mathbf{S} \mathbf{A} - \mathbf{S} \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{S} + \mathbf{Q} \mathbf{x} ; \end{aligned}$$

and since this must be true for all \mathbf{x} we get

$$-\dot{\mathbf{S}} = \mathbf{A}^T \mathbf{S} + \mathbf{S} \mathbf{A} - \mathbf{S} \mathbf{B} \mathbf{R}^{-1} \mathbf{B}^T \mathbf{S} + \mathbf{Q} ;$$

with

$$\dot{S}(t) = -F S(t)$$

This is called a Riccati matrix differential equation. Therefore, we can obtain $S(t)$ by backwards integration of the Riccati matrix differential equation, and then obtain the closed loop optimal control law by

$$u = -R^{-1}B^T S(t)x;$$

a linear state feedback with time varying gains.

For the case of constant A, B, Q, R matrices and $t_f \rightarrow \infty$, we have the steady state problem $\dot{S} = 0$. In this case the optimal closed loop control law is

$$u = -R^{-1}B^T S x;$$

where S is found by solving the algebraic Riccati equation (ARE) for the positive definite S ,

$$A^T S + S A - S B R^{-1} B^T S + Q = 0;$$

This is a nonlinear algebraic equation in the elements of S and it may admit multiple solutions, only one of them is positive definite though, and this is the one that we seek. See the `lqr` command for solution of the LQR problem using MATLAB.

Recall that previously we were using pole (eigenvalue) placement to produce arbitrary closed loop eigenvalues. Here we have a technique more suited for large, multivariable systems in which we choose the weighting matrices Q and R . The mathematics then yields a set of closed loop eigenvalues which are guaranteed to be stable (we will see why shortly) but over which we have no direct control. If the closed loop eigenvalues are not acceptable, it may be necessary to change the weighting matrices Q and R and iterate. If the errors in the state x_i are too large, it would be necessary to raise q_{ii} . If there is excessive use of control u_j , it would be necessary to raise r_{jj} . This would cause the state or control with the increased weighting in J to be reduced in the next design (iteration) at the expense of increased errors in the other states and/or increased usage of the other controls.

How do we know that the LQR design yields a stable system though? We can show stability by using Lyapunov's method. Choose

$$V(x) = x^T S x;$$

as a Lyapunov function candidate, where S is the positive definite solution of the Riccati equation. Since S is a positive definite matrix, $V(x) > 0$. Its time derivative is

$$\begin{aligned} \dot{V}(x) &= \dot{x}^T S x + x^T \dot{S} x + x^T S \dot{x} \\ &= (Ax + Bu)^T S x + x^T \dot{S} x + x^T S (Ax + Bu) \\ &= x^T (S - A^T S + S A - 2S B R^{-1} B^T S)x \\ &= x^T (S B R^{-1} B^T S - Q)x \\ &= -x^T S B R^{-1} B^T S x - x^T Q x : \end{aligned}$$

Let $z = R^{-1}B^T S x$ be some vector, then

$$V(x) = \int_0^{\infty} z^T R z + x^T Q x < 0 ;$$

since Q, R are positive definite matrices. Therefore, $V(x)$ is a Lyapunov function for the LQR design, and since

$$\begin{aligned} V(x) &> 0 \quad \text{and} \\ V(x) &< 0 ; \end{aligned}$$

the design will always yield a stable system (as long as the Riccati equation supplies the positive definite solution matrix S).

As an example, say we have $\dot{x} = 2x + u$, a scalar system. The open loop pole is $s = 2 = 0$ or $s = 2$, so it is unstable. We wish to control x near zero and minimize

$$J = \int_0^{\infty} (qx^2 + ru^2) dt ;$$

Suppose we want to use $q = 0.25$ and $r = 1$. Then the ARE is $2k + 2k + k^2 + 0.25 = 0$, or $k^2 + 4k + 0.25 = 0$. The positive root is $k = 4.06$ and the optimal control is

$$u = -1^{-1} \cdot 4.06x = -4.06x ;$$

The closed loop eigenvalue is $\det(2 - 4.06 - s) = 0$ or $s = -2.06$, and the closed loop response is $x(t) = x(t_0)e^{i 2.06t}$. If we wish to reduce the error in x faster at the expense of using more control we can raise q . If we redesign for $q = 4, r = 1$ we get $k = 4.83, u = -4.83x$, and $x(t) = x(t_0)e^{i 2.83t}$. If we wish to reduce the amount of control used at the expense of slower response, we can raise r . If we redesign for $q = 0.25$ and $r = 10$, we get $k = 40.06, u = -4x$, and $x(t) = x(t_0)e^{i 2t}$.

Example: Consider the submarine equations of motion

$$\begin{aligned} \dot{\mu} &= q ; \\ \dot{w} &= a_{11}Uw + a_{12}Uq + a_{13}Z_G B \mu + b_1 U^2 \pm ; \\ \dot{q} &= a_{21}Uw + a_{22}Uq + a_{23}Z_G B \mu + b_2 U^2 \pm ; \\ \dot{z} &= -j U \mu + w ; \end{aligned}$$

One common logic in selecting the weighting matrices Q and R in the performance index J is to say that we are willing to use control u_{j_0} when state error x_{i_0} is reached. We can make Q and R diagonal with

$$\begin{aligned} q_{ii} &= \frac{1}{x_{i_0}^2} ; \quad i = 1; 2; \dots; n \quad (n \text{ states}) ; \\ r_{jj} &= \frac{1}{u_{j_0}^2} ; \quad j = 1; 2; \dots; m \quad (m \text{ controls}) ; \end{aligned}$$

In our case the performance index is, in general,

$$J = \int_0^{\infty} (q_{11}\mu^2 + q_{22}w^2 + q_{33}q^2 + q_{44}z^2 + r_{\pm}^2) dt ;$$

In this case we want to control μ and z near zero (their nominal values) and use a reasonable amount of dive planes to do the job. We assume it would be reasonable to use 5° dive planes for depth control when the pitch angle deviates 3° from zero or the boat reaches a depth deviation of 1.5 feet (about one tenth of the length). We, therefore, assume all terms in Q and R to be zero except,

$$q_{11} = \frac{\mu^3}{57.3} = 364.8 \text{ weighting on } \mu^2 ;$$

$$q_{44} = (1.5)^2 = 0.444 \text{ weighting on } z^2 ;$$

$$r_{11} = \frac{\mu^5}{57.3} = 133.3 \text{ weighting on } \pm^2 ;$$

The performance index is

$$J = \int_0^{\infty} (q_{11}\mu^2 + q_{44}z^2 + r_{11}\pm^2) dt ;$$

and the control law then becomes

$$\pm = -(2.7570\mu - 0.5457w - 2.7657q + 0.0577z) ;$$

and the closed loop poles are

$$-0.5207 \pm 0.2841i \quad \text{and} \quad -0.1197 \pm 0.0704i ;$$

A numerical simulation in terms of z and \pm is shown in Figure 29 by the solid curves. If we decide to use 5° dive planes for depth control when the pitch angle deviates 3° from zero or the boat reaches a depth deviation 0.5 feet from zero, we expect a tighter control law: the same dive plane angle is commanded for one third the error in z . In this case the control law is

$$\pm = -(4.6187\mu - 0.5177w - 4.5379q + 0.1732z) ;$$

and the closed loop poles are

$$-0.4901 \pm 0.2819i \quad \text{and} \quad -0.2267 \pm 0.1111i ;$$

The dominant pole is more negative in this case, as it should. The results of this simulation are also shown in Figure 29 with the dotted curves, the response is faster at the expense of more plane activity.

Other performance indices are also possible. Suppose the objective is to keep the submarine at constant depth, $z = 0$, while minimizing the added drag due to dive plane activity. The design is then for a depth controller which will minimize the added drag on the boat due to its deviations from the equilibrium (nominal) level flight path $x = [\mu; w; q; z]^T = [0; 0; 0; 0]^T$ and control $\pm = 0$. To formulate the problem we need the longitudinal (surge) equation of motion, which is (see ME 4823 for details)

$$(m - X_{\dot{u}})\dot{u} = X_{qq}q^2 + (X_{wq} - m)wq + X_{ww}w^2 + X_{UU}U^2 + X_{\pm\pm}\pm^2 + T_{prop} ;$$

where X_{UU} represents the drag coefficient in straight line motion, T_{prop} is the propulsive force, and the terms X_{qq} , X_{wq} , X_{ww} , $X_{\pm\pm}$ produce the added drag due to nonzero w , q , \pm . The control objectives here are:

- depth control : minimize z^2 , deviation from desired ;
- added drag : minimize $\int F_d$;

where

$$\int F_d = \int (X_{qq}q^2 + (X_{wq} + m)wq + X_{ww}w^2 + X_{\pm\pm\pm}^2) dt ;$$

The weighting index is then

$$J = \int_0^Z (q_{44}z^2 + \int F_d) dt ;$$

or

$$J = \int_0^Z ((X_{qq}q^2 + (X_{wq} + m)wq + X_{ww}w^2 + X_{\pm\pm\pm}^2) dt ;$$

Therefore, we can use

$$Q = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & X_{ww} & \frac{1}{2}(X_{wq} + m) & 0 \\ 0 & \frac{1}{2}(X_{wq} + m) & X_{qq} & 0 \\ 0 & 0 & 0 & q_{44} \end{bmatrix} ;$$

and

$$R = \int X_{\pm\pm} ;$$

where q_{44} is the weighting factor between minimizing depth deviations and minimizing drag. Relatively large values of q_{44} will penalize depth deviations heavily and will result in tight control with increased plane activity (this may be required in operations at periscope depth, for example). On the other hand, if q_{44} is chosen small, the resulting control law will penalize control activity more resulting in minimizing drag and fuel efficiency, with larger depth deviations from nominal.

6.7 Time Optimal Control of a Double Integral Plant

Consider the dynamical system,

$$M \ddot{\mathbf{X}} = \mathbf{F} ;$$

If we define,

$$x_1 = \mathbf{x} ; \quad x_2 = \dot{\mathbf{x}} ; \quad u = \frac{\mathbf{F}}{M} ;$$

we can write it in state space form as,

$$\begin{aligned} \dot{x}_1 &= x_2 ; \\ \dot{x}_2 &= u ; \end{aligned}$$

We also assume the control constraints

$$|u| \leq 1 ;$$

and the initial conditions,

$$x_1(0) = x_{10}; \quad x_2(0) = x_{20}; \quad x_1(T) = x_2(T) = 0;$$

We want to minimize the time to ϕ ,

$$\min J = \int_0^T dt;$$

The Hamiltonian is

$$H(x; p; u; t) = p^T f(x; u; t) - L(x; u; t) = p_1 x_2 + p_2 u - 1;$$

The necessary conditions for optimality are

$$\begin{aligned} \dot{x}_1 &= \frac{\partial H}{\partial p_1} = x_2; \\ \dot{x}_2 &= \frac{\partial H}{\partial p_2} = u; \\ \dot{p}_1 &= - \frac{\partial H}{\partial x_1} = 0; \\ \dot{p}_2 &= - \frac{\partial H}{\partial x_2} = - p_1; \end{aligned}$$

Pontryagin's maximum principle states that u must maximize $H = p_1 x_2 + p_2 u - 1$. Therefore, the optimal control needs to maximize $p_2 u$ (since the rest of H does not depend on u). We can see that if p_2 is positive, u must get the maximum positive value (in this case $+1$), while if p_2 is negative, u must be -1 . Therefore, the optimal control is given by

$$u = \text{sgn}[p_2(t)] = +1 \text{ if } p_2 > 0 \text{ and } -1 \text{ if } p_2 < 0;$$

The optimal trajectory is given by the solution to,

$$\begin{aligned} \dot{x}_1 &= x_2; \\ \dot{x}_2 &= \text{sgn}(p_2); \\ \dot{p}_1 &= 0; \\ \dot{p}_2 &= - p_1; \\ x_1(0) &= x_{10}; \quad x_2(0) = x_{20}; \quad x_1(T) = 0; \quad x_2(T) = 0; \end{aligned}$$

This is a reduced system of equations, since u is eliminated by maximizing H .

To solve this system we observe that since $\dot{p}_1 = 0$ we have that $p_1 = \text{const}$: and this means that p_2 is a first-order polynomial in t . Therefore, it can only go from positive to negative at most once in its life, which means that there are only four possible control sequences,

$$+1g; \quad -1g; \quad +1;-1g; \quad -1;+1g;$$

If we let $U = \pm 1$ be the control, we have

$$\begin{aligned}x_1 &= x_{10} + x_{20}t + \frac{1}{2}U t^2 ; \\x_2 &= x_{20} + U t ;\end{aligned}$$

If we eliminate t we can get

$$x_1 - x_{10} - \frac{1}{2}U x_{20}^2 = \frac{1}{2}U x_2^2 ;$$

which represents a family of parabolas as shown in Figure 30. If $u = +1$ we are located on branch A while if $u = -1$ we are on branch B. The branch that goes through the origin is called the switching line and it is given by

$$x_1 = -\frac{1}{2}x_2^2 ;$$

To see how this optimal control works, suppose we start from an initial condition with both x_1 and x_2 positive. We apply control $u = -1$ until we hit the switching line, there we switch to $u = +1$ and we land at the origin with zero velocity.

A feedback control implementation is shown in Figure 31. We define

$$z = x_1 + \frac{1}{2}x_2^2 ;$$

which means that the switching line is $z = 0$. Therefore, we get the optimal control through a switch $u = -1$ when $z > 0$ and $u = +1$ when $z < 0$. We should point out that in this case the final portion of the state trajectory follows the switching curve, this is not typical for all systems though. Since the optimal control switches from positive to negative we call it bang-bang control. Most minimum time control problems lead to bang-bang controllers. Pontryagin has shown that for a system of order n with negative real poles and scalar u , $|u| \leq 1$, the optimal control switches at most $n - 1$ times.

7 DISCRETE AND STOCHASTIC SYSTEMS

7.1 Discrete Systems

Recall our basic continuous system in state space form,

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

A control system that is to be implemented using a digital computer, as is usually the case, is in a discrete state space form,

$$\begin{aligned}x_{n+1} &= A_d x_n + B_d u_n ; \\y_n &= C_d x_n ;\end{aligned}$$

The first thing we have to do is to be able to go from the continuous to the discrete model. We start with the solution to the state equations in the form

$$x(t) = e^{A(t-t_0)}x(t_0) + \int_{t_0}^t e^{A(t-\zeta)}B u(\zeta) d\zeta ;$$

We can use this solution over one sample period T to obtain a difference equation. Let

$$\begin{aligned} t &= nT + T ; \\ t_0 &= nT ; \end{aligned}$$

and we get

$$x(nT + T) = e^{AT}x(nT) + \int_{nT}^{nT+T} e^{A(nT+T-\zeta)}B u(\zeta) d\zeta ;$$

Now assume that the input does not change within one sample period,

$$u(\zeta) = u(nT) \quad \text{for} \quad nT \leq \zeta < nT + T ;$$

We refer to this operation as the zero-order hold with no delay. Then, by defining the auxiliary variable

$$\zeta' = nT + T - \zeta ;$$

we get

$$x(nT + T) = e^{AT}x(nT) + \int_0^T e^{A\zeta'}B u(nT) d\zeta' ;$$

Therefore, the system

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

becomes

$$\begin{aligned} x_{n+1} &= A_d x_n + B_d u_n ; \\ y_n &= C_d x_n ; \end{aligned}$$

where

$$\begin{aligned} A_d &= e^{AT} ; \\ B_d &= \int_0^T e^{A\zeta'} B d\zeta' ; \\ C_d &= C ; \end{aligned}$$

and T is the sample period. The MATLAB command **cd** automates the above conversion from continuous to discrete form.

A low sample period T ; i.e., high sample rate, is in general desirable for good performance so that we can approximate the continuous model as closely as possible. This, however, will demand a fast computer and A/D and D/A converters. It should be emphasized here that

low T is always with respect to the response time of the physical system. Low T for one system may be high for a different system. Low sample rate, high T , may lead to instabilities when the design is based on the continuous system. In such a case we should switch to a direct discrete design. This means that the continuous system is discretized first, and any compensator design is based on the discrete version. Fortunately this parallels the continuous design we have already developed.

We can place the poles of a discrete system to desirable locations by linear state variable feedback,

$$u_n = -K x_n ;$$

and if not all states are measurable we can use a discrete full-order estimator,

$$\hat{x}_{n+1} = A_d \hat{x}_n + B_d u_n + L (y_n - C_d \hat{x}_n) :$$

We can find the gain matrices K and L by poleplacement of

$$A_d - B_d K ;$$

and

$$A_d - L C_d :$$

We already know how to do the poleplacement design, the only thing we need to know is: When is a discrete system $x_{n+1} = A x_n$ stable? We can see this by considering a scalar system. Consider the continuous system

$$\dot{x} = ax :$$

The solution is $x(t) = e^{at}x(0)$ so if $a < 0$ the system will be stable. The discrete system

$$x_{n+1} = ax_n ;$$

has

$$\begin{aligned} x_1 &= ax_0 ; \\ x_2 &= ax_1 = a^2x_0 ; \\ x_3 &= ax_2 = a^3x_0 ; \end{aligned}$$

and, finally,

$$x_n = a^n x_0 :$$

For stability, we want $x_n \rightarrow 0$ as $n \rightarrow \infty$, or $a^n \rightarrow 0$, which means that we want

$$|a| < 1 :$$

Therefore, the discrete time system $x_{n+1} = A x_n$ is stable if and only if all eigenvalues of A have absolute value less than one; i.e., they are located inside the unit circle in the s -plane, see Figure 32. Since the continuous matrix A becomes e^{AT} when discretized, we can argue that an eigenvalue which is equal to s , for a continuous system, corresponds to an eigenvalue

equal to e^{-T} for a discrete system with sample period T . By keeping this analogy in mind we can do in discrete time everything we did in continuous time. The corresponding MATLAB commands have the same names with simply the pre- x **d** in front, for example **dlqr** will do the discrete LQR design.

As an example, consider the system

$$\dot{x} = -x + u;$$

which is open-loop unstable. A control law of the form $u = -2x$ places the closed loop pole of the continuous system at -1 , this means that the continuous system has a time constant 1 second. Now let's discretize the system using a sample period T , we set the closed loop pole of the discrete system at e^{-T} . How different will be the discrete gain from the continuous gain? This should depend strictly on T . If T is very small compared to 1, the time constant of the system, then the two gains must be relatively close. Ten times smaller should be small enough. On the other hand, if T is of the same order of magnitude as 1, we have to compute the gain from the discrete design. This is illustrated by the results of Figure 33 where we present the discrete time gain for a discrete closed loop pole at e^{-T} , versus T for T from 0.01 sec to 1 sec. This corresponds to sample rates from 100 Hz to 1 Hz, respectively.

7.2 Stochastic Processes

To this point we have treated the entire control/estimation problem as deterministic; everything had a known value at each time. In real world problems, however, there are quantities which we can only describe probabilistically, for example sensor characteristics or sea waves. There are unpredictable disturbances and measurement noise which occur during operation of real systems. These disturbances and noise can be modeled as stochastic processes. A very useful special class of stochastic processes is the Gauss-Markov process which can be completely described by the following:

1. Its mean value vector \bar{x} ,

$$\bar{x} = E[x(t)];$$

($n \times 1$)

which gives the expected value or ensemble average of all possible observations at time t ; this is the most likely value.

2. Its correlation matrix C ,

$$C_{(i,j)} = E[(x(t) - \bar{x}(t))(x(i) - \bar{x}(i))^T];$$

($n \times n$)

which is a symmetric matrix and gives the relationship between the deviation from the mean at time t to the deviation from the mean at a different time i .

When $t = \zeta$, this correlation matrix becomes the covariance matrix which measures the mean square deviation of the state vector from the mean; i.e.,

$$X_{\mathbf{z}}(t) \quad C(t; t) = E_{\mathbf{n}} [x(t) - \bar{x}(t)][x(t) - \bar{x}(t)]^T \quad \mathbf{o} :$$

At any time t , the state $x(t)$ is normally distributed (Gaussian distribution) about the mean and the diagonal elements of $X(t)$ give the variance (standard deviation squared) for the associated elements of x .

A special Gaussian Markov process is the purely random process. This is an idealized, very jittery process which is completely uncorrelated from one time to the next. This is a useful model for disturbances or noise which change very rapidly compared with the time response of a system. The correlation matrix for a purely random process is

$$C(t; \zeta) = Q(t) \delta(t - \zeta) ;$$

where $Q(t)$ is the power spectral density, and $\delta(t - \zeta)$ is the Dirac delta function; this is zero everywhere except at $t = \zeta$ where it assumes a "value" such that $\int_{-\infty}^{\infty} \delta(t - \zeta) d\zeta = 1$. This can be viewed as the limit of a sequence of impulses of random magnitude (equal plus and minus so the mean is zero; average square magnitude is $\frac{1}{2} Q(t)$ and random time of occurrence. For such a sequence,

$$Q(t) = 2 \bar{\nu} \bar{w}^2 \delta(t) ;$$

where $\bar{\nu}$ is the average number of occurrences per unit time.

The key behind using Gaussian Markov processes is that a Gaussian Markov process can always be represented by a state vector of a linear dynamical system forced by a Gaussian purely random process where the initial state vector is Gaussian. Thus,

$$\dot{\mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{w} ;$$

where

$$\begin{aligned} E_{\mathbf{n}} [\mathbf{w}(t)] &= \bar{\mathbf{w}} = \mathbf{0} ; \\ E_{\mathbf{n}} [\mathbf{w}(t) \mathbf{w}^T] &= Q(t) \delta(t - \zeta) ; \\ E_{\mathbf{n}} [\mathbf{x}(t_0)] &= \bar{\mathbf{x}}_0 ; \\ E_{\mathbf{n}} [x(t_0) - \bar{x}_0][x(t_0) - \bar{x}_0]^T &= X_0 ; \\ E_{\mathbf{n}} [w(t) - \bar{w}][x(t_0) - \bar{x}_0]^T &= 0 ; \end{aligned}$$

The forcing disturbance w and the initial state $x(t_0)$ are completely independent or uncorrelated. Recall the state property for deterministic systems: knowing the current state and the state equation completely defines the future for zero control. The Markov property is completely parallel to this: knowing the current state mean \bar{x}_0 and covariance matrix X_0 completely defines the future mean and covariance for zero control when subjected to the

disturbance described by $\bar{w} = 0$ and Q . The Gaussian property states that the state will always be normally distributed about the mean value in accordance with the variance (standard deviation squared) given by the diagonal elements of the covariance matrix. Thus for one state x , it will be within one standard deviation $\frac{3}{4}$ of \bar{x} 68.3% of the time; within $2\frac{3}{4}$ of \bar{x} 95.5% of the time; within $3\frac{3}{4}$ of \bar{x} 99.7% of the time. For multiple states these percentages decrease as shown in the following table:

n	$\frac{3}{4}$	$2\frac{3}{4}$	$3\frac{3}{4}$
1	68:3	95:5	99:7
2	39:4	86:5	98:9
3	20:0	73:9	97:1

The mean value vector of a Gauss{Markov process obeys the state differential equation

$$\dot{\bar{x}} = A\bar{x} + \bar{w}; \quad \bar{x}(t_0) = \bar{x}_0;$$

The covariance matrix obeys equation

$$\dot{X} = A X + X A^T + \bar{w} \bar{w}^T; \quad X(t_0) = X_0;$$

which is completely independent and which could be calculated in advance. Note that the term $A X + X A^T$ represents the effect of the system dynamics and it may decrease X for a stable system, while the other term $\bar{w} \bar{w}^T$ represents the effect of the disturbance and it always increases X since we have a positive definite Q .

We can visualize this by considering a simple first order system so that all the above matrices are scalars. A stable first order system with an initial mean \bar{x}_0 and small standard deviation $\frac{3}{4}_0$ could be released while subjected to noise. It could respond as shown in Figure 34 for large noise. As an example suppose we have the system

$$\dot{x} + 2x = w;$$

where w is zero mean, purely random (white noise), x is exactly 1 at $t = 0$. At this time, x is released and the disturbance w with power spectral density $Q = q = 3$ begins to act on the system. We want to determine the mean and the covariance of the response. The mean will follow the state equation

$$\dot{\bar{x}} = A\bar{x} + \bar{w} = -2\bar{x}; \quad \bar{x}(0) = 1; \quad A = -2;$$

and $\bar{w} = 0$ since w is white noise. The solution for the mean is

$$\bar{x}(t) = e^{-2t};$$

The covariance will follow equation

$$\dot{X} = A X + X A^T + \bar{w} \bar{w}^T; \quad Q = q = 3; \quad \bar{w} = 1;$$

or

$$\dot{X} = -2X - 2X + q;$$

and, with exact knowledge at $t = 0$, the initial condition is $X(0) = 0$. The solution is

$$X(t) = 0.75 + 1 \int_0^t e^{-4t} dt = 3/4 + t e^{-4t};$$

the variance of $x(t)$ about its mean $\bar{x}(t)$, refer to Figure 35.

normally distributed			
t	\bar{x}	X	σ^2
0	1	0	0
0.5	0.368	0.648	0.805
1	0	0.750	0.866

Most physical disturbances can be modeled by one of the following special cases:

1. White noise: A stationary, purely random Gauss-Markov process with zero correlation time (see below) and constant power spectral density,

$$C = Q \pm(t_j - t_k):$$

2. Random bias: A random, unpredictable constant with infinite correlation time and constant correlation. In this case we introduce

$$\dot{x}_{n+1} = 0; \quad x_{n+1}(t_0) = \text{random}:$$

We add a new, constant state to the system of equations; i.e., we augment the state equations and estimate x_{n+1} along with the rest of the states x_i , $i = 1; \dots; n$, as we have already seen before. As examples, a disturbance which changes rapidly compared to the dominant dynamics of the system can be modeled as white noise; e.g. wave effects on the steering of a large tanker. A disturbance which changes very slowly compared to the dominant dynamics of the system can be modeled as a random bias; e.g. tidal current on ship steering.

3. Exponentially correlated noise: Between the two extremes where white noise and random bias models are appropriate, are disturbances which change on the same time scale as the dominant dynamics of the system. These disturbances have finite, non-zero correlation times ζ_c . The simplest can be modeled as a first order system driven by white noise; i.e.,

$$\zeta_c \dot{x}_{n+1} + x_{n+1} = w:$$

In these cases the state vector can be augmented with x_{n+1} . Disturbances which change with about the same dynamics as the system must be modeled with a finite ζ_c ; e.g. the force and moment produced by a passing ship during underway replenishment. The above equation is called a shaping filter because it "shapes" white noise w to produce another disturbance x_{n+1} which is called "colored" noise. The correlation time is the same as the time constant of the disturbance variation, this can be obtained by considering the physics of the problem. For example, if the disturbance is the force produced by a passing ship we can take ζ_c to be approximately the time it takes to travel a ship length.

To complete the model for the exponentially correlated disturbance it is necessary to specify the power spectral density of the white noise w . This is given by,

$$q = 2\zeta_c^2 \zeta_c;$$

where

$$\begin{aligned} \sigma &= \text{root mean square (RMS) noise level,} \\ \tau_c &= \text{correlation time:} \end{aligned}$$

The same formula is also used in design to establish the power spectral density of disturbances modeled as white noise. In that case the correlation times (modeled as zero) are actually small nonzero quantities compared to the time constants of the system. In practice this can be the integration time step in simulations, or the sample time in experiments.

More complex models for modeling disturbances are also possible, this is a trade-off between accuracy and simplicity. Of great interest to naval engineering is the modeling of the disturbance due to waves. The simplest approach would be to model this as white noise, this is very accurate for large ships. For smaller vessels it might be worth modeling the periodic nature of the disturbance. There are a couple of ways to do this. If we assume a sinusoidal wave as the dominant model for waves in the area, we can use a second order model driven by white noise w ,

$$\ddot{y} + \omega^2 y = w ;$$

where ω is the assumed frequency of the dominant wave (usual periods of sea waves are in the 6 to 15 sec range), and y is the amplitude of the disturbance. In state space form then we need to augment our system with two additional equations

$$\begin{aligned} \underline{x}_{n+1} &= \underline{x}_{n+2} ; \\ \underline{x}_{n+2} &= j \omega \underline{x}_{n+1} + w ; \end{aligned}$$

where $y = x_{n+1}$ and $\dot{y} = x_{n+2}$. More accurate descriptions of the seaway are also used. A typical description follows the so-called Pierson-Moskowitz wave spectrum given by

$$S(\omega) = \frac{a}{\omega^5} e^{-b\omega^4} ;$$

where a, b are constants describing the particular seaway. Such a spectrum can be simulated by feeding a white noise signal into a suitable shaping filter. As an example, for a significant wave height (the average of the highest one third of all wave heights) of 7 m and a mean wave period of 9.4 seconds we have $a = 0.78$ and $b = 0.063$. Then the rational spectrum

$$S_R(\omega) = \frac{b_2^2 \omega^2}{\omega^6 + (a_1^2 + 2a_2)\omega^4 + (a_2^2 + 2a_1a_3)\omega^2 + a_3^2} ;$$

with $a_1 = 0.5$, $a_2 = 0.33$, $a_3 = 0.07$, and $b_2 = 0.415$ can be used as an approximation of $S(\omega)$ for the chosen sea state. When both S and S_R are plotted versus ω the agreement is good. For details see the article "Control of yaw and roll by a rudder-roll stabilization system" by Kallstrom in the Proceedings of the Sixth Ship Control Systems Symposium, 1981. A stochastic process with spectral density given by S_R can be obtained as output from the filter

$$G(s) = \frac{b_2 s}{s^3 + a_1 s^2 + a_2 s + a_3} ;$$

with white noise as input. A similar model can be built for approximating the wave slope spectrum,

$$S_s(\omega) = \frac{\omega^4}{g^2} S(\omega);$$

and either one or both wave height and wave slope models can then be used for realistic design and simulations.

7.3 Kalman Filter

We present now the maximum likelihood, stochastic observer or filter for a nonstationary Gauss-Markov process. This will be seen to be completely parallel to the deterministic observer discussed in Section 3. We will sketch the derivation of the continuous time Kalman filter using calculus of variations in a manner which parallels our derivation of the optimal control law in 6:6.

Recall our classical full order observer design

$$\dot{\mathbf{b}} = \mathbf{A}\mathbf{b} + \mathbf{B}u + \mathbf{L}(y - \mathbf{C}\mathbf{b});$$

In general, we would like to place the observer poles as negative as possible, this will create large elements of the observer gain matrix \mathbf{L} . The larger the \mathbf{L} , the faster the error in the observer dynamics will decay to zero. A very large \mathbf{L} , however, will amplify undesirable noise which is always present in real systems. Therefore, there seems to be a limit on \mathbf{L} which should depend on the level of noise in the system; this in turn should be directly related to the quality of our sensors and the disturbances. The Kalman filter is this best value for \mathbf{L} and it provides an optimal stochastic observer, just like the linear quadratic regulator provided an optimal controller.

Consider the system

$$\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}u + \mathbf{w};$$

where \mathbf{w} is a purely random process, and

$$\begin{aligned} E[\mathbf{x}(t_0)] &= \bar{\mathbf{x}}_0; \\ E[\mathbf{x}(t_0) - \bar{\mathbf{x}}_0][\mathbf{x}(t_0) - \bar{\mathbf{x}}_0]^T &= \mathbf{P}_0; \end{aligned}$$

which is the covariance of the error in the estimate of the state $\mathbf{b}(t_0)$ at t_0 . Initially we assume that $\mathbf{b}(t_0) = \bar{\mathbf{x}}_0$: the most likely estimate at t_0 is the mean value at that time. In general,

$$\mathbf{P}(t) = E[\mathbf{e}(t) - \hat{\mathbf{x}}(t)][\mathbf{e}(t) - \hat{\mathbf{x}}(t)]^T = E[\mathbf{e}(t)\mathbf{e}^T(t)];$$

where $\mathbf{e} = \mathbf{b} - \hat{\mathbf{x}}$ is the error in the estimate of the state. The disturbance \mathbf{w} in the state equations is a purely random process with

$$\begin{aligned} E[\mathbf{w}(t)] &= \mathbf{0}; \quad \text{zero mean;} \\ E[\mathbf{w}(t)\mathbf{w}^T(t)] &= \mathbf{Q}(t)\delta(t - \tau); \end{aligned}$$

where Q is the power spectral density matrix. We want to estimate the state vector $\mathbf{x}(t)$ using a set of noisy measurements,

$$y = Cx + v;$$

where the measurement noise v is another purely random process with

$$\begin{aligned} E[v(t)] &= 0; \quad \text{zero mean;} \\ E[v(t)v^T(t)] &= R(t)\delta(t - \tau): \end{aligned}$$

What we want to do is to generate an estimate of both x and w which enter the state equations. This can be done in a least square sense if we minimize the cost function

$$J = \frac{1}{2} (\mathbf{x}_0 - \bar{\mathbf{x}}_0)^T P_0^{-1} (\mathbf{x}_0 - \bar{\mathbf{x}}_0) + \frac{1}{2} \int_{t_0}^{t_f} \mathbf{w}^T Q^{-1} \mathbf{w} + (y - Cx)^T R^{-1} (y - Cx) dt:$$

Observe that the first term minimizes the error in the initial estimate; the second term minimizes the error in the estimate of w ; and the third term minimizes the error in the estimate of x . The minimization is subject to the constraints

$$\begin{aligned} \dot{x} &= Ax + Bu + w; \\ y &= Cx + v: \end{aligned}$$

Following a process similar to the LQR design, we can define the Hamiltonian

$$H = \frac{1}{2} \mathbf{w}^T Q^{-1} \mathbf{w} + (y - Cx)^T R^{-1} (y - Cx) + \lambda^T (Ax + Bu + w);$$

and formulate the Euler-Lagrange equations, as before. We can find then that the optimal observer has the familiar form,

$$\dot{\mathbf{b}} = A\mathbf{b} + Bu + L(y - C\mathbf{b}); \quad \mathbf{b}(t_0) = \bar{\mathbf{x}}_0;$$

where L is the Kalman filter gain matrix

$$L = PC^T R^{-1};$$

and P is the solution of the forward matrix Riccati differential equation

$$\begin{aligned} \dot{P} &= AP + PA^T + Q - PC^T R^{-1} CP; \\ P(t_0) &= P_0: \end{aligned}$$

In the steady state case, these results become

$$L = PC^T R^{-1};$$

where now P is the solution to the algebraic Riccati equation

$$AP + PA^T + Q - PC^T R^{-1} CP = 0:$$

The positive definite solution defines P , the covariance of the error in the estimate of the state \mathbf{b} .

As an example, consider the system

$$\begin{aligned} \dot{x} &= -2x + w ; \quad \text{so } A = -2; \quad b = 1; \\ y &= x + v ; \quad \text{so } C = 1; \end{aligned}$$

The disturbance w is exponentially correlated with a correlation time

$$\tau_w = 0.01 ;$$

and root mean square value

$$\sigma_w = 1.2 ;$$

The measurement noise v is also exponentially correlated with correlation time

$$\tau_v = 0.01 ;$$

but with an RMS value

$$\sigma_v = 0.2 ;$$

We want to design a Kalman filter to produce a best estimate of x from y . The system has the time constant

$$T = 0.5 \text{ s} ;$$

so we can model both the disturbance and noise as white noise compared with the dynamics of the system. The power spectral densities are estimated as

$$\begin{aligned} \text{for } w : \quad Q &= \frac{1}{4} 2 \sigma_w^2 \tau_w = 2(1.2)^2 0.01 = 0.0288 ; \\ \text{for } v : \quad R &= \frac{1}{4} 2 \sigma_v^2 \tau_v = 2(0.2)^2 0.01 = 0.0008 ; \end{aligned}$$

Our filter is given by

$$\dot{\hat{x}} = -2\hat{x} + L(y - \hat{x}) ; \quad L = P R^{-1} ;$$

To find P we use the algebraic Riccati equation

$$\begin{aligned} -2P + P(-2) + 0.0288 + P \frac{1}{0.0008} P &= 0 \\ P^2 + 0.0032P - 0.0002304 &= 0 \\ P &= 0.00346 ; \end{aligned}$$

the positive root. Then

$$L = P R^{-1} = \frac{0.00346}{0.0008} = 4.3246 ;$$

giving

$$\dot{\hat{x}} = -2\hat{x} + 4.3246(y - \hat{x}) = -6.3246\hat{x} + 4.3246y ;$$

The error in the estimate produced by the filter is

$$\begin{aligned} \dot{e} &= \dot{\hat{x}} - \dot{x} \\ &= A\hat{x} + Bu + L(Cx + v - C\hat{x}) - Ax - Bu - \dot{w} \\ &= (A - LC)e + Lv - \dot{w} \\ &= (-2 - 4.3246)e + 4.3246v - \dot{w} \\ &= -6.3246e + 4.3246v - \dot{w} ; \end{aligned}$$

The eigenvalue of the filter is at -6.3246 which is well to the left of the system eigenvalue, -2 , so the estimate will converge fast compared to the system.

7.4 The LQG Compensator

Recall that the separation principle allowed us to design the controller and the estimator separately and then use $\hat{\mathbf{x}}$ instead of \mathbf{x} in the control law. The same principle states here that the optimal way to control a system

$$\dot{\mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u} + \mathbf{j} w ;$$

is to use a Kalman stochastic observer to estimate the state from the noisy measurements

$$y = \mathbf{C} \mathbf{x} + v ;$$

and then use this estimate $\hat{\mathbf{x}}$ with the optimal deterministic linear controller we have already developed. The optimal controller can be derived from the LQR design, or we can use any kind of state feedback and feedforward we desire. The key is that we have no control over the poles of the observer here, nor can we choose the \mathbf{Q} and \mathbf{R} matrices that enter the Kalman filter design. These are set by the quality of our sensors and the level of the disturbances. After computing \mathbf{L} from the Riccati equation, we should find the observer poles from the eigenvalues of $(\mathbf{A} - \mathbf{L} \mathbf{C})$ and make sure that they are more negative (the dominant pole) than the dominant poles of the controller. This can be done directly if we use poleplacement or indirectly by changing the weighting matrices in the LQR design. In case that the controller poles are not satisfactory, it is time to get better sensors!

The above combination of the optimal controller (LQR) and the optimal stochastic observer (Kalman filter) is called the Linear Quadratic Gaussian (LQG) compensator. This theoretical result produces a control system which is completely parallel to the deterministic observer and controller derived previously, except that now the controller and observer gain matrices are theoretically derived to yield optimal performance in the presence of stochastic disturbances w and measurement noise v .

Summarizing the total design problem, we have:

² State \mathbf{x} ,

$$\dot{\mathbf{x}} = \mathbf{A} \mathbf{x} + \mathbf{B} \mathbf{u} + \mathbf{j} w ; \quad \mathbf{x}(t_0) = \mathbf{x}_0 ;$$

with

$$E [w w^T] = \mathbf{Q} \delta(t - \tau) ; \quad \text{white noise ;}$$

$$E [w] = 0 ;$$

$$\text{covariance } \mathbf{X} = \begin{matrix} \mathbf{h} & \mathbf{i} \\ (\mathbf{x} - \bar{\mathbf{x}}) & (\mathbf{x} - \bar{\mathbf{x}})^T \end{matrix} ;$$

² Estimate $\hat{\mathbf{x}}$,

$$\dot{\hat{\mathbf{x}}} = \mathbf{A} \hat{\mathbf{x}} + \mathbf{B} \mathbf{u} + \mathbf{L} (y - \mathbf{C} \hat{\mathbf{x}}) ; \quad \hat{\mathbf{x}}(t_0) = \bar{\mathbf{x}}_0 ;$$

with covariance

$$\mathbf{X} = \begin{matrix} \mathbf{h} & \mathbf{i} \\ (\hat{\mathbf{x}} - \bar{\mathbf{x}}) & (\hat{\mathbf{x}} - \bar{\mathbf{x}})^T \end{matrix} ;$$

² Error in estimate \mathbf{x} ,

$$\mathbf{e} = \mathbf{b}_j \mathbf{x};$$

with covariance

$$P = \mathbf{h} (\mathbf{b}_j \mathbf{x}) (\mathbf{b}_j \mathbf{x})^T = E \mathbf{e} \mathbf{e}^T;$$

² Measurements y ,

$$y = C \mathbf{x} + v;$$

with

$$E v v^T = R \pm (t_j \ j); \text{ white noise};$$

$$E [v] = 0;$$

² Controller,

$$u = -K \mathbf{b};$$

² Controller gain K ,

$$K = R^{-1} B^T S;$$

where

$$A^T S + S A - S B R^{-1} B^T S + Q = 0;$$

² Estimator gain L ,

$$L = P C^T R^{-1};$$

where

$$A P + P A^T - P C^T R^{-1} C P = 0;$$

It should of course be emphasized that the matrices Q and R that enter the controller design are completely different than those in the observer design. The block diagram of the LQG design is shown in Figure 38.

7.5 Linear Quadratic Gaussian Compensator block diagram

With the optimal design developed above the performance can be evaluated either probabilistically or deterministically using computer simulation. Here we will develop the root mean square (RMS) response which can be easily computed for linear systems and can serve as a comparison index for different designs. If we wish to establish the response of the state about zero (the states are defined as deviations from nominal), we can begin with the free response. Using the previous equations,

$$\dot{\mathbf{b}} = (A - B K) \mathbf{b} + L [C (\mathbf{x} - \mathbf{b}) + v];$$

$$\dot{\mathbf{x}} = (A - B K) \mathbf{x} - L C \mathbf{x} + L v; \quad \mathbf{b}(t_0) = \bar{\mathbf{x}}_0 = 0;$$

The dynamics in the error are governed by,

$$\begin{aligned} \dot{\mathbf{x}} &= \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u} + \mathbf{L}(C\mathbf{x} + \mathbf{v} - C\hat{\mathbf{x}}) - \mathbf{A}\hat{\mathbf{x}} + \mathbf{B}\mathbf{u} - \dot{\hat{\mathbf{x}}} \\ &= (\mathbf{A} - \mathbf{L}C)\mathbf{x} + \mathbf{L}\mathbf{v} - \dot{\hat{\mathbf{x}}} \end{aligned}$$

with

$$\mathbf{x}(t_0) = \bar{\mathbf{x}}_0, \quad \hat{\mathbf{x}}(t_0) = \hat{\mathbf{x}}_0$$

From the \mathbf{b} and \mathbf{x} equations we can see that \mathbf{x} is statistically independent of \mathbf{b} , so

$$E \mathbf{x}(t_0) \mathbf{b}^T(t_0) = \mathbf{0};$$

and

$$E \mathbf{x}(t) \mathbf{b}^T(t) = \mathbf{0};$$

We can, therefore, establish the covariance of the state to be given by

$$\begin{aligned} X &= E \mathbf{x}(t) \mathbf{x}^T(t) \\ &= E (\mathbf{b} + \mathbf{x}) (\mathbf{b} + \mathbf{x})^T \\ &= E \mathbf{b} \mathbf{b}^T + E \mathbf{x} \mathbf{x}^T + E \mathbf{b} \mathbf{x}^T + E \mathbf{x} \mathbf{b}^T \\ &= E \mathbf{b} \mathbf{b}^T + E \mathbf{x} \mathbf{x}^T : \end{aligned}$$

This gives

$$X(t) = \hat{X}(t) + P(t);$$

or, at steady state,

$$X = \hat{X} + P;$$

which says that

$$\begin{aligned} (\text{covariance of state}) &= (\text{covariance of estimate of state}) \\ &+ (\text{covariance of error in estimate of state}); \end{aligned}$$

We already know how to obtain P and thus we need \hat{X} to obtain X , and the RMS response of the state x which is given by the square root of the diagonal terms in X . If we use the above equation in the definition of the covariance \hat{X} we can finally obtain the following differential equation for \hat{X} ,

$$\dot{\hat{X}} = (\mathbf{A} - \mathbf{B}\mathbf{K})\hat{X} + \hat{X}(\mathbf{A} - \mathbf{B}\mathbf{K})^T + \mathbf{P}C^T R^{-1} C \mathbf{P} = \mathbf{0}; \quad \hat{X}(t_0) = \mathbf{0};$$

which in the steady state yields the linear matrix equation,

$$(\mathbf{A} - \mathbf{B}\mathbf{K})\hat{X} + \hat{X}(\mathbf{A} - \mathbf{B}\mathbf{K})^T + \mathbf{P}C^T R^{-1} C \mathbf{P} = \mathbf{0};$$

which can be solved for \hat{X} and then used in $X = \hat{X} + P$ to obtain X .

The root mean square (RMS) use of the controls u can be derived directly from the definition of its covariance,

$$U = E \{ uu^T \} = E \{ (j K b)(j K b)^T \} = K E \{ bb^T \} K^T = K X K^T :$$

The square root of the associated diagonal elements of X and U give the RMS value of the states and controls, respectively, when the system is subjected to the disturbances w described by Q and the measurement noise described by R . The above equations are estimates of the RMS value of the response of a system and can be used for comparing different control and estimator designs. It should be borne in mind that they are not valid for nonlinear systems; they can not be used when the control effort saturates, for example. In these cases the associated RMS values of the variables of interest should be computed numerically by simulation.

8 NONLINEAR SYSTEMS

We introduce here a few concepts and analysis techniques for nonlinear systems. The analysis and control of linear systems is a necessary step in understanding nonlinear dynamics. Although, as we have seen, almost every nonlinear system can be locally approximated by a linearized system, this correlation should not be pushed too far. For nonlinear systems the principle of superposition of solutions does not hold. There are no separate natural and forced motions. Twice the input does not mean twice the output. For nonlinear systems there may be a significant dependence of the response on the magnitude and type of the excitation. For example, a nonlinear system may have completely different behavior under step inputs of different magnitude, or sinusoidal inputs of different frequencies. The response may also depend drastically on the initial conditions. In fact for some systems it may happen that the long term behavior of the solutions may be effectively random, even though both the system and the input are purely deterministic, as a result of extreme sensitivity to initial conditions. Since one can never be exactly certain about the initial state, the final state of such a system may very well be unpredictable. Such essentially unpredictable deterministic systems are known as chaotic systems.

8.1 Introduction

As a first example of what may happen when nonlinearities are present in a physical system, consider the so called Duffing's equation. This is nothing but a spring-mass-damper system with nonlinear spring force characteristics,

$$m \ddot{x} + b \dot{x} + kx + \alpha x^3 = 0 :$$

The spring force is $kx + \alpha x^3$ instead of kx that would be if the spring were linear. We call the case of $\alpha > 0$ a hardening spring, and $\alpha < 0$ a softening spring. A typical example would be the familiar $GZ(\dot{A})$ curve: it has the characteristics of a hardening spring for small \dot{A} for

a surface ship, and a softening spring for a submarine. The plot of spring force vs. spring displacement would typically appear as shown in Figure 39.

¶ We know that the natural frequency of oscillation of the linear spring system is $\omega_n = \sqrt{k/m}$, in other words it depends only on k and not on the amplitude of oscillation. For a hardening spring, it can be seen that the equivalent linearized spring constant is $k + 3\alpha x^2$, which means that it increases with the displacement x . Therefore, we expect the natural frequency of the hardening spring system to increase with the amplitude of oscillation, as well. The opposite is true for the softening spring case, $\alpha < 0$, see Figure 40.

Now consider Du± ng's equation with forcing,

$$m\ddot{x} + b\dot{x} + kx + \alpha x^3 = P \cos \omega t :$$

We know that the frequency response curve has the familiar shape of Figure 41. It starts from 1, it may reach a maximum at about ω_n depending on the amount of damping, and then it approaches zero. We can observe that the frequency response curve "wraps around" the amplitude vs. frequency curve we had before. Therefore, we can guess that the frequency response curves for the hardening and softening nonlinear springs will take one of the two forms shown in Figure 42.

We can see that depending on the frequency of excitation and upon increasing or decreasing this frequency, the system may experience oscillations with different amplitude, or sudden changes in the amplitude of the response. These phenomena are characteristic of nonlinear restoring forces and moments, and are called jump phenomena or hysteresis.

A different type of phenomena of nonlinear systems may occur when the system is excited with input of frequency ω . A linear system would respond only with the same frequency, but a nonlinear system may experience responses, besides ω , at frequencies $\omega = n\omega$ where n is an integer. These are called subharmonic oscillations. Superharmonic oscillations at frequencies $n\omega$ are also possible although they are not as severe as the subharmonics. This is because higher frequencies are usually associated with more damping. The generation of the above oscillations depends upon the initial conditions, as well as the amplitude and frequency of the excitation.

One question that one may ask is, how many types of behavior are possible for nonlinear systems? The answer to this depends mainly on the system dimensionality. Suppose we have a first order, scalar, system. This involves one variable only, and this can be represented on a straight line. Since it is restricted to move on this line, the system can only experience one or more equilibrium points, and these can be either stable or unstable. Now consider a second order system, this involves two variables x_1, x_2 , and if we want to plot these together we need to use a two-dimensional graph, a plane. The solutions in time on this plane can do whatever they desire except cross each other: this would violate uniqueness of solutions for all subsequent times, since different response would be obtained from identical starting conditions. Solutions of dynamic systems, linear or nonlinear, exist and are unique. We can see that two types of behavior are possible here: the solutions can either approach a point asymptotically (equilibrium point), or a closed curve on which they may be constrained to move for ever. This represents a periodic solution. Such an isolated periodic solution is

called a limit cycle and occurs without any periodic excitation! The study of limit cycles is a very tough but nice problem in nonlinear systems. Now let's imagine a system with three or more state variables. We need at least a three-dimensional graph to plot all of our solutions together here. It is clear that such a system may exhibit both equilibrium points and isolated periodic solutions or limit cycles. In three or more dimensions, the restriction that trajectories may not cross does not constrain the solutions to be simple. There is enough room in three dimensional spaces and beyond so that the solutions they can wrap around each other, twist, turn, and tangle themselves into fantastic knots as they develop in time, forming complicated patterns. Therefore, some complex dynamic behavior is possible for third or higher order systems. Forced and/or discrete systems are usually more complicated. To summarize we can have the following possible types of behavior for nonlinear systems:

- ² First order unforced systems: Equilibrium points only.
- ² Second order unforced systems: Equilibrium points and limit cycles.
- ² Third order or higher unforced systems: Equilibrium points, limit cycles, possible complicated behavior.
- ² Second order or higher forced systems: Equilibrium points, periodic solutions, possible complicated behavior.
- ² Discrete systems of any order: Equilibrium points, periodic solutions, possible complicated behavior.

Let's consider as an example, a Van der Pol equation; a spring-mass-damper system with nonlinear damping and no forcing,

$$m\ddot{x} + b(1 - x^2)\dot{x} + kx = 0 :$$

The equilibrium point of this equation is $x = 0$, the origin. By linearization we can easily see that the origin is unstable. The linearized system is $m\ddot{x} + b\dot{x} + kx = 0$, and we see that $x = 0$ is unstable because of the negative damping term $-b$. So where are the solutions going? We have seen that for small x the solutions move away from $x = 0$. For large x we can see that the term $-b(1 - x^2)$ will become positive, so the damping will be positive and the solutions will have to move towards $x = 0$. Therefore, solutions which originate from large x will move towards the origin. Since they cannot cross each other and there are no other equilibrium points to attract them, they have to approach a limit cycle which should be located somewhere around the origin. This argument, which is known as the Poincaré-Bendixon theorem, holds for second order systems only and it will reveal the existence of a limit cycle but it cannot provide any information about its size or frequency. The sketch of Figure 43 illustrates Poincaré's argument.

Another phenomenon typical in nonlinear systems is the frequency entrainment. Suppose we have a system which is capable of exhibiting a limit cycle of frequency ω_0 . If a periodic force of frequency ω is applied to this system we have the phenomenon of beats. As the difference between the two frequencies decreases, the beat frequency also decreases and, for

a linear system, it is zero only if $\omega = \omega_0$. In a self excited nonlinear system, however, it is found that the frequency ω_0 of the limit cycle falls in synchronization with, in other words it is entrained by, the forcing frequency ω within a certain band of frequencies. This phenomenon is illustrated in Figure 44.

8.2 A Simple Zero Eigenvalue

Suppose we have the nonlinear system of state equations,

$$\dot{\underline{x}} = f(\underline{x}) :$$

We know that the equilibrium points, \bar{x} , of the system are defined by

$$f(\bar{x}) = 0 :$$

This is a nonlinear system of algebraic equations and it may have multiple solutions in \bar{x} , which means that the nonlinear system may have more than one positions of static equilibrium. If we pick one equilibrium, \bar{x} , we can establish its stability properties by linearization. The linearized system becomes

$$\dot{\underline{x}} = A \underline{x} ;$$

where A is the Jacobian matrix of $f(\underline{x})$ evaluated at \bar{x} ,

$$A = \left. \frac{\partial f}{\partial \underline{x}} \right|_{\bar{x}} ;$$

and the state \underline{x} has been redefined to designate small deviations from the equilibrium \bar{x} ,

$$\underline{x} = \underline{x} - \bar{x} :$$

As long as all eigenvalues of A have negative real parts, we know that the linear system will be stable. This means that the equilibrium \bar{x} will be stable for the nonlinear system as well. No surprises so far, in fact what we have just said is nothing but Lyapunov's linearization technique.

The question we ask ourselves next is, what happens if one real eigenvalue of the linearized matrix A is zero? The interesting case here is when the rest of the eigenvalues have all negative real parts, otherwise \bar{x} is unstable and the problem is solved. If the case of a zero eigenvalue appears to be too specialized to be of any practical use consider this: Assume that $f(\underline{x})$ depends on one physical parameter (and there will be plenty of physical parameters in any problem) and that physical parameter is allowed to vary over some range; aren't they all? Then it is clear that A will depend on that parameter and as the parameter varies, it is possible that one real eigenvalue of A will become zero for a specific value of the parameter. Our problem is then to establish the dynamics of the nonlinear system as one real eigenvalue of A crosses zero; i.e., goes from negative to positive. As the solutions evolve in time, things are interesting only along the direction of the eigenvector that corresponds to the critical eigenvalue (the one that crosses zero). Along the rest of the directions in the state space,

everything should converge back to the equilibrium; remember that we assumed that all remaining eigenvalues of A have negative real parts. The above statement should be clear for those of us who haven't forgotten our ME 2801 or O.D.E. material. Although, strictly speaking, it is a true statement for linear systems, there are technical reasons that force it to be true for nonlinear systems as well, the only difference is that the corresponding directions in the state space are curved instead of straight.

We can see then that it is possible to approximate our original system by a one-dimensional system, which is much easier to analyze. The dynamics of the two systems will be qualitatively similar. The formalization of the above reduction procedure constitutes what is known as center manifold reduction, or normal form computation in nonlinear analysis. So let's see what happens for the case of a zero eigenvalue by using a (typical) first order system,

$$\dot{x} = \mu x - x^3;$$

where x is scalar and μ is our distinguished parameter which is allowed to vary between -1 and $+1$. The equilibrium points of the system can be found from

$$\mu x - x^3 = 0 \Rightarrow x(\mu - x^2) = 0;$$

and we can see that, depending on the sign of μ , the equilibria are

$$\bar{x} = 0;$$

if $\mu < 0$, and

$$\bar{x} = 0 \quad \text{and} \quad \bar{x} = \pm \sqrt{\mu};$$

if $\mu > 0$. There is only one equilibrium for negative μ , this is $\bar{x} = 0$, the trivial equilibrium. However, as μ crosses zero moving towards positive values a new pair of equilibria appears out of thin air. These two new equilibria are symmetric (equal plus and minus values), they are close to the trivial equilibrium initially, but as μ moves away from its critical value, $\mu = 0$, they move further away from zero. To analyze the stability properties of these equilibria, let's pick $\bar{x} = 0$ first. The Jacobian is,

$$\frac{\partial f}{\partial x} \Big|_{\bar{x}} = \mu - 3\bar{x}^2;$$

At $\bar{x} = 0$ we get the linearized system

$$\dot{x} = \mu x;$$

and we see that $\bar{x} = 0$ is stable if $\mu < 0$ and unstable if $\mu > 0$. For $\bar{x} = \pm \sqrt{\mu}$ we get the linearized system

$$\dot{x} = \mu - 3(\pm \sqrt{\mu})^2 x = \pm 2\mu x;$$

We can see then that for $\mu > 0$, the equilibrium $\bar{x} = \pm \sqrt{\mu}$ is stable. Remember that, for $\mu < 0$ this equilibrium does not exist. The same is true for the other equilibrium $\bar{x} = \mp \sqrt{\mu}$. Therefore, we can summarize our findings as follows:

- ² For $\mu < 0$ only the trivial equilibrium exists and is stable.
- ² For $\mu > 0$ the trivial equilibrium becomes unstable and a pair of symmetric stable equilibria are generated.

This phenomenon, the loss of stability of an equilibrium and the generation of additional equilibrium states, is called a pitchfork bifurcation and is very common in nature; Euler buckling of a beam is a very typical example. In particular, we refer to the above case as the supercritical pitchfork, this is a rather benign loss of stability since upon loss of stability of the trivial equilibrium the additional nearby equilibrium states are stable. Graphically, we can represent this case as shown in Figure 45 where solid curves represent stable and dotted curves unstable equilibria. We have also indicated the direction of solutions in time of our system for different values of μ . Occasionally, the above case is referred to as a soft loss of stability since for small values of μ , beyond its critical value, the final steady state of the system does not differ much from the nominal (trivial) steady state.

As a second example, consider a "similar" system as before, the linear part remains the same, and the nonlinear part x^3 suffers a sign change,

$$\dot{x} = \mu x - x^3 :$$

We can analyze this in exactly the same way as before, and we can draw the following conclusions (verify these),

- ² For $\mu > 0$ only the trivial equilibrium exists and is unstable.
- ² For $\mu < 0$ the trivial equilibrium becomes stable and a pair of symmetric unstable equilibria are generated.

This case, which is also shown in Figure 45, is called a subcritical pitchfork. A comparison with the previous case reveals that this is a much more serious loss of stability case. Upon loss of stability of the trivial equilibrium position, there is no other stable equilibrium in its vicinity to attract the solutions, which may therefore assume a different state of motion with what could be observed as a discontinuous jump. Furthermore, even before the trivial equilibrium loses its stability the domain of attraction becomes very small and a random perturbation can always throw the system to a different state of motion. This new steady state may be a limit cycle or, depending on the dimensionality of the system, a more complicated response pattern. This loss of stability, sometimes called a hard loss of stability, demonstrates the significance of nonlinear terms in the equations of motion.

8.3 A Purely Imaginary Pair of Eigenvalues

Assume now that our nonlinear system has one pair of purely imaginary eigenvalues for some value of the parameter μ . In other words, this means that as μ is varied over some range, one pair of complex conjugate eigenvalues of the linearized system matrix A crosses the imaginary

axis. It is assumed that the rest of the eigenvalues of A remain negative or have negative real parts. We wish to investigate what happens to the nonlinear system during this process. More specifically, in the previous section we saw that the case of one real eigenvalue crossing zero is associated with the generation or exchange of stability of additional equilibrium points for the nonlinear system. The purpose of this section is to show that the corresponding case of the real part of one complex conjugate pair of eigenvalues crossing zero is associated with the generation of periodic solutions or limit cycles for the nonlinear system.

Following similar arguments as before, we can convince ourselves that in the case of a purely imaginary pair of eigenvalues, the only interesting dynamics of $\underline{x} = f(x)$ will be concentrated on a two dimensional space spanned by the eigenvectors which correspond to the critical pair of eigenvalues of A . We start, therefore, with a two dimensional system in the rather special form,

$$\begin{aligned} \dot{x}_1 &= -\omega x_1 - \eta x_2 + a x_1 (x_1^2 + x_2^2); \\ \dot{x}_2 &= \omega x_1 + \eta x_2 + a x_2 (x_1^2 + x_2^2); \end{aligned}$$

The system admits the trivial equilibrium $\bar{x}_1 = \bar{x}_2 = 0$. The linearized equations around the trivial equilibrium are

$$\begin{pmatrix} \dot{\bar{x}}_1 \\ \dot{\bar{x}}_2 \end{pmatrix} = \begin{pmatrix} -\omega & -\eta \\ \eta & \omega \end{pmatrix} \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \end{pmatrix};$$

with eigenvalues $\pm i\omega$. Therefore, for $\omega = 0$ the eigenvalues are purely imaginary (we assume that $\eta \neq 0$). As ω crosses zero, the trivial equilibrium becomes unstable. To compute other potential equilibrium points for our nonlinear system we use

$$\begin{aligned} -\omega \bar{x}_1 - \eta \bar{x}_2 + a \bar{x}_1 (\bar{x}_1^2 + \bar{x}_2^2) &= 0; \\ \eta \bar{x}_1 + \omega \bar{x}_2 + a \bar{x}_2 (\bar{x}_1^2 + \bar{x}_2^2) &= 0; \end{aligned}$$

If we multiply the first equation by \bar{x}_2 , the second by \bar{x}_1 , and we add them up, we get

$$\omega (\bar{x}_1^2 + \bar{x}_2^2) = 0;$$

Therefore, since $\omega \neq 0$, the only equilibrium solution is the trivial equilibrium $\bar{x}_1 = \bar{x}_2 = 0$. To proceed with the analysis we introduce polar coordinates, $(r; \mu)$, by using the transformation,

$$\begin{aligned} x_1 &= r \cos \mu; \\ x_2 &= r \sin \mu; \end{aligned}$$

The equations of motion are then written as

$$\begin{aligned} \dot{r} \cos \mu - r \dot{\mu} \sin \mu &= -\omega r \cos \mu - \eta r \sin \mu + ar^3 \cos \mu; \\ \dot{r} \sin \mu + r \dot{\mu} \cos \mu &= \eta r \cos \mu + \omega r \sin \mu + ar^3 \sin \mu; \end{aligned}$$

which reduce to

$$\begin{aligned} \dot{r} &= -\omega r + ar^3; \\ \dot{\mu} &= \eta r; \end{aligned}$$

It is clear that an equilibrium point, r , of the r equation will correspond to a limit cycle back in the original coordinates x_1 and x_2 . We can see that the r equation has equilibria given by

$$\mu r + ar^3 = 0 :$$

Let us assume that $a < 0$. Then for $\mu < 0$, the trivial equilibrium is stable. For $\mu > 0$ there is a stable limit cycle of radius proportional to the square root of μ , surrounding the unstable trivial equilibrium. If $a > 0$, then the limit cycle occurs for $\mu < 0$; it is unstable and surrounds a stable equilibrium point. The two cases are shown schematically in Figure 46. This resembles our pitchfork bifurcation of the previous section. Therefore, we can summarize our conclusions about the x_1, x_2 system as follows:

² If $a < 0$, then:

{ If $\mu < 0$ the trivial equilibrium is stable.

{ If $\mu > 0$ the trivial equilibrium is unstable, and a family of stable limit cycles with amplitude $\sqrt{\mu/a}$ exists.

² If $a > 0$, then:

{ If $\mu > 0$ the trivial equilibrium is unstable.

{ If $\mu < 0$ the trivial equilibrium is stable, and a family of unstable limit cycles with amplitude $\sqrt{\mu/a}$ exists.

We can see that the situation is similar to our pitchfork case; here we have the generation of periodic solutions except of equilibrium points. This bifurcation to periodic solutions is normally called the Poincaré-A ndronov-Hopf bifurcation. Analogously to the pitchfork case, we distinguish here the two major cases, supercritical and subcritical Hopf bifurcation. For more complicated systems, the reduction to the above two dimensional form and the computation of the leading nonlinear coefficient a which dictates limit cycle stability can be a significant undertaking.

8.4 Popov and Circle Criteria

Quite often, we need to analyze a control loop which contains a nonlinearity. Such a typical loop is shown in Figure 47. The two methods that we describe here enclose the nonlinearity in a linear envelope. The linear envelope rather than the particular nonlinearity is then used in the subsequent analysis. This approach leads to sufficient but not necessary stability conditions. Before proceeding to describe graphical techniques for the analysis of a feedback loop containing a nonlinearity, it is instructive to consider two celebrated conjectures, by two of the best minds of control theory.

1. The Aizerman and Kalman conjectures:

Aizerman postulated that the system of Figure 47 will be stable provided that the linear

system of Figure 48 is stable for all values of k in the interval $[k_1; k_2]$ where k_1, k_2 are defined by the relation

$$k_1 \leq \frac{N(e)}{e} \leq k_2 ;$$

for all $e \neq 0$. In this notation k_1, k_2 represent a linear envelope surrounding the nonlinearity, see Figure 51 where A stands for k_1 and B for k_2 . Aizerman's conjecture, reasonable as it might sound, is false as has been shown by counterexamples.

Kalman suggested that the system of Figure 47 will be stable provided that the linear system of Figure 48 is stable for all k in the interval $[\hat{k}_1; \hat{k}_2]$ where

$$\hat{k}_1 \leq \frac{dN(e)}{de} \leq \hat{k}_2 ;$$

and where

$$k_1 \leq \frac{N(e)}{e} \leq k_2 ;$$

and

$$\hat{k}_1 \leq k_1 \leq k_2 \leq \hat{k}_2 ;$$

Kalman's conjecture imposes additional requirements on the nonlinear characteristics but nevertheless it is also false — again shown by counterexamples. The failure of the two conjectures shows that intuitive reasoning cannot be relied on in nonlinear systems. One reason for the failure of the conjectures is that instabilities may arise in nonlinear systems due to the effects of harmonics. These are, of course, absent in linear systems. In the following, we discuss briefly two techniques for dealing with the problem of Figure 47. These two techniques, Popov's and circle criteria, can be viewed as extensions to Nyquist's stability criterion for linear systems.

2. Popov's stability criterion:

Popov developed a graphical Nyquist-like criterion to examine the stability of the loop shown in Figure 47. It is assumed that $G(s)$ is a stable transfer function. The nonlinearity $N(e)$ must be time-invariant and piecewise continuous function of e . The derivative $dN(e)/de$ must be bounded and $N(e)$ must satisfy the condition

$$0 < \frac{N(e)}{e} < k ;$$

for some positive constant k . Graphically, the last condition means that the curve representing N must lie within a particular linear envelope. A sufficient condition for global asymptotic stability of the feedback loop may then be stated as:

If there exists any real number q and an arbitrarily small number $\pm > 0$ such that

$$\operatorname{Re} \{ (1 + j! q)G(j!)g + \frac{1}{k} \} \geq \pm > 0 ;$$

for all $!$ then for any initial state the system output tends to zero as $t \rightarrow \infty$.

The proof can be found in most textbooks on nonlinear control and it makes use of Lyapunov's direct method.

To carry out a graphical test based on the above equation, a modified transfer function $G^*(j\omega)$ is defined by

$$G^*(j\omega) = (fG(j\omega)g + j\omega = fG(j\omega)g + X(j\omega) + jY(j\omega)) :$$

The criterion then, in terms of X and Y , becomes

$$X(j\omega) + qY(j\omega) + \frac{1}{k} \neq 0 :$$

The $G^*(j\omega)$ curve (the so called Popov locus) is plotted in the complex plane. The system is then stable if some straight line, at an arbitrary slope $1=q$, and passing through the $j\omega=0$ point avoids intersecting the $G^*(j\omega)$ locus. Figures 49 and 50 show two possible graphical results for stable and not necessarily stable situations respectively. Recall that the test gives a sufficient condition for stability and that the feedback loop whose result is given in Figure 50 is not necessarily unstable.

3. The circle method:

The circle method of stability analysis can be considered as a generalization of Popov's method. Compared with that method it has two important advantages:

1. It allows $G(s)$ to be open loop unstable;
2. It allows the nonlinearity to be time varying.

The nonlinearity N is assumed to lie within an envelope such that,

$$Ae < N(e;t) < Be ;$$

as shown in Figure 51. Then it is a sufficient condition for asymptotic stability that the Nyquist plot $G(j\omega)$ lies outside a circle in the complex plane that crosses the real axis at the points $\omega=1/A$ and $\omega=1/B$ and has its center at the point,

$$\frac{1}{2} \left(\frac{1}{A} + \frac{1}{B} \right) + \frac{j\omega}{2} \left(\frac{1}{A} - \frac{1}{B} \right) ;$$

for some real value of q . Here it is assumed that $A < B$.

This is the so called generalized circle criterion. Notice that the center of the circle depends on both frequency and choice of the value of q . In return for a loss of sharpness in the result (remembering that the method gives a sufficient criterion), q can be set equal to zero and then a single frequency invariant circle results (Figure 52). The circle can be considered as the generalization of the $(j\omega=0)$ point in the Nyquist test for linear systems.