MANEUVERING AND CONTROL OF MARINE VEHICLES

Michael S. Triantafyllou

Franz S. Hover

Department of Ocean Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts USA

Maneuvering and Control of Marine Vehicles
Latest Revision: November 5, 2003

Contents

1 KINEMATICS OF MOVING FRAMES

1.1 Rotation of Reference Frames

We denote through a subscript the specific reference system of a vector. Let a vector expressed in the inertial frame be denoted as $\vec{x}$, and in a body-reference frame $\vec{x}_b$. For the moment, we assume that the origins of these frames are coincident, but that the body frame has a different angular orientation. The angular orientation has several well-known descriptions, including the Euler angles and the Euler parameters (quaternions). The former method involves successive rotations about the principle axes, and has a solid link with the intuitive notions of roll, pitch, and yaw. One of the problems with Euler angles is that for certain specific values the transformation exhibits discontinuities. Quaternions present a more elegant and robust method, but with more abstraction. We will develop the equations of motion using Euler angles.

Tape three pencils together to form a right-handed three-dimensional coordinate system. Successively rotating the system about three of *its own* principal axes, it is easy to see that any possible orientation can be achieved. For example, consider the sequence of [yaw, pitch, roll]: starting from an orientation identical to some inertial frame, rotate the movable system about its yaw axis, then about the *new* pitch axis, then about the *newer still* roll axis. Needless to say, there are many valid Euler angle rotation sets possible to reach a given orientation; some of them might use the same axis twice.

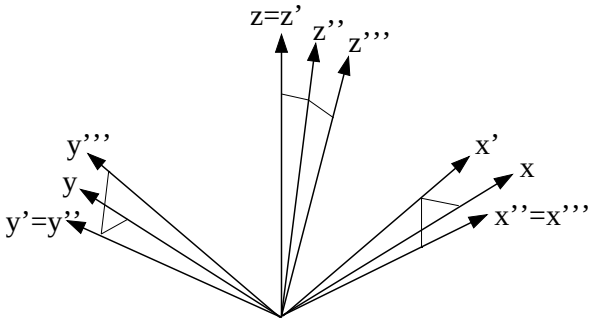


Figure 1: Successive application of three Euler angles transforms the original coordinate frame into an arbitrary orientation.

A first question is: what is the coordinate of a point fixed in inertial space, referenced to a rotated *body* frame? The transformation takes the form of a 3×3 matrix, which we now derive through successive rotations of the three Euler angles. Before the first rotation, the body-referenced coordinate matches that of the inertial frame: $\vec{x}_b^{\,0} = \vec{x}$. Now rotate the movable frame yaw axis (z) through an angle ϕ. We have

$$\vec{x}_b^{\,1} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \vec{x}_b^{\,0} = R(\phi)\vec{x}_b^{\,0}. \tag{1}$$

Rotation about the z-axis does not change the z-coordinate of the point; the other axes are modified according to basic trigonometry. Now apply the second rotation, pitch about the *new* y-axis by the angle θ:

$$\vec{x}_b^{\,2} = \begin{bmatrix} \cos\theta & 0 & -\sin\theta \\ 0 & 1 & 0 \\ \sin\theta & 0 & \cos\theta \end{bmatrix} \vec{x}_b^{\,1} = R(\theta)\vec{x}_b^{\,1}. \tag{2}$$

Finally, rotate the body system an angle ψ about its *newest* x-axis:

$$\vec{x}_b^{\,3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\psi & \sin\psi \\ 0 & -\sin\psi & \cos\psi \end{bmatrix} \vec{x}_b^{\,2} = R(\psi)\vec{x}_b^{\,2}. \tag{3}$$

This represents the location of the original point, in the fully-transformed body-reference frame, i.e., $\vec{x}_b^{\,3}$. We will use the notation $\vec{x}_b$ instead of $\vec{x}_b^{\,3}$ from here on. The three independent rotations can be cascaded through matrix multiplication (order matters!):

$$\begin{aligned} \vec{x}_b &= R(\psi)R(\theta)R(\phi)\vec{x} \\ &= \begin{bmatrix} c\theta c\phi & c\theta s\phi & -s\theta \\ -c\psi s\phi + s\psi s\theta c\phi & c\psi c\phi + s\psi s\theta s\phi & s\psi c\theta \\ s\psi s\phi + c\psi s\theta c\phi & -s\psi c\phi + c\psi s\theta s\phi & c\psi c\theta \end{bmatrix} \vec{x} \\ &= R(\phi,\theta,\psi)\vec{x}. \end{aligned} \tag{4}$$

All of the transformation matrices, including $R(\phi,\theta,\psi)$, are orthonormal: their inverse is equivalent to their transpose. Additionally, we should note that the rotation matrix R is universal to *all* representations of orientation, including quaternions. The roles of the trigonometric functions, as written, are specific to Euler angles, and to the order in which we performed the rotations.

In the case that the movable (body) reference frame has a different origin than the inertial frame, we have

$$\vec{x} = \vec{x}_0 + R^T\vec{x}_b, \tag{5}$$

where $\vec{x}_0$ is the location of the moving origin, expressed in inertial coordinates.

1.2 Differential Rotations

Now consider small rotations from one frame to another; using the small angle assumption to ignore higher-order terms gives

$$R \simeq \begin{bmatrix} 1 & \delta\phi & -\delta\theta \\ -\delta\phi & 1 & \delta\psi \\ \delta\theta & -\delta\psi & 1 \end{bmatrix} \tag{6}$$

$$= \begin{bmatrix} 0 & \delta\phi & -\delta\theta \\ -\delta\phi & 0 & \delta\psi \\ \delta\theta & -\delta\psi & 0 \end{bmatrix} + I_{3\times 3},$$

where $I_{3\times 3}$ donotes the identity matrix. R comprises the identity plus a part equal to the (negative) cross-product operator $[-\delta\vec{E}\times]$, where $\delta\vec{E} = [\delta\psi, \delta\theta, \delta\phi]$, the vector of Euler angles ordered with the axes $[x, y, z]$. Small rotations are completely decoupled; the order of the small rotations does not matter. Since $R^{-1} = R^T$, we have also $R^{-1} = I_{3\times 3} + \delta\vec{E}\times$;

$$\vec{x}_b = \vec{x} - \delta\vec{E} \times \vec{x} \tag{7}$$
$$\vec{x} = \vec{x}_b + \delta\vec{E} \times \vec{x}_b. \tag{8}$$

We now fix the point of interest on the *body*, instead of in inertial space, calling its location in the body frame $\vec{r}$ (radius). The differential rotations occur over a time step δt, so that we can write the location of the point before and after the rotation, with respect to the first frame as follows:

$$\begin{aligned} \vec{x}(t) &= \vec{r} \\ \vec{x}(t+\delta t) &= R^T\vec{r} = \vec{r} + \delta\vec{E} \times \vec{r}. \end{aligned} \tag{9}$$

Dividing by the differential time step gives

$$\begin{aligned} \frac{\delta\vec{x}}{\delta t} &= \frac{\delta\vec{E}}{\delta t} \times \vec{r} \\ &= \vec{\omega} \times \vec{r}, \end{aligned} \tag{10}$$

where the *rotation rate* vector $\omega \simeq d\vec{E}/dt$ because the Euler angles for this infinitesimal rotation are small and decoupled. This same cross-product relationship can be derived in the second frame as well:

$$\begin{aligned} \vec{x}_b(t) &= R\vec{r} = \vec{r} - \delta\vec{E} \times \vec{r} \\ \vec{x}_b(t+\delta t) &= \vec{r}. \end{aligned} \tag{11}$$

such that

$$\begin{aligned} \frac{\delta\vec{x}_b}{\delta t} &= \frac{\delta\vec{E}}{\delta t} \times \vec{r} \\ &= \vec{\omega} \times \vec{r}, \end{aligned} \tag{12}$$

On a rotating body whose origin point is fixed, the time rate of change of a constant radius vector is the cross-product of the rotation rate vector $\vec{\omega}$ and the radius vector itself. The resultant derivative is in the moving body frame.

In the case that the radius vector changes with respect to the body frame, we need an additional term:

$$\frac{d\vec{x}_b}{dt} = \vec{\omega} \times \vec{r} + \frac{\partial \vec{r}}{\partial t}. \tag{13}$$

Finally, allowing the origin to move as well gives

$$\frac{d\vec{x}_b}{dt} = \vec{\omega} \times \vec{r} + \frac{\partial \vec{r}}{\partial t} + \frac{d\vec{x}_o}{dt}. \tag{14}$$

This result is often written in terms of body-referenced velocity $\vec{v}$:

$$\vec{v} = \vec{\omega} \times \vec{r} + \frac{\partial \vec{r}}{\partial t} + \vec{v}_o, \tag{15}$$

where $\vec{v}_o$ is the body-referenced velocity of the origin. The total velocity of the particle is equal to the velocity of the reference frame origin, plus a component due to rotation of this frame. The velocity equation can be generalized to *any* body-referenced vector $\vec{f}$:

$$\frac{d\vec{f}}{dt} = \frac{\partial f}{\partial t} + \vec{\omega} \times \vec{f}. \tag{16}$$

1.3 Rate of Change of Euler Angles

Only for the case of infinitesimal Euler angles is it true that the time rate of change of the Euler angles equals the body-referenced rotation rate. For example, with the sequence [yaw,pitch,roll], the Euler yaw angle (applied first) is definitely not about the final body yaw axis; the pitch and roll rotations moved the axis. An important part of any simulation is the evolution of the Euler angles. Since the physics determine rotation rate $\vec{\omega}$, we seek a mapping $\vec{\omega} \to d\vec{E}/dt$.

The idea is to consider small changes in each Euler angle, and determine the effects on the rotation vector. The first Euler angle undergoes two additional rotations, the second angle one rotation, and the final Euler angle no additional rotations:

$$\begin{aligned} \vec{\omega} &= R(\psi)R(\theta) \begin{Bmatrix} 0 \\ 0 \\ \frac{d\phi}{dt} \end{Bmatrix} + R(\psi) \begin{Bmatrix} 0 \\ \frac{d\theta}{dt} \\ 0 \end{Bmatrix} + \begin{Bmatrix} \frac{d\psi}{dt} \\ 0 \\ 0 \end{Bmatrix} \\ &= \begin{bmatrix} 1 & 0 & -\sin\theta \\ 0 & \cos\psi & \sin\psi\cos\theta \\ 0 & -\sin\psi & \cos\psi\cos\theta \end{bmatrix} \begin{Bmatrix} \frac{d\psi}{dt} \\ \frac{d\theta}{dt} \\ \frac{d\phi}{dt} \end{Bmatrix}. \end{aligned} \tag{17}$$

Taking the inverse gives

$$\begin{aligned} \frac{d\vec{E}}{dt} &= \begin{bmatrix} 1 & \sin\psi\tan\theta & \cos\psi\tan\theta \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi/\cos\theta & \cos\psi/\cos\theta \end{bmatrix} \vec{\omega} \\ &= \Gamma(\vec{E})\vec{\omega}. \end{aligned} \tag{18}$$

Singularities exist in Γ at $\theta = \{\pi/2, 3\pi/2\}$, because of the division by $\cos\theta$, and hence this otherwise useful equation for propagating the angular orientation of a body fails when the vehicle rotates about the intermediate y-axis by ninety degrees. In applications where this is a real possibility, for example in orbiting satellites and robotic arms, quaternions provide a seamless mapping. For most ocean vessels, the singularity is acceptable, as long as it is not on the yaw axis!

1.4 Dead Reckoning

The measurement of heading and longitudinal speed gives rise to one of the oldest methods of navigation: dead reckoning. Quite simply, if the estimated longitudinal speed over ground is U, and the estimated heading is ϕ, ignoring the lateral velocity leads to the evolution of Cartesian coordinates:

$$\begin{aligned} \dot{x} &= U\cos\phi \\ \dot{y} &= U\sin\phi. \end{aligned}$$

Needless to say, currents and vehicle sideslip will cause this to be in error. Nonetheless, some of the most remarkable feats of navigation in history have depended on dead reckoning.

2 VESSEL INERTIAL DYNAMICS

We consider the rigid body dynamics with a coordinate system affixed on the body. A common frame for ships, submarines, and other marine vehicles has the body-referenced x-axis forward, y-axis to port (left), and z-axis up. This will be the sense of our body-referenced coordinate system here.

2.1 Momentum of a Particle

Since the body moves with respect to an inertial frame, dynamics expressed in the body-referenced frame need extra attention. First, linear momentum for a particle obeys the equality

$$\vec{F} = \frac{d}{dt}\left(m\vec{v}\right) \tag{19}$$

A rigid body consists of a large number of these small particles, which can be indexed. The summations we use below can be generalized to integrals quite easily. We have

$$\vec{F_i} + \vec{R_i} = \frac{d}{dt}\left(m_i\vec{v_i}\right), \tag{20}$$

where $\vec{F_i}$ is the external force acting on the particle and $\vec{R_i}$ is the net force exerted by all the other surrounding particles (internal forces). Since the collection of particles is not driven apart by the internal forces, we must have equal and opposite internal forces such that

$$\sum_{i=1}^{N} \vec{R}_i = 0. \tag{21}$$

Then summing up all the particle momentum equations gives

$$\sum_{i=1}^{N} \vec{F}_i = \sum_{i=1}^{N} \frac{d}{dt} \left(m_i \vec{v}_i \right). \tag{22}$$

Note that the particle velocities are *not* independent, because the particles are rigidly attached.
Now consider a body reference frame, with origin $\mathbf{0}$, in which the particle i resides at body-referenced radius vector $\vec{r}$; the body translates and rotates, and we now consider how the momentum equation depends on this motion.

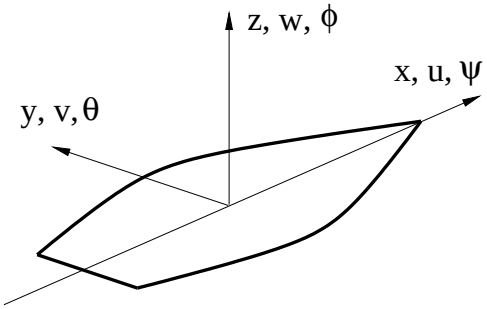


Figure 2: Convention for the body-referenced coordinate system on a vessel: x is forward, y is sway to the left, and z is heave upwards. Looking forward from the vessel bridge, roll about the x axis is positive counterclockwise, pitch about the y-axis is positive bow-down, and yaw about the z-axis is positive turning left.

2.2 Linear Momentum in a Moving Frame

The expression for total velocity may be inserted into the summed linear momentum equation to give

$$\begin{aligned} \sum_{i=1}^{N} \vec{F}_i &= \sum_{i=1}^{N} \frac{d}{dt} (m_i (\vec{v}_o + \vec{\omega} \times \vec{r}_i)) \\ &= m \frac{\partial \vec{v}_o}{\partial t} + \frac{d}{dt} \left[\vec{\omega} \times \sum_{i=1}^{N} m_i \vec{r}_i \right], \end{aligned} \tag{23}$$

where $m = \sum_{i=1}^{N} m_i$, and $\vec{v}_i = \vec{v}_o + \vec{\omega} \times \vec{r}_i$. Further defining the center of gravity vector $\vec{r}_G$ such that

$$m\vec{r}_G = \sum_{i=1}^{N} m_i \vec{r}_i, \tag{24}$$

we have

$$\sum_{i=1}^{N} \vec{F}_i = m\frac{\partial \vec{v}_o}{\partial t} + m\frac{d}{dt}(\vec{\omega} \times \vec{r}_G). \tag{25}$$

Using the expansion for total derivative again, the complete vector equation in body coordinates is

$$\vec{F} = \sum_{i=1} N = m\left(\frac{\partial \vec{v}_o}{\partial t} + \vec{\omega} \times \vec{v}_o + \frac{d\vec{\omega}}{dt} \times \vec{r}_G + \vec{\omega} \times (\vec{\omega} \times \vec{r}_G)\right). \tag{26}$$

Now we list some conventions that will be used from here on:

$$\begin{aligned}
\vec{v}_o &= \{u, v, w\} \text{ (body-referenced velocity)} \\
\vec{r}_G &= \{x_G, y_G, z_g\} \text{ (body-referenced location of center of mass)} \\
\vec{\omega} &= \{p, q, r\} \text{ (rotation vector, in body coordinates)} \\
\vec{F} &= \{X, Y, Z\} \text{ (external force, body coordinates).}
\end{aligned}$$

The last term in the previous equation simplifies using the vector triple product identity

$$\vec{\omega} \times (\vec{\omega} \times \vec{r}_G) = (\vec{\omega} \cdot \vec{r}_G)\vec{\omega} - (\vec{\omega} \cdot \vec{\omega})\vec{r}_G,$$

and the resulting three linear momentum equations are

$$\begin{aligned}
X &= m\left[\frac{\partial u}{\partial t} + qw - rv + \frac{dq}{dt}z_G - \frac{dr}{dt}y_G + (qy_G + rz_G)p - (q^2 + r^2)x_G\right] \\
Y &= m\left[\frac{\partial v}{\partial t} + ru - pw + \frac{dr}{dt}x_G - \frac{dp}{dt}z_G + (rz_G + px_G)q - (r^2 + p^2)y_G\right] \\
Z &= m\left[\frac{\partial w}{\partial t} + pv - qu + \frac{dp}{dt}y_G - \frac{dq}{dt}x_G + (px_G + qy_G)r - (p^2 + q^2)z_G\right].
\end{aligned} \tag{27}$$

Note that about half of the terms here are due to the mass center being in a different location than the reference frame origin, i.e., $\vec{r}_G \neq \vec{0}$.

2.3 Example: Mass on a String

Consider a mass on a string, being swung around around in a circle at speed U, with radius r. The centrifugal force can be computed in at least three different ways. The vector equation at the start is

$$\vec{F} = m\left(\frac{\partial \vec{v}_o}{\partial t} + \vec{\omega} \times \vec{v}_o + \frac{d\vec{\omega}}{dt} \times \vec{r}_G + \vec{\omega} \times (\vec{\omega} \times \vec{r}_G)\right).$$

2.3.1 Moving Frame Affixed to Mass

Affixing a reference frame *on* the mass, with the local x oriented forward and y inward towards the circle center, gives

$$\begin{aligned} \vec{v}_o &= \{U,0,0\}^T \\ \vec{\omega} &= \{0,0,U/r\}^T \\ \vec{r}_G &= \{0,0,0\}^T \\ \frac{\partial \vec{v}_o}{\partial t} &= \{0,0,0\}^T \\ \frac{\partial \vec{\omega}}{\partial t} &= \{0,0,0\}^T, \end{aligned}$$

such that

$$\vec{F} = m\vec{\omega} \times \vec{v}_o = m\{0, U^2/r, 0\}^T.$$

The force of the string pulls in on the mass to create the circular motion.

2.3.2 Rotating Frame Attached to Pivot Point

Affixing the moving reference frame to the pivot point of the string, with the same orientation as above but allowing it to rotate with the string, we have

$$\begin{aligned} \vec{v}_o &= \{0,0,0\}^T \\ \vec{\omega} &= \{0,0,U/r\}^T \\ \vec{r}_G &= \{0,r,0\}^T \\ \frac{\partial \vec{v}_o}{\partial t} &= \{0,0,0\}^T \\ \frac{\partial \vec{\omega}}{\partial t} &= \{0,0,0\}^T, \end{aligned}$$

giving the same result:

$$\vec{F} = m\vec{\omega} \times (\vec{\omega} \times \vec{r}_G) = m\{0, U^2/r, 0\}^T.$$

2.3.3 Stationary Frame

A frame fixed in inertial space, and momentarily coincident with the frame on the mass (2.3.1), can also be used for the calculation. In this case, as the string travels through a small arc $\delta\psi$, vector subtraction gives

$$\delta\vec{v} = \{0, U\sin\delta\psi, 0\}^T \simeq \{0, U\delta\psi, 0\}^T.$$

Since $\dot{\psi} = U/r$, it follows easily that in the fixed frame $d\vec{v}/dt = \{0, U^2/r, 0\}^T$, as before.

2.4 Angular Momentum

For angular momentum, the summed particle equation is

$$\sum_{i=1}^{N}(\vec{M}_i + \vec{r}_i \times \vec{F}_i) = \sum_{i=1}^{N} \vec{r}_i \times \frac{d}{dt}(m_i \vec{v}_i), \tag{28}$$

where $\vec{M}_i$ is an external moment on the particle i. Similar to the case for linear momentum, summed internal moments cancel. We have

$$\begin{aligned}\sum_{i=1}^{N}(\vec{M}_i + \vec{r}_i \times \vec{F}_i) &= \sum_{i=1}^{N} m_i \vec{r}_i \times \left[\frac{\partial \vec{v}_o}{\partial t} + \vec{\omega} \times \vec{v}_o\right] + \sum_{i=1}^{N} m_i \vec{r}_i \times \left(\frac{\partial \vec{\omega}}{\partial t} \times \vec{r}_i\right) + \\ &\quad \sum_{i=1}^{N} m_i \vec{r}_i \times (\vec{\omega} \times (\vec{\omega} \times \vec{r}_i)).\end{aligned}$$

The summation in the first term of the right-hand side is recognized simply as $m\vec{r}_G$, and the first term becomes

$$m\vec{r}_G \times \left[\frac{\partial \vec{v}_o}{\partial t} + \vec{\omega} \times \vec{v}_o\right]. \tag{29}$$

The second term expands as (using the triple product)

$$\begin{aligned}\sum_{i=1}^{N} m_i \vec{r}_i \times \left(\frac{\partial \vec{\omega}}{\partial t} \times \vec{r}_i\right) &= \sum_{i=1}^{N} m_i \left((\vec{r}_i \cdot \vec{r}_i)\frac{\partial \vec{\omega}}{\partial t} - \left(\frac{\partial \vec{\omega}}{\partial t} \cdot \vec{r}_i\right)\vec{r}_i\right) \\ &= \left\{ \begin{array}{l} \sum_{i=1}^{N} m_i \left((y_i^2 + z_i^2)\dot{p} - (y_i \dot{q} + z_i \dot{r})x_i\right) \\ \sum_{i=1}^{N} m_i \left((x_i^2 + z_i^2)\dot{q} - (x_i \dot{p} + z_i \dot{r})y_i\right) \\ \sum_{i=1}^{N} m_i \left((x_i^2 + y_i^2)\dot{r} - (x_i \dot{p} + y_i \dot{q})z_i\right) \end{array} \right\}.\end{aligned} \tag{30}$$

Employing the definitions of moments of inertia,

$$\begin{aligned} I &= \begin{bmatrix} I_{xx} & I_{xy} & I_{xz} \\ I_{yx} & I_{yy} & I_{yz} \\ I_{zx} & I_{zy} & I_{zz} \end{bmatrix} \qquad \text{(inertia matrix)} \\ I_{xx} &= \sum_{i=1}^{N} m_i (y_i^2 + z_i^2) \\ I_{yy} &= \sum_{i=1}^{N} m_i (x_i^2 + z_i^2) \\ I_{zz} &= \sum_{i=1}^{N} m_i (x_i^2 + y_i^2) \\ I_{xy} &= I_{yx} = -\sum_{i=1}^{N} m_i x_i y_i \qquad \text{(cross-inertia)} \end{aligned}$$

$$
\begin{aligned}
I_{xz} &= I_{zx} = -\sum_{i=1}^{N} m_i x_i z_i \\
I_{yz} &= I_{zy} = -\sum_{i=1}^{N} m_i y_i z_i,
\end{aligned}
$$

the second term of the angular momentum right-hand side collapses neatly into $I\partial\vec{\omega}/\partial t$. The third term can be worked out along the same lines, but offers no similar condensation:

$$
\begin{aligned}
\sum_{i=1}^{N} m_i \vec{r}_i \times ((\vec{\omega}\cdot\vec{r}_i)\vec{\omega} - (\vec{\omega}\cdot\vec{\omega})\vec{r}_i) &= \sum_{i=1}^{N} m_i \vec{r}_i \times \vec{\omega}(\vec{\omega}\cdot\vec{r}_i) \\
&= \left\{ \begin{array}{l} \sum_{i=1}^{N} m_i (y_i r - z_i q)(x_i p + y_i q + z_i r) \\ \sum_{i=1}^{N} m_i (z_i p - x_i r)(x_i p + y_i q + z_i r) \\ \sum_{i=1}^{N} m_i (x_i q - y_i p)(x_i p + y_i q + z_i r) \end{array} \right\} \\
&= \left\{ \begin{array}{l} I_{yz}(q^2 - r^2) + I_{xz} pq - I_{xy} pr \\ I_{xz}(r^2 - p^2) + I_{xy} rq - I_{yz} pq \\ I_{xy}(p^2 - q^2) + I_{yz} pr - I_{xz} qr \end{array} \right\} + \\
&\quad \left\{ \begin{array}{l} (I_{zz} - I_{yy}) rq \\ (I_{xx} - I_{zz}) rp \\ (I_{yy} - I_{xx}) qp \end{array} \right\}.
\end{aligned} \tag{31}
$$

Letting $\vec{M} = \{K, M, N\}$ be the total moment acting on the body, i.e., the left side of Equation 28, the complete moment equations are

$$
\begin{aligned}
K &= I_{xx}\dot{p} + I_{xy}\dot{q} + I_{xz}\dot{r} + \\
&\quad (I_{zz} - I_{yy}) rq + I_{yz}(q^2 - r^2) + I_{xz} pq - I_{xy} pr + \\
&\quad m\left[y_G(\dot{w} + pv - qu) - z_G(\dot{v} + ru - pw)\right]
\end{aligned} \tag{32}
$$

$$
\begin{aligned}
M &= I_{yx}\dot{p} + I_{yy}\dot{q} + I_{yz}\dot{r} + \\
&\quad (I_{xx} - I_{zz}) pr + I_{xz}(r^2 - p^2) + I_{xy} qr - I_{yz} qp + \\
&\quad m\left[z_G(\dot{u} + qw - rv) - x_G(\dot{w} + pv - qu)\right]
\end{aligned}
$$

$$
\begin{aligned}
N &= I_{zx}\dot{p} + I_{zy}\dot{q} + I_{zz}\dot{r} + \\
&\quad (I_{yy} - I_{xx}) pq + I_{xy}(p^2 - q^2) + I_{yz} pr - I_{xz} qr + \\
&\quad m\left[x_G(\dot{v} + ru - pw) - y_G(\dot{u} + qw - rv)\right].
\end{aligned}
$$

2.5 Example: Spinning Book

Consider a homogeneous rectangular block with $I_{xx} < I_{yy} < I_{zz}$ and all off-diagonal moments of inertia are zero. The linearized angular momentum equations, with no external forces or moments, are

$$I_{xx}\frac{dp}{dt} + (I_{zz} - I_{yy})rq = 0$$
$$I_{yy}\frac{dq}{dt} + (I_{xx} - I_{zz})pr = 0$$
$$I_{zz}\frac{dr}{dt} + (I_{yy} - I_{xx})qp = 0.$$

We consider in turn the stability of rotations about each of the main axes, with constant angular rate Ω. The interesting result is that rotations about the x and z axes are stable, while rotation about the y axis is not. This is easily demonstrated experimentally with a book or a tennis racket.

2.5.1 x-axis

In the case of the x-axis, $p = \Omega + \delta p$, $q = \delta q$, and $r = \delta r$, where the δ prefix indicates a small value compared to Ω. The first equation above is uncoupled from the others, and indicates no change in δp, since the small term $\delta q \delta r$ can be ignored. Differentiate the second equation to obtain

$$I_{yy}\frac{\partial^2 \delta q}{\partial t^2} + (I_{xx} - I_{zz})\Omega\frac{\partial \delta r}{\partial t} = 0$$

Substitution of this result into the third equation yields

$$I_{yy}I_{zz}\frac{\partial^2 \delta q}{\partial t^2} + (I_{xx} - I_{zz})(I_{xx} - I_{yy})\Omega^2 \delta q = 0.$$

A simpler expression is $\delta\ddot{q} + \alpha\delta q = 0$, which has response $\delta q(t) = \delta q(0)e^{\sqrt{-\alpha}t}$, when $\delta\dot{q}(0) = 0$. For spin about the x-axis, both coefficients of the differential equation are positive, and hence $\alpha > 0$. The imaginary exponent indicates that the solution is of the form $\delta q(t) = \delta q(0) cos\sqrt{\alpha}t$, that is, it oscillates but does not grow. Since the perturbation δr is coupled, it too oscillates.

2.5.2 y-axis

Now suppose $q = \Omega + \delta q$: differentiate the first equation and substitute into the third equation to obtain

$$I_{zz}I_{xx}\frac{\partial^2 \delta p}{\partial t^2} + (I_{yy} - I_{xx})(I_{yy} - I_{zz})\Omega^2 \delta p = 0.$$

Here the second coefficient has negative sign, and therefore $\alpha < 0$. The exponent is real now, and the solution grows without bound, following $\delta p(t) = \delta p(0)e^{\sqrt{-\alpha}t}$.

2.5.3 z-axis

Finally, let $r = \Omega + \delta r$: differentiate the first equation and substitute into the second equation to obtain

$$I_{yy} I_{xx} \frac{\partial^2 \delta p}{\partial t^2} + (I_{xx} - I_{zz})(I_{yy} - I_{zz})\Omega^2 \delta p = 0.$$

The coefficients are positive, so bounded oscillations occur.

2.6 Parallel Axis Theorem

Often, the mass center of an body is at a different location than a more convenient measurement point, the geometric center of a vessel for example. The parallel axis theorem allows one to translate the mass moments of inertia referenced to the mass center into another frame with parallel orientation, and vice versa. Sometimes a translation of coordinates to the mass center will make the cross-inertial terms I_{xy}, I_{yz}, I_{xz} small enough that they can be ignored; in this case $\vec{r}_G = \vec{0}$ also, so that the equations of motion are significantly reduced, as in the spinning book example.
The formulas are:

$$\begin{aligned} I_{xx} &= \bar{I}_{xx} + m(\delta y^2 + \delta z^2) \\ I_{yy} &= \bar{I}_{yy} + m(\delta x^2 + \delta z^2) \\ I_{zz} &= \bar{I}_{zz} + m(\delta x^2 + \delta y^2) \\ I_{yz} &= \bar{I}_{yz} - m\delta y \delta z \\ I_{xz} &= \bar{I}_{xz} - m\delta x \delta z \\ I_{xy} &= \bar{I}_{xy} - m\delta x \delta y, \end{aligned} \tag{33}$$

where $\bar{I}$ represents an MMOI in the axes of the mass center, and δx, for example, is the translation of the x-axis to the new frame. Note that translation of MMOI using the parallel axis theorem *must* be either to or from a frame resting exactly at the center of gravity.

2.7 Basis for Simulation

Except for external forces and moments $\vec{F}$ and $\vec{M}$, we now have the necessary terms for writing a full nonlinear simulation of a rigid body, in body coordinates. There are twelve states, comprising the following components:

- $\vec{v}_o$, the vector of body-referenced velocities.
- $\vec{\omega}$, body rotation rate vector.
- $\vec{x}$, location of the body origin, in *inertial* space.

- $\vec{E}$, Euler angle vector.

The derivatives of body-referenced velocity and rotation rate come from Equations 27 and 32, with some coupling which generally requires a 6×6 matrix inverse. The Cartesian position propagates according to

$$\dot{\vec{x}} = R^T(\vec{E})\vec{v}_o, \tag{34}$$

while the Euler angles follow:

$$\dot{\vec{E}} = \Gamma(\vec{E})\vec{\omega}. \tag{35}$$

3 NONLINEAR COEFFICIENTS IN DETAIL

The method of hydrodynamic coefficients is a somewhat blind series expansion of the fluid force in an attempt to provide a framework on which to base experiments and calculations to evaluate these terms. The basic dificulty, i.e. the intractability of the governing equations of motion of viscous fluid prohibits, at least today and in the near future, a computation of these forces.
Still, we are not totally ignorant about these forces, since a number of symmetries and basic laws can be applied to reduce the number of unknown coefficients. This is the purpose of this section.
The basic assumption in using the method of hydrodynamic coefficients is that the forces have no memory effects, i.e. past motions have no impact on the fluid forces at the present moment. This is not correct when the flow separates, or when large vortices are shed, because then the vorticity in the fluid affects the fluid forces for a considerable time after they have been shed – i.e. until they move sufficiently far away from the body. We will show later some methods which allow us to incorporate the effect of shed vorticity, because under certain conditions such effects can not be ignored.
We employ the following basic facts and assumptions to derive the fluid forces acting on a ship, submarine or vehicle:

1. We retain only first order acceleration terms. Based on Newton's second law, we expect the inertia terms from the fluid to be linearly dependent on acceleration.

2. We do not include terms coupling velocities and accelerations. Again, based on Newton's second law, we expect inertia forces to depend on acceleration alone.

3. We consider port/starboard symmetry. Unless there is a reason not to, this is a useful property to us in order to eliminate a certain number of coefficients which are either zero or very small. In ships, a propeller introduces an asymmetry port/starboard since it rotates in a certain direction (unless it is a twin-screw ship with counter-rotating propellers), but in such cases we limit this asymmetry to only propeller-related terms.

4. We retain up to third order terms. This is a practical consideration and has been found to, in general, serve well the purpose of deriving equations which are sufficiently accurate in a wide parametric range. It does not constitute an absolute rule, but most existing models employ this assumption

Finally, we find sometimes necessary to use coefficients such as $Y_{|v|v}$, providing a drag-related term, whose strict definition would be:

$$Y_{|v|v} = \frac{\partial^2 Y}{\partial v \partial |v|}(v = 0) \tag{36}$$

A Taylor series expansion would not include such terms, so their inclusion is motivated by physical arguments.

3.1 Helpful Facts

To exploit symmetries, we consider the following simple facts:

- A function $f(x)$ which is symmetric in x has zero odd derivatives with respect to x at $x = 0$. The proof is to consider the Taylor series expansion of $f(x)$ about $x = 0$:

$$f(x) = f(0) + \frac{df}{dx}(0)x + \frac{1}{2!}\frac{d^2 f}{dx^2}(0)x^2 + \frac{1}{3!}\frac{d^3 f}{dx^3}(0)x^3 + \quad ... \tag{37}$$

  Symmetry gives that for all x:

$$f(x) = f(-x) \tag{38}$$

  The only way that (38) is satisfied given the expression (37) is that all odd derivatives of $f(x)$ must be zero, i.e., for n odd:

$$\frac{d^n f}{dx^n}(0) = 0 \tag{39}$$

- A function $f(x)$ which is anti-symmetric in x has zero even derivatives with respect to x at $x = 0$. The proof is exactly analogous to the result above, using the anti-symmetry condition:

$$f(x) = -f(-x) \tag{40}$$

  to find that, for n even:

$$\frac{d^n f}{dx^n}(0) = 0 \tag{41}$$

3.2 Nonlinear Equations in the Horizontal Plane

To demonstrate the methodology, we will derive the governing nonlinear equations of motion in the horizontal plane (surge, sway and yaw), employing the assumptions above. We will not include drag related terms of the form of equation 36 for the time being. We start by re-stating the inertia terms driven by external forces, which include the fluid forces:

$$\begin{aligned} m(\dot{u} - vr - x_G r^2) &= X \\ m(\dot{v} + ur + x_G \dot{r}) &= Y \\ I_{zz}\dot{r} + m x_G(\dot{v} + ur) &= N \end{aligned} \tag{42}$$

where we have assumed that $w = 0$, $p = 0$, $q = 0$, $y_G = 0$, $z_G = 0$.

3.2.1 Fluid Force X

By denoting the rudder angle as δ, we derive the following expression for the fluid force X, valid up to third order:

$$\begin{aligned} X = {} & X_e + X_{\dot{u}}\dot{u} + X_u u + X_{uu}u^2 + X_{uuu}u^3 + X_{vv}v^2 + X_{rr}r^2 + X_{\delta\delta}\delta^2 + \\ & X_{rv}rv + X_{r\delta}r\delta + X_{v\delta}v\delta + X_{vvu}v^2u + + X_{rru}r^2u + X_{\delta\delta u}\delta^2 u + \\ & X_{r\delta u}r\delta u + X_{rvu}rvu + + X_{v\delta u}v\delta u + X_{rv\delta}rv\delta \end{aligned} \tag{43}$$

We have used three basic properties, i.e., that the fluid force X, independent of the forward velocity, must be:

1. a symmetric function of v when $r = 0$ and $\delta = 0$;
2. a symmetric function of r when $v = 0$ and $\delta = 0$;
3. a symmetric function of δ when $r = 0$ and $v = 0$;

This is a result of port/starboard symmetry and is expressed as:

$$X(u, v, r = 0, \delta = 0) = X(u, -v, r = 0, \delta = 0) \tag{44}$$
$$X(u, v = 0, r, \delta = 0) = X(u, v = 0, -r, \delta = 0) \tag{45}$$
$$X(u, v = 0, r = 0, \delta) = X(u, v = 0, r = 0, -\delta) \tag{46}$$

These relations imply that all odd derivatives of X with respect to v at $v = 0$ are zero, when $r = 0$ and $\delta = 0$; and similarly for r and δ. For example:

$$\frac{\partial^3 X}{\partial v^3}(u, v = 0, r = 0, \delta = 0) = 0 \tag{47}$$

which implies that $X_{vvv} = 0$. Also, since relations such as (44) hold for all forward velocities u, it means that derivatives with respect to u of expression (44) are aso true. For example, from (47) we derive:

$$\frac{\partial^4 X}{\partial v^3 \partial u}(u, v = 0, r = 0, \delta = 0) = 0 \tag{48}$$

or equivalently $X_{vvvu} = 0$.
In summary, the symmetries provide the following zero coefficients:

$$\begin{array}{llll} X_v = 0; & X_{vvv} = 0; & X_{vu} = 0; & X_{vuu} = 0; \\ X_r = 0; & X_{rrr} = 0; & X_{ru} = 0; & X_{ruu} = 0; \\ X_\delta = 0; & X_{\delta\delta\delta} = 0; & X_{\delta u} = 0; & X_{\delta uu} = 0 \end{array} \tag{49}$$

3.2.2 Fluid Force Y

In the case of the fluid foce Y, the symmetry implies that this force must be an antisymmetric function of v when $r = 0$, $\delta = 0$; and likewise for r and δ, i.e.:

$$\begin{aligned} Y(u, v, r = 0, \delta = 0) &= -Y(u, -v, r = 0, \delta = 0) \\ Y(u, v = 0, r, \delta = 0) &= -Y(u, v = 0, -r, \delta = 0) \\ Y(u, v = 0, r = 0, \delta) &= -Y(u, v = 0, r = 0, -\delta) \end{aligned} \tag{50}$$

Hence, in analogy with the previous section, the even derivatives of Y with respect to v (and then r, and δ) must be zero, or:

$$\begin{array}{ll} Y_{vv} = 0; & Y_{vvu} = 0 \\ Y_{rr} = 0; & Y_{rru} = 0 \\ Y_{\delta\delta} = 0; & Y_{\delta\delta u} = 0 \end{array} \tag{51}$$

Finally, due to port/starboard symmetry, the force Y should not be affected by u when $v = 0$, $r = 0$, $\delta = 0$, except for propeller effects, which break the symmetry. For this reason, we will include terms such as Y_u to allow for the propeller asymmetry (twin propeller ships with counter-rotating propellers have zero such terms).
Finally, we derive the following expansion for Y:

$$\begin{aligned} Y = {} & Y_e + Y_u u + Y_{uu} u^2 + Y_{\dot{v}} \dot{v} + Y_{\dot{r}} \dot{r} + Y_v v + Y_r r + Y_\delta \delta + Y_{\delta u} \delta u + Y_{vu} vu + \\ & Y_{ru} ru + Y_{vuu} vu^2 + Y_{ruu} ru^2 + Y_{\delta uu} \delta u^2 + Y_{vvv} v^3 + Y_{rrr} r^3 + Y_{\delta\delta\delta} \delta^3 + \\ & Y_{rr\delta} r^2 \delta + Y_{rrv} r^2 v + Y_{vvr} v^2 r + Y_{vv\delta} v^2 \delta + Y_{vr\delta} vr\delta + Y_{\delta\delta r} \delta^2 r + Y_{\delta\delta v} \delta^2 v \end{aligned} \tag{52}$$

3.2.3 Fluid Moment N

The derivation for N follows the same exact steps as for the side force Y, i.e. the same symmetries apply. As a result, we first find that the following coefficients are zero:

$$\begin{aligned} N_{vv} = 0; \quad & N_{vvu} = 0 \\ N_{rr} = 0; \quad & N_{rru} = 0 \\ N_{\delta\delta} = 0; \quad & N_{\delta\delta u} = 0 \end{aligned} \tag{53}$$

Hence, we derive an analogous expansion for the moment:

$$\begin{aligned} N \;=\; & N_e + N_u u + N_{uu}u^2 + N_{\dot{v}}\dot{v} + N_{\dot{r}}\dot{r} + N_v v + N_r r + N_\delta \delta + N_{\delta u}\delta u + \\ & N_{vu}vu + N_{ru}ru + N_{vuu}vu^2 + N_{ruu}ru^2 + N_{\delta uu}\delta u^2 + N_{vvv}v^3 + N_{rrr}r^3 + \\ & N_{\delta\delta\delta}\delta^3 + N_{rr\delta}r^2\delta + N_{rrv}r^2 v + N_{vvr}v^2 r + N_{vv\delta}v^2\delta + N_{vr\delta}vr\delta + \\ & N_{\delta\delta r}\delta^2 r + N_{\delta\delta v}\delta^2 v. \end{aligned} \tag{54}$$

4 VESSEL DYNAMICS: LINEAR CASE

4.1 Surface Vessel Linear Model

We first discuss some of the hydrodynamic parameters which govern a ship maneuvering in the horizontal plane. The body x-axis is forward and the y-axis is to port, so positive r has the vessel turning left. We will consider motions only in the horizontal plane, which means $\theta = \psi = p = q = w = 0$. Since the vessel is symmetric about the $x - z$ plane, $y_G = 0$; z_G is inconsequential. We then have at the outset

$$\begin{aligned} X &= m\left(\frac{\partial u}{\partial t} - rv - x_G r^2\right) \\ Y &= m\left(\frac{\partial v}{\partial t} + ru + x_G\frac{\partial r}{\partial t}\right) \\ N &= I_{zz}\frac{\partial r}{\partial t} + mx_G\left(\frac{\partial v}{\partial t} + ru\right). \end{aligned} \tag{55}$$

Letting $u = U + u$, where $U >> u$, and eliminating higher-order terms, this set is

$$\begin{aligned} X &= m\frac{\partial u}{\partial t} \\ Y &= m\left(\frac{\partial v}{\partial t} + rU + x_G\frac{\partial r}{\partial t}\right) \\ N &= I_{zz}\frac{\partial r}{\partial t} + mx_G\left(\frac{\partial v}{\partial t} + rU\right). \end{aligned} \tag{56}$$

A number of coefficients can be discounted, as noted in the last chapter. First, in a homogeneous sea, with no current, wave, or wind effects, $\{X_x, X_y, X_\phi, Y_x, Y_y, Y_\phi, N_x, N_y, N_\phi\}$ are all zero. We assume that no hydrodynamic forces depend on the position of the vessel.[1] Second, consider X_v: since this longitudinal force would have the same sign regardless of the sign of v (because of side-to-side hull symmetry), it must have zero slope with v at the origin. Thus $X_v = 0$. The same argument shows that $\{X_r, X_{\dot{v}}, X_{\dot{r}}, Y_u, Y_{\dot{u}}, N_u, N_{\dot{u}}\} = 0$. Finally, since fluid particle acceleration relates linearly with pressure or force, we do not consider nonlinear acceleration terms, or higher time derivatives. It should be noted that some nonlinear terms related to those we have eliminated above are *not* zero. For instance, $Y_{uu} = 0$ because of hull symmetry, but in general $X_{vv} = 0$ only if the vessel is bow-stern symmetric.
We have so far, considering only the linear hydrodynamic terms,

$$\begin{aligned}
(m - X_{\dot{u}})\dot{u} &= X_u u + X' && (57)\\
(m - Y_{\dot{v}})\dot{v} + (mx_G - Y_{\dot{r}})\dot{r} &= Y_v v + (Y_r - mU)r + Y' && (58)\\
(mx_G - N_{\dot{v}})\dot{v} + (I_{zz} - N_{\dot{r}})\dot{r} &= N_v v - (N_r - mx_G U)r + N'. && (59)
\end{aligned}$$

The right side here carries also the imposed forces from a thruster(s) and rudder(s) $\{X', Y', N'\}$. Note that the surge equation is *decoupled* from the sway and yaw, but that sway and yaw themselves are coupled, and therefore are of immediate interest. With the state vector $\vec{s} = \{v, r\}$ and external force/moment vector $\vec{F} = \{Y', N'\}$, a state-space representation of the sway/yaw system is

$$\begin{aligned}
\begin{bmatrix} m - Y_{\dot{v}} & mx_G - Y_{\dot{r}} \\ mx_G - N_{\dot{v}} & I_{zz} - N_{\dot{r}} \end{bmatrix} \frac{d\vec{s}}{dt} &= \begin{bmatrix} Y_v & Y_r - mU \\ N_v & N_r - mx_G U \end{bmatrix} \vec{s} + \vec{F}, \text{ or} && (60)\\
M\dot{\vec{s}} &= P\vec{s} + \vec{F} \\
\dot{\vec{s}} &= M^{-1}P\vec{s} + M^{-1}\vec{F} \\
\dot{\vec{s}} &= A\vec{s} + B\vec{F}. && (61)
\end{aligned}$$

The matrix M is a mass or inertia matrix, which is always invertible. The last form of the equation is a standard one wherein A represents the internal dynamics of the system, and B is a gain matrix for the control and disturbance inputs.

4.2 Stability of the Sway/Yaw System

Consider the homogeneous system $\dot{\vec{s}} = A\vec{s}$:

$$\begin{aligned}
\dot{s}_1 &= A_{11}s_1 + A_{12}s_2 \\
\dot{s}_2 &= A_{21}s_1 + A_{22}s_2.
\end{aligned}$$

[1] Note that the linearized heave/pitch dynamics of a submarine do depend on the pitch angle; this topic will be discussed later.

We can rewrite the second equation as

$$s_2 = \left(\frac{d(\cdot)}{dt} - A_{22} \right)^{-1} A_{21} s_1 \tag{62}$$

and substitute into the first equation to give

$$\ddot{s}_1 + (-A_{11} - A_{22})\dot{s}_1 + (A_{11}A_{22} - A_{12}A_{21})s_1 = 0. \tag{63}$$

Note that these operations are allowed because the derivative operator is linear; in the language of the Laplace transform, we would simply use s. A necessary and sufficient condition for stability of this ODE system is that each coefficient must be greater than zero:

$$\begin{aligned} -A_{11} - A_{22} &> 0 \\ A_{11}A_{22} - A_{12}A_{21} &> 0 \end{aligned} \tag{64}$$

The components of A for the sway/yaw problem are

$$\begin{aligned} A_{11} &= \frac{(I_{zz} - N_{\dot{r}})Y_v + (Y_{\dot{r}} - mx_G)N_v}{(m - Y_{\dot{v}})(I_{zz} - N_{\dot{r}}) - (mx_G - Y_{\dot{r}})(mx_G - N_{\dot{v}})} \\ A_{12} &= \frac{-(I_{zz} - N_{\dot{r}})(mU - Y_r) - (Y_{\dot{r}} - mx_G)(mx_G U - N_r)}{(m - Y_{\dot{v}})(I_{zz} - N_{\dot{r}}) - (mx_G - Y_{\dot{r}})(mx_G - N_{\dot{v}})} \\ A_{21} &= \frac{(N_{\dot{v}} - mx_G)Y_v + (m - Y_{\dot{v}})N_v}{(m - Y_{\dot{v}})(I_{zz} - N_{\dot{r}}) - (mx_G - Y_{\dot{r}})(mx_G - N_{\dot{v}})} \\ A_{22} &= \frac{-(N_{\dot{v}} - mx_G)(mU - Y_r) - (m - Y_{\dot{v}})(mx_G U - N_r)}{(m - Y_{\dot{v}})(I_{zz} - N_{\dot{r}}) - (mx_G - Y_{\dot{r}})(mx_G - N_{\dot{v}})}. \end{aligned} \tag{65}$$

The denominators are identical, and can be simplified. First, let $x_G \simeq 0$; valid for many vessels with the origin is at the geometric center. If the origin is at the center of mass, $x_G = 0$. Next, if the vessel is reasonably balanced with regard to forward and aft areas with respect to the origin, the terms $\{N_{\dot{v}}, Y_{\dot{r}}, N_v, Y_r\}$ take very small values in comparison with the others. To wit, the added mass term $-Y_{\dot{v}}$ is of the order of the vessel's material mass m, and similarly $N_{\dot{r}} \simeq -I_{zz}$. Both $Y_{\dot{v}}$ and $N_{\dot{r}}$ take large negative values. Linear drag and rotational drag are significant also; these are the terms Y_v and N_r, both large and negative. The denominator for A's components reduces to $(m - Y_{\dot{v}})(I_{zz} - N_{\dot{r}})$, and

$$\begin{aligned} A_{11} &= \tfrac{Y_v}{m - Y_{\dot{v}}} &< 0 \\ A_{22} &= \tfrac{N_r}{I_{zz} - N_{\dot{r}}} &< 0. \end{aligned}$$

Hence the first condition for stability is met: $-A_{11} - A_{22} > 0$. For the second condition, since the denominators of the A_{ij} are identical, we have only to look at the numerators. For stability, we require

$$\begin{aligned}(I_{zz} - N_{\dot{r}})Y_v(m - Y_{\dot{v}})N_r \\ -\left[N_{\dot{v}}Y_v + (m - Y_{\dot{v}})N_v\right]\left[-(I_{zz} - N_{\dot{r}})(mU - Y_r) + Y_{\dot{r}}N_r\right] \ &> 0.\end{aligned} \tag{66}$$

The first term is the product of two large negative and two large positive numbers. The second part of the second term contains mU, which has a large positive value, generally making stability critical on the (usually negative) N_v. When only the largest terms are considered for a vessel, a simpler form is common:

$$C = Y_v N_r + N_v(mU - Y_r) > 0. \tag{67}$$

C is called the vessels *stability parameter*. The terms of C compete, and yaw/sway stability depends closely on the magnitude and sign of N_v. Adding more surface area aft drives N_v more positive, increasing stability as expected. Stability can also be improved by moving the center of gravity forward. Nonzero x_G shows up as follows:

$$C = Y_v(N_r - mx_G U) + N_v(mU - Y_r) > 0. \tag{68}$$

Since N_r and Y_v are both negative, positive x_G increases the (positive) influence of C's first term.

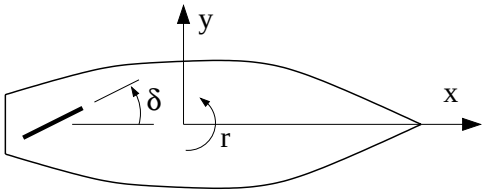


Figure 3: Convention for positive rudder angle in the vessel reference system.

4.3 Basic Rudder Action in the Sway/Yaw Model

Rudders are devices which develop large lift forces due to an angle of attack with respect to the oncoming fluid. The basic form is as follows: $L = \frac{1}{2}\rho A U^2 C_l(\alpha)$, where α is the angle of attack. The lift coefficient C_l is normally linear with α near $\alpha = 0$, but the rudder stalls when the angle of attack reaches a critical value, and thereafter develops much less lift. We will assume that α is small enough that the linear relationship applies:

$$C_l(\alpha) = \left.\frac{\partial C_l}{\partial \alpha}\right|_{\alpha=0} \alpha. \tag{69}$$

Since the rudder develops force (and a small moment) far away from the body origin, say a distance l aft, the moment equation is quite simple. We have

$$Y_\alpha = \frac{1}{2}\rho A \left.\frac{\partial C_l}{\partial \alpha}\right|_{\alpha=0} U^2 \tag{70}$$

$$N_\alpha = -\frac{1}{2}\rho A \left.\frac{\partial C_l}{\partial \alpha}\right|_{\alpha=0} lU^2. \tag{71}$$

Note the difference between the rudder angle expressed in the body frame, δ, and the total angle of attack α. Angle of attack is influenced by δ, as well as v/U and lr. Thus, in tank testing with $v = 0$, $\delta = \alpha$ and $N_\delta = N_\alpha$, etc., but in real conditions, other hydrodynamic derivatives are augmented to capture the necessary effects, for example N_v and N_r. Generally speaking, the hydrodynamic characteristics of the vessel depend strongly on the rudder, even when $\delta = 0$. In this case the rudder still opposes yaw and sway perturbations and acts to stabilize the vessel.
A positive rudder deflection (defined to have the same sense as the yaw angle) causes a negative yaw perturbation, and a very small positive sway perturbation.

4.3.1 Adding Yaw Damping through Feedback

The stability coefficient C resulting from the addition of a control law $\delta = k_r r$, where $k_r > 0$ is a feedback gain, is

$$C = Y_v(N_r - mx_G U + k_r N_\delta) + N_v(mU - Y_r - k_r Y_\delta). \tag{72}$$

Y_δ is small positive, but N_δ is large and negative. Hence C becomes more positive, since Y_v is negative.
Control system limitations and the stalling of rudders make obvious the fact that even a very large control gain k_r cannot completely solve stability problems of a poorly-designed vessel with an inadequate rudder. On the other hand, a vessel which is overly stable ($C >> 0$ with no rudder action) is unmaneuverable. A properly-balanced vessel just achieves stability with zero rudder action, so that a reasonable amount of control will provide good maneuvering capabilities.

4.3.2 Heading Control in the Sway/Yaw Model

Considering just the yaw equation of motion, i.e., $v = 0$, with a rudder, we have

$$(I_{zz} - N_{\dot{r}})\ddot{\phi} + (mx_G U - N_r)\dot{\phi} = N_\delta \delta. \tag{73}$$

Employing the control law $\delta = k_\phi \phi$, the system equation becomes a homogeneous, second-order ODE:

$$(I_{zz} - N_{\dot{r}})\ddot{\phi} + (mx_G U - N_r)\dot{\phi} - N_\delta k_\phi \phi = 0. \tag{74}$$

Since all coefficients are positive (recall $N_\delta < 0$), the equation gives a stable θ response, settling under second-order dynamics to $\theta(\infty) = 0$. The control law $\delta = k_\phi(\phi - \phi_{desired}) + k_r r$ is the basis for heading autopilots, which are used to track $\phi_{desired}$. This use of an error signal

to drive an actuator is in fact the essence of feedback control. In this case, we require *sensors* to obtain r and ϕ, a *controller* to calculate δ, and an *actuator* to implement the corrective action.

4.4 Response of the Vessel to Step Rudder Input

4.4.1 Phase 1: Accelerations Dominate

When the rudder first moves, acceleration terms dominate, since the velocities are zero. The equation looks like this:

$$\begin{bmatrix} m - Y_{\dot{v}} & mx_G - Y_{\dot{r}} \\ mx_G - N_{\dot{v}} & I_{zz} - N_{\dot{r}} \end{bmatrix} \begin{Bmatrix} \dot{v} \\ \dot{r} \end{Bmatrix} = \begin{Bmatrix} Y_\delta \\ N_\delta \end{Bmatrix} \delta. \tag{75}$$

Since $Y_{\dot{r}}$ and mx_G are comparatively small in the first row, we have

$$\dot{v}(0) = \frac{Y_\delta \delta}{m - Y_{\dot{v}}}, \tag{76}$$

and the vessel moves to the left, the positive v-direction. The initial yaw is in the negative r-direction, since $N_\delta < 0$:

$$\dot{r}(0) = \frac{N_\delta \delta}{I_{zz} - N_{\dot{r}}}. \tag{77}$$

The first phase is followed by a period (**Phase 2**), in which many terms are competing and contributing to the transient response.

4.4.2 Phase 3: Steady State

When the transients have decayed, the vessel is in a steady turning condition, and the accelerations are zero. The system equations reduce to

$$\begin{Bmatrix} v \\ r \end{Bmatrix} = \frac{\delta}{C} \begin{Bmatrix} (mx_G U - N_r)Y_\delta + (Y_r - mU)N_\delta \\ N_v Y_\delta - Y_v N_\delta \end{Bmatrix}. \tag{78}$$

Note that the denominator is the stability parameter. The steady turning rate is thus approximated by

$$r = -\frac{Y_v N_\delta}{C}\delta. \tag{79}$$

With $C > 0$, the steady-state yaw rate is negative. If the vessel is unstable ($C < 0$), it turns in the opposite direction than expected. This turning rate equation can also be used to estimate turning radius R:

$$R = \frac{U}{r} = \frac{UC}{-Y_v N_\delta \delta}. \tag{80}$$

The radius goes up directly with C, indicating that too stable a ship has poor turning performance. We see also that increasing the rudder area increases N_δ, decreasing R as desired. Increasing the deflection δ to reduce R works only to the point of stalling.

4.5 Summary of the Linear Maneuvering Model

We conclude our discussion of the yaw/sway model by noting that

1. The linearized sway/yaw dynamics of a surface vessel are strongly coupled, and they are independent of the longitudinal dynamics.

2. The design parameter C should be slightly greater than zero for easy turning, and " hands-off" stability. The case $C < 0$ should only be considered under active feedback control.

3. The analysis is valid only up to small angles of attack and turning rates. Very tight maneuvering requires the nonlinear inertial components and hydrodynamic terms. Among other effects, the nonlinear equations couple surge to the other motions, and the actual vessel loses forward speed during maneuvering.

4.6 Stability in the Vertical Plane

Stability in the horizontal plane changes very little as a function of speed, because drag and lift effects generally scale with U^2. This fact is *not* true in the vertical plane, for which the dimensional weight/buoyancy forces and moments are invariant with speed. For example, consider the case of heave and pitch, with $x_G = 0$ and no actuation:

$$m\left(\frac{\partial w}{\partial t} - Uq\right) = Z_{\dot{w}}\dot{w} + Z_w w + Z_{\dot{q}}\dot{q} + Z_q q + (B - W) \quad (81)$$

$$I_{yy}\frac{dq}{dt} = M_{\dot{w}}\dot{w} + M_w w + M_{\dot{q}}\dot{q} + M_q q - Bl_b \sin\theta. \quad (82)$$

The last term in each equation is a hydrostatic effect induced by opposing net buoyancy B and weight W. l_b denotes the vertical separation of the center of gravity and the center of buoyancy, creating the so-called righting moment which nearly all underwater vehicles possess. Because buoyancy effects do not change with speed, the dynamic properties and hence stability of the vehicle may change with speed.

5 SIMILITUDE

5.1 Use of Nondimensional Groups

For a consistent description of physical processes, we require that all terms in an equation must have the same units. On the basis of physical laws, some quantities are dependent on other, independent quantities. We form nondimensional groups out of the dimensional ones in this section, and apply the technique to maneuvering.
The Buckingham π-theorem provides a basis for all nondimensionalization. Let a quantity Q_n be given as a function of a set of $n - 1$ other quantities:

$$Q_n = f_Q(Q_1, Q_2, \cdots, Q_{n-1}). \tag{83}$$

There are n variables here, but suppose also that there are only k independent ones; k is equivalent to the number of physical unit types encountered. The theorem asserts that there are $n - k$ dimensionless groups π_i that can be formed, and the functional equivalence is reduced to

$$\pi_{n-k} = f_\pi(\pi_1, \pi_2, \cdots, \pi_{n-k-1}). \tag{84}$$

Example. Suppose we have a block of mass m resting on a frictionless horizontal surface. At time zero, a steady force of magnitude F is applied. We want to know $X(T)$, the distance that the block has moved as of time T. The dimensional function is $X(T) = f_Q(m, F, T)$, so $n = 4$. The (MKS) units are

$$\begin{aligned} [X(\cdot)] &= m \\ [m] &= kg \\ [F] &= kgm/s^2 \\ [T] &= s, \end{aligned}$$

and therefore $k = 3$. There is just one nondimensional group in this relationship; π_1 assumes only a constant (but unknown) value. Simple term-cancellation gives $\pi_1 = X(T)m/FT^2$, not far at all from the known result that $X(T) = FT^2/2m$!

Example. Consider the flow rate Q of water from an open bucket of height h, through a drain nozzle of diameter d. We have

$$Q = f_Q(h, d, \rho, \mu, g),$$

where the water density is ρ, and its absolute viscosity μ; g is the acceleration due to gravity. No other parameters affect the flow rate. We have $n = 6$, and the (MKS) units of these quantities are:

$$\begin{aligned} [Q] &= m^3/s \\ [h] &= m \\ [d] &= m \\ [\rho] &= kg/m^3 \\ [\mu] &= kg/ms \\ [g] &= m/s^2 \end{aligned}$$

There are only three units that appear: [length, time, mass], and thus $k = 3$. Hence, only three non-dimensional groups exist, and one is a unique function of the other two. To

arrive at a set of valid groups, we must create three nondimensional quantities, making sure that each of the original (dimensional) quantities is represented. Intuition and additional manipulations come in handy, as we now show.
Three plausible first groups are: $\pi_1 = \rho Q/d\mu$, $\pi_2 = d\rho\sqrt{gh}/\mu$, and $\pi_3 = h/d$. Note that all six quantities appear at least once. Since h and d have the same units, they could easily change places in the first two groups. However, π_1 is recognized as a Reynolds number pertaining to the orifice flow. π_2 is more awkward, but products and fractions of groups are themselves valid groups, and we may construct $\pi_4 = \pi_1/\pi_2 = Q/d^2\sqrt{gh}$ to nondimensionalize Q with a pressure velocity, and then $\pi_5 = \pi_1/\pi_4 = \rho d\sqrt{gh}/\mu$ to establish an orifice Reynolds number independent of Q. We finally have the useful result

$$\pi_4 = f_\pi(\pi_5, \pi_2) \longrightarrow$$
$$\frac{Q}{d^2\sqrt{gh}} = f_\pi\left(\frac{\rho d\sqrt{gh}}{\mu}, \frac{h}{d}\right).$$

The uncertainty about where to use h and d, and the questionable importance of h/d as a group are remnants of the theorem. Intuition is that h/d is immaterial, and the other two terms have a nice physical meaning, e.g., π_5 is a Reynolds number.
The power of the π-theorem is primarily in reducing the number of parameters which must be considered independently to characterize a process. In the flow example, the theorem reduced the number of independent parameters from five to two, with no constraints about the actual physics taking place.

5.2 Common Groups in Marine Engineering

One frequently encounters the following groups in fluid mechanics and marine engineering:

1. Froude number:

$$Fr = \frac{U}{\sqrt{gL}}, \tag{85}$$

   where U is the speed of the vessel, g is the acceleration due to gravity, and L is the waterline length of the vessel. The Froude number appears in problems involving pressure boundary conditions, such as in waves on the ocean surface. Roughly speaking, it relates the vessel speed U to water wave speeds of wavelength L; the phase speed of a surface wave is $V = \sqrt{\lambda g/2\pi}$, where λ is the wavelength.

2. Cavitation number:

$$\delta = \frac{P_\infty - P_v}{\frac{1}{2}\rho U^2}, \tag{86}$$

   where P_∞ represents the ambient total pressure, P_v the vapor pressure of the fluid, and U the propeller inlet velocity. A low cavitation number means that the Bernoulli pressure loss across the lifting surface will cause the fluid to vaporize, causing bubbles, degradation of performance, and possible deterioration of the material.

3. Reynolds number:

$$Re = \frac{Ul}{\mu/\rho}, \tag{87}$$

where U is velocity, μ is absolute viscosity, and ρ is density. Since Re appears in many applications, l represents one of many length scales. Reynolds number is a ratio of fluid inertial pressures to viscous pressures: When Re is high, viscous effects are negligible. Re can be used to characterize pipe flow, bluff body wakes, and flow across a plate, among others.

4. Weber number:

$$W = \frac{\rho U^2 l}{\sigma}, \tag{88}$$

where σ is the surface tension of a fluid. Given that $[\sigma] = N/m$ (MKS), ρU^2 normalizes pressure, and l normalizes length. The Weber number indicates the importance of surface tension.

To appreciate the origins of these terms from a fluid particle's point of view, consider a box having side lengths $[dx, dy, dz]$. Various forces on the box scale as

$$\begin{aligned}
\text{(inertia)} \quad F_i &= \rho\frac{\partial v}{\partial t}dxdydz + \rho v\frac{\partial v}{\partial x}dxdydz \simeq \rho U^2 l^2 \\
\text{(gravity)} \quad F_g &= \rho g dxdydz \simeq \rho g l^3 \\
\text{(pressure)} \quad F_p &= P dxdy \simeq P l^2 \\
\text{(shear)} \quad F_s &= \mu\frac{\partial v}{\partial z}dxdy \simeq \mu U l \\
\text{(surface tension)} \quad F_\sigma &= \sigma dx \simeq \sigma l.
\end{aligned}$$

Thus the groups listed above can be written as

$$\begin{aligned}
Fr = \frac{F_i}{F_g} &\simeq \frac{U^2}{gl} \\
\delta = \frac{F_p}{F_i} &\simeq \frac{P}{\rho U^2} \\
Re = \frac{F_i}{F_s} &\simeq \frac{\rho U l}{\mu} \\
W = \frac{F_i}{F_\sigma} &\simeq \frac{\rho U^2 l}{\sigma}
\end{aligned}$$

When testing models, it is imperative to maintain as many of the nondimensional groups as possible of the full-scale system. This holds for the geometry of the body and the kinematics

of the flow, the surface roughness, and all of the relevant groups governing fluid dynamics. Consider the example of nozzle flow from a bucket. Suppose that we conduct a model test in which Re is abnormally large, i.e., the viscous effects are negligible. Under inviscid conditions, the flow rate is $Q = \pi d^2\sqrt{2gh}/4$. This rate cannot be achieved for lower-Re conditions because of fluid drag in the orifice, however.
In a vessel, we write the functional relationship for drag as a starting point:

$$\begin{aligned} C_r = \frac{D}{\frac{1}{2}\rho A U^2} &= f_Q(\rho, \mu, g, \sigma, U, l) \\ &= f_\pi(Re, Fr, W). \end{aligned}$$

First, since l is large, W is very large, and hence surface tension plays no role. Next, we look at $Re = Ul/\nu$ and $Fr = U/\sqrt{gl}$, both of which are important for surface vessels. Suppose that $l_{ship} = \lambda l_{model}$, so that usually $\lambda >> 1$; additionally, we set $g_{model} = g_{ship}$, i.e., the model and the true vessel operate in the same gravity field.
Froude number similitude requires $U_{model} = U_{ship}/\sqrt{\lambda}$. Then Reynolds number scaling implies directly $\nu_{model} = \nu_{ship}/\lambda^{3/2}$. Unfortunately, few fluids with this property are workable in a large testing tank. As a result, accurate scaling of Re for large vessels to model scale is quite difficult.
For surface vessels, and submarines near the surface, it is a routine procedure to employ turbulence stimulators to achieve flow that would normally occur with ship-scale Re. Above a critical value $Re \simeq 500,000$, C_f is not sensitive to Re. With this achieved, one then tries to match Fr closely.

5.3 Similitude in Maneuvering

The linear equations of motion for the horizontal yaw/sway problem are:

$$\begin{aligned} (m - Y_{\dot v})\dot v - Y_v v + (mU - Y_r)r + (mx_G - Y_{\dot r})\dot r &= Y \\ (I_{zz} - N_{\dot r})\dot r + (mx_G U - N_r)r + (mx_G - N_{\dot v})\dot v - N_v v &= N. \end{aligned}$$

These equations can be nondimensionalized in a standard way, by using the quantities $[U, L, \rho]$: these three values provide the necessary units of length, time, and mass, and furthermore are readily accessible to the user. First, we create nondimensional states, denoted with a prime symbol:

$$\begin{aligned} \dot v' &= \frac{L}{U^2}\dot v \\ v' &= \frac{1}{U}v \\ \dot r' &= \frac{L^2}{U^2}\dot r \end{aligned} \tag{89}$$

$$r' = \frac{L}{U} r.$$

We follow a similar procedure for the constant terms as follows, including a factor of 1/2 with ρ, for consistency with our previous expressions:

$$\begin{aligned}
m' &= \frac{m}{\frac{1}{2}\rho L^3} \\
I'_{zz} &= \frac{I_{zz}}{\frac{1}{2}\rho L^5} \\
x'_G &= \frac{x_G}{L} \\
U' &= \frac{U}{U} = 1 \\
Y'_{\dot{v}} &= \frac{Y_{\dot{v}}}{\frac{1}{2}\rho L^3} \\
Y'_v &= \frac{Y_v}{\frac{1}{2}\rho U L^2} \\
Y'_{\dot{r}} &= \frac{Y_{\dot{r}}}{\frac{1}{2}\rho L^4} \\
Y'_r &= \frac{Y_r}{\frac{1}{2}\rho U L^3} \\
Y' &= \frac{Y}{\frac{1}{2}\rho U^2 L^2} \\
N'_{\dot{v}} &= \frac{N_{\dot{v}}}{\frac{1}{2}\rho L^4} \\
N'_v &= \frac{N_v}{\frac{1}{2}\rho U L^3} \\
N'_{\dot{r}} &= \frac{N_{\dot{r}}}{\frac{1}{2}\rho L^5} \\
N'_r &= \frac{N_r}{\frac{1}{2}\rho U L^4} \\
N' &= \frac{N}{\frac{1}{2}\rho U^2 L^3}.
\end{aligned} \tag{90}$$

Note that every force has been normalized with $\frac{1}{2}\rho U^2 L^2$, and every moment with $\frac{1}{2}\rho U^2 L^3$; time has been also nondimensionalized with L/U. Thus we arrive at a completely equivalent set of nondimensional system equations,

$$\begin{aligned}
(m' - Y'_{\dot{v}})\dot{v}' - Y'_v v' + (m'U' - Y'_r)r' + (m'x'_G - Y'_{\dot{r}})\dot{r}' &= Y' \\
(I'_{zz} - N'_{\dot{r}})\dot{r}' + (m'x'_G U' - N'_r)r' + (m'x'_G - N'_{\dot{v}})\dot{v}' - N'_v v' &= N'.
\end{aligned} \tag{91}$$

Since fluid forces and moments generally scale with U^2, the nondimensionalized description holds for a range of velocities.

5.4 Roll Equation Similitude

Certain nondimensional coefficients may arise which depend explicitly on U, and therefore require special attention. Let us carry out a similar normalization of the simplified roll equation

$$(I_{xx} - K_{\dot{p}})\dot{p} - K_p p - K_\psi \psi = K. \tag{92}$$

For a surface vessel, the roll moment K_ψ is based on metacentric stability, and has the form $K_\psi = -\rho g \nabla (GM)$, where ∇ is the displaced fluid volume of the vessel, and GM is the metacentric height. The nondimensional terms are

$$\begin{aligned} I'_{xx} &= \frac{I_{xx}}{\frac{1}{2}\rho L^5} \\ K'_{\dot{p}} &= \frac{K_{\dot{p}}}{\frac{1}{2}\rho L^5} \\ K'_p &= \frac{K_p}{\frac{1}{2}\rho U L^4} \\ K'_\psi &= \frac{K_\psi}{\frac{1}{2}\rho U^2 L^3} \\ \dot{p}' &= \frac{L^2}{U^2}\dot{p} \\ p' &= \frac{L}{U}p, \end{aligned} \tag{93}$$

leading to the equivalent system

$$(I'_{xx} - K'_{\dot{p}})\dot{p}' - K'_p p' - \left(\frac{2gL}{U^2}\right)\nabla'(GM')\psi = K'. \tag{94}$$

Note that the roll angle ϕ was not nondimensionalized. The Froude number has a very strong influence on roll stability, since it appears explicitly in the nondimensional righting moment term, and also has a strong influence on K'_p.
In the case of a submarine, the righting moment has the form $K_\psi = -Bh$, where B is the buoyant force, and h is the righting arm. The nondimensional coefficient becomes

$$K'_\psi = -\frac{Bh}{\frac{1}{2}\rho U^2 L^3}.$$

K'_ψ again depends strongly on U, since B and h are fixed; this K'_ψ needs to be maintained in model tests.

6 CAPTIVE MEASUREMENTS

Before making the decision to measure hydrodynamic derivatives, a preliminary search of the literature may turn up useful estimates. For example, test results for many hull-forms have already been published, and the basic lifting surface models are not difficult. The available computational approaches should be considered as well; these are very good for predicting added mass in particular. Finally, modern sensors and computer control systems make possible the estimation of certain coefficients based on open-water tests of a model or a full-scale design.
In model tests, the Froude number $Fr = \frac{U}{\sqrt{gL}}$, which scales the influence of surface waves, must be maintained between model and full-scale surface vessels. Reynolds number $Re = \frac{UL}{\nu}$, which scales the effect of viscosity, need not be matched as long as the scale model attains turbulent flow (supercritical Re). One can use turbulence stimulators near the bow if necessary. Since the control surface(s) and propeller(s) affect the coefficients, they should both be implemented in model testing.

6.1 Towtank

In a towtank, tow the vehicle at different angles of attack, measuring sway force and yaw moment. The slope of the curve at zero angle determines Y_v and N_v respectively; higher-order terms can be generated if the points deviate from a straight line. Rudder derivatives can be computed also by towing with various rudder angles.

6.2 Rotating Arm Device

On a rotating arm device, the vessel is fixed on an arm of length R, rotating at constant rate r: the vessel forward speed is $U = rR$. The idea is to measure the crossbody force and yaw moment as a function of r, giving the coefficients Y_r and N_r. Note that the lateral force also contains the component $(m - Y_{\dot{v}})r^2R$. The coefficients Y_v and N_v can also be obtained by running with a fixed angle of attack. Finally, the measurement is made over one rotation only, so that the vessel does not re-enter its own wake.

6.3 Planar-Motion Mechanism

With a planar motion mechanism, the vessel is towed at constant forward speed U, but is held by two posts, one forward and one aft, which can each impose independent sway motions, therefore producing variable yaw. The model moves in pure sway if $y_a(t) = y_b(t)$, in a pure yaw motion about the mid-length point if $y_a(t) = -y_b(t)$, or in a combination sway and yaw motion. The connection points are a distance l forward and aft from the vessel origin.
Usually a sinusoidal motion is imposed:

$$\begin{aligned} y_a(t) &= a\cos\omega t \\ y_b(t) &= b\cos(\omega t + \psi), \end{aligned} \tag{95}$$

and the transverse forces on the posts are measured and approximated as

$$\begin{aligned} Y_a(t) &= F_a \cos(\omega t + \theta_a) \\ Y_b(t) &= F_b \cos(\omega t + \theta_b). \end{aligned} \tag{96}$$

If linearity holds, then

$$\begin{aligned} (m - Y_{\dot{v}})\dot{v} - Y_v v + (mU - Y_r)r + (mx_G - Y_{\dot{r}})\dot{r} &= Y_a + Y_b \\ (I_{zz} - N_{\dot{r}})\dot{r} + (mx_G U - N_r)r + (mx_G - N_{\dot{v}})\dot{v} - N_v v &= (Y_b - Y_a)l. \end{aligned} \tag{97}$$

We have $v = (\dot{y}_a + \dot{y}_a)/2$ and $r = (\dot{y}_b - \dot{y}_a)/2l$. When $a = b$, these become

$$\begin{aligned} v &= -\frac{a\omega}{2}\left(\sin\omega t(1 + \cos\psi) + \cos\omega t \sin\psi\right) \\ \dot{v} &= -\frac{a\omega^2}{2}\left(\cos\omega t(1 + \cos\psi) - sin\omega t \sin\psi\right) \\ r &= -\frac{a\omega}{2l}\left(\sin\omega t(\cos\psi - 1) + \cos\omega t \sin\psi\right) \\ \dot{r} &= -\frac{a\omega^2}{2l}\left(\cos\omega t(\cos\psi - 1) - \sin\omega t \sin\psi\right). \end{aligned} \tag{98}$$

Equating the sine terms and then the cosine terms, we obtain four independent equations:

$$\begin{aligned} (m - Y_{\dot{v}})\left(-\frac{a\omega^2}{2}\right)(1 + \cos\psi) - \\ Y_v\left(-\frac{a\omega}{2}\right)\sin\psi + \\ (mU - Y_r)\left(-\frac{a\omega}{2l}\right)\sin\psi + \\ (mx_G - Y_{\dot{r}})\left(-\frac{a\omega^2}{2l}\right)(\cos\psi - 1) &= F_a\cos\theta_a + F_b\cos\theta_b \end{aligned} \tag{99}$$

$$\begin{aligned} (m - Y_{\dot{v}})\left(-\frac{a\omega^2}{2}\right)(-\sin\psi) - \\ Y_v\left(-\frac{a\omega}{2}\right)(1 + \cos\psi) + \\ (mU - Y_r)\left(-\frac{a\omega}{2l}\right)(\cos\psi - 1) + \\ (mx_G - Y_{\dot{r}})\left(-\frac{a\omega^2}{2l}\right)(-\sin\psi) &= -F_a\sin\theta_a - F_b\sin\theta_b \end{aligned} \tag{100}$$

$$(I_{zz} - N_{\dot{r}})\left(-\frac{a\omega^2}{2l}\right)(\cos\psi - 1) + \tag{101}$$

$$
\begin{aligned}
(mx_G U - N_r)\left(-\frac{a\omega}{2l}\right)\sin\psi + \\
(mx_G - N_{\dot{v}})\left(-\frac{a\omega^2}{2}\right)(1+\cos\psi) - \\
N_v\left(-\frac{a\omega}{2}\right)\sin\psi &= l(F_b\cos\theta_b - F_a\cos\theta_a) \\
(I_{zz} - N_{\dot{r}})\left(-\frac{a\omega^2}{2l}\right)(-\sin\psi) + \\
(mx_G U - N_r)\left(-\frac{a\omega}{2l}\right)(\cos\psi - 1) + \\
(mx_G - N_{\dot{v}})\left(-\frac{a\omega^2}{2}\right)(-\sin\psi) - \\
N_v\left(-\frac{a\omega}{2}\right)(1+\cos\psi) &= l(-F_b\sin\theta_b + F_a\sin\theta_a)
\end{aligned}
\tag{102}
$$

In this set of four equations, we know from the imposed motion the values $[U, \psi, a, \omega]$. From the experiment, we obtain $[F_a, F_b, \theta_a, \theta_b]$, and from the rigid-body model we have $[m, I_{zz}, x_G]$. It turns out that the two cases of $\psi = 0$ (pure sway motion) and $\psi = 180^o$ (pure yaw motion) yield a total of eight independent equations, exactly what is required to find the eight coefficients $[Y_{\dot{v}}, Y_v, Y_{\dot{r}}, Y_r, N_{\dot{v}}, N_v, N_{\dot{r}}, N_r]$. Remarkably, we can write the eight solutions directly: For $\psi = 0$,

$$
\begin{aligned}
(m - Y_{\dot{v}})\left(-\frac{a\omega^2}{2}\right)(2) &= F_a\cos\theta_a + F_b\cos\theta_b \\
-Y_v\left(-\frac{a\omega}{2}\right)(2) &= -F_a\sin\theta_a - F_b\sin\theta_b \\
(mx_G - N_{\dot{v}})\left(-\frac{a\omega^2}{2}\right)(2) &= l(F_b\cos\theta_b - F_a\cos\theta_a) \\
-N_v\left(-\frac{a\omega}{2}\right)(2) &= l(-F_b\sin\theta_b + F_a\sin\theta_a),
\end{aligned}
\tag{103}
$$

to be solved respectively for $[Y_{\dot{v}}, Y_v, N_{\dot{v}}, N_v]$. For $\psi = 180^o$, we have

$$
\begin{aligned}
(mx_G - Y_{\dot{r}})\left(-\frac{a\omega^2}{2l}\right)(-2) &= F_a\cos\theta_a + F_b\cos\theta_b \\
(mU - Y_r)\left(-\frac{a\omega}{2l}\right)(-2) &= -F_a\sin\theta_a - F_b\sin\theta_b \\
(I_{zz} - N_{\dot{r}})\left(-\frac{a\omega^2}{2l}\right)(-2) &= l(F_b\cos\theta_b - F_a\cos\theta_a) \\
(mx_G U - N_r)\left(-\frac{a\omega}{2l}\right)(-2) &= l(-F_b\sin\theta_b + F_a\sin\theta_a),
\end{aligned}
\tag{104}
$$

to be solved for $[Y_{\dot{r}}, Y_r, N_{\dot{r}}, N_r]$. Thus, the eight linear coefficients for a surface vessel maneuvering, for a given speed, can be deduced from two tests with a planar motion mechanism. We note that the nonlinear terms will play a significant role if the motions are too large, and that some curve fitting will be needed in any event. The PMM can be driven with more complex trajectories which will target specific nonlinear terms.

7 STANDARD MANEUVERING TESTS

This section describes some of the typical maneuvering tests which are performed on full-scale vessels, to assess stability and performance.

7.1 Dieudonné Spiral

1. Achieve steady speed and direction for one minute. No changes in speed setting are made after this point.
2. Turn rudder quickly by 15°, and keep it there until steady yaw rate is maintained for one minute.
3. Reduce rudder angle by 5°, and keep it there until steady yaw rate is maintained for one minute.
4. Repeat in decrements of -5°, to -15°.
5. Proceed back up to 15°.

The Dieudonné maneuver has the vessel path following a growing spiral, and then a contracting spiral in the opposite direction. The test reveals if the vessel has a memory effect, manifested as a hysteresis in yaw rate r. For example, suppose that the first 15° rudder deflection causes the vessel to turn right, but that the yaw rate at zero rudder, on the way negative, is still to the right. The vessel has gotten "stuck" here, and will require a negative rudder action to pull out of the turn. But if the corrective action causes the vessel to turn left at all, the same memory effect may occur. It is easy to see that the rudder in this case has to be used excessively driving the vessel back and forth. We say that the vessel is unstable, and clearly a poor design.

7.2 Zig-Zag Maneuver

1. With zero rudder, achieve steady speed for one minute.
2. Deflect the rudder to 20°, and hold until the vessel turns 20°.
3. Deflect the rudder to -20°, and hold until the vessel turns to -20∘ with respect to the starting heading.
4. Repeat.

This maneuver establishes several important characteristics of the yaw response. These are: the response time (time to reach a given heading), the yaw overshoot (amount the vessel exceeds $\pm 20^\circ$ when the rudder has turned the other way), and the total period for the 20° oscillations. Of course, similar tests can be made with different rudder angles and different threshold vessel headings.

7.3 Circle Maneuver

From a steady speed, zero yaw rate condition, the rudder is moved to a new setting. The vessel responds by turning in a circle. After steady state is reached again, parameters of interest are the turning diameter, the drift angle β, the speed loss, and the angle of heel ψ.

7.3.1 Drift Angle

The drift angle is the equivalent to angle of attack for lifting surfaces, and describes how the vessel "skids" during a turn. If the turning circle has radius R (measured from the vessel origin), then the speed *tangential* to the circle is $U = rR$. The vessel-reference velocity components are thus $u = U\cos\beta$ and $v = -U\sin\beta$. A line along the vessel centerline reaches closest to the true center of the turning circle at a point termed the *turning center*. At this location, which may or may not exist on the physical vessel, there is no apparent lateral velocity, and it appears to an observer there that the vessel turns about this point.

7.3.2 Speed Loss

Speed loss occurs primarily because of drag induced by the drift angle. A vessel which drifts very little may have very little speed loss.

7.3.3 Heel Angle

Heel during turning occurs as a result of the intrinsic coupling of sway, yaw, and roll caused by the center of gravity. In a surface vessel, the fluid forces act below the waterline, but the center of gravity is near the waterline or above. When the rudder is first deflected, inertial terms dominate (Phase 1) and the sway equation is

$$(m - Y_{\dot{v}})\dot{v} - (Y_{\dot{r}} - mx_G)\dot{r} = Y_\delta \delta. \tag{105}$$

The coefficients for $\dot{r}$ are quite small, and thus the vessel first rolls to starboard (positive) for a positive rudder action.
When steady turning conditions are reached (Phase 3), hydrodynamic forces equalize the centrifugal force mUr and the rudder force $Y_\delta\delta$. The sway equation is

$$-Y_v v + (mU - Y_r)r = Y_\delta \delta, \tag{106}$$

with Y_r small, $v < 0$ when $r > 0$ for most vessels, and $|Y_v v| > |Y_\delta \delta|$. Because the centrifugal force acts above the waterline, the vessel ultimately rolls to port (negative) for a positive rudder action.

The transition between the inertially-dominated and steady-turning regimes includes an overshoot: in the above formulas, the vessel overshoots the final port roll angle as the vessel slows. From the sway equation, we see that if the rudder is straightened out at this point, the roll will momentarily be even worse!
In summary, the vessel rolls into the turn initially, but then out of the turn in the steady state.

7.3.4 Heeling in Submarines with Sails

Submarines typically roll into a turn during all phases. Unlike surface vessels, which have the rigid mass center above the center of fluid forcing, submarines have the mass center below the rudder action point, and additionally feel the effects of a large sail above both. The inertial equation

$$(m - Y_{\dot{v}})\dot{v} - (Y_{\dot{r}} - mx_G)\dot{r} = Y_\delta \delta \tag{107}$$

is dominated by $m\dot{v}$ (acting low), $-Y_{\dot{v}}\dot{v}$ (acting high), and $Y_\delta\delta$ (intermediate). Because $|Y_{\dot{v}}| > m$, the vessel rolls under the sail, the keel out of the turn. In the steady state,

$$-Y_v v + (mU - Y_r)r = Y_\delta \delta. \tag{108}$$

The drift angle β keeps the Y_v-force, acting high, toward the center of the turn, and again centrifugal force mUr causes the bottom of the submarine to move out of the turn. Hence, the roll angle of a submarine with a sail is always into the turn, both initially and in the steady state. The heel angle declines as the speed of the submarine drops.

8 STREAMLINED BODIES

8.1 Nominal Drag Force

A symmetric streamlined body at zero angle of attack experiences only a drag force, which has the form

$$F_A = -\frac{1}{2}\rho C_A A_o U^2. \tag{109}$$

The drag coefficient C_A has both pressure and skin friction components, and hence area A_o is usually that of the wetted surface. Note that the A-subscript will be used to denote zero angle of attack conditions; also, the sign of F_A is negative, because it opposes the vehicle's x-axis.

8.2 Munk Moment

Any shape other than a sphere generates a moment when inclined in an inviscid flow. d'Alembert's paradox predicts zero net *force*, but not necessarily a zero moment. This *Munk moment* arises for a simple reason, the asymmetric location of the stagnation points,

where pressure is highest on the front of the body (decelerating flow) and lowest on the back (accelerating flow). The Munk moment is always destabilizing, in the sense that it acts to turn the vehicle perpendicular to the flow.
Consider a symmetric body with added mass components A_{xx} along the vehicle (slender) x-axis (forward), and A_{zz} along the vehicle's z-axis z (up). We will limit the present discussion to the vertical plane, but similar arguments can be used to describe the horizontal plane. Let α represent the angle of attack, taken to be positive with the nose up – this equates to a negative pitch angle ϕ in vehicle coordinates, if it is moving horizontally. The Munk moment is:

$$\begin{aligned} M_m &= -\frac{1}{2}(A_{zz} - A_{xx})U^2 \sin 2\alpha \\ &\simeq -(A_{zz} - A_{xx})U^2\alpha. \end{aligned} \tag{110}$$

$A_{zz} > A_{xx}$ for a slender body, and the negative sign indicates a negative pitch with respect to the vehicle's pitch axis. The added mass terms A_{zz} and A_{xx} can be estimated from analytical expressions (available only for regular shapes such as ellipsoids), from numerical calculation, or from slender body approximation (to follow).

8.3 Separation Moment

In a viscous fluid, flow over a streamlined body is similar to that of potential flow, with the exceptions of the boundary layer, and a small region near the trailing end. In this latter area, a helical vortex may form and convect downstream. Since vortices correlate with low pressure, the effect of such a vortex is stabilizing, but it also induces drag. The formation of the vortex depends on the angle of attack, and it may cover a larger area (increasing the stabilizing moment and drag) for a larger angle of attack. For a small angle of attack, the transverse force F_n can be written in the same form as for control surfaces:

$$\begin{aligned} F_n &= \frac{1}{2}\rho C_n A_o U^2 \\ &\simeq \frac{1}{2}\rho \frac{\partial C_n}{\partial \alpha}\alpha A_o U^2. \end{aligned} \tag{111}$$

With a positive angle of attack, this force is in the positive z-direction. The zero-α drag force F_A is modified by the vortex shedding:

$$\begin{aligned} F_a &= -\frac{1}{2}\rho C_a A_o U^2 \text{ , where} \\ C_a &= C_A \cos^2\phi. \end{aligned} \tag{112}$$

The last relation is based on writing $C_A(U\cos\phi)^2$ as $(C_A\cos^2\phi)U^2$, i.e., a decomposition using apparent velocity.

8.4 Net Effects: Aerodynamic Center

The Munk moment and the moment induced by separation are competing, and their magnitudes determine the stability of a hullform. First we simplify:

$$\begin{aligned} F_a &= -\gamma_a \\ F_n &= \gamma_n \alpha \\ M_m &= -\gamma_m \alpha. \end{aligned}$$

Each constant γ is taken as positive, and the signs reflect orientation in the vehicle reference frame, with a nose-up angle of attack. The Munk moment is a pure couple which does not depend on a reference point. We pick a temporary origin O for F_n however, and write the total pitch moment about O as:

$$\begin{aligned} M &= M_m + F_n l_n \\ &= (-\gamma_m + \gamma_n l_n)\alpha. \end{aligned} \tag{113}$$

where l_n denotes the (positive) distance between O and the application point of F_n. The net moment about O is zero if we select

$$l_n = \frac{\gamma_m}{\gamma_n}, \tag{114}$$

and the location of O is then called the aerodynamic center or AC.
The point AC has an intuitive explanation: it is the location on the hull where F_n would act to create the total moment. Hence, if the vehicle's origin lies in front of AC, the net moment is stabilizing. If the origin lies behind AC, the moment is destabilizing. For self-propelled vehicles, the mass center must be forward of AC for stability. Similarly, for towed vehicles, the towpoint must be located forward of AC. In many cases with very streamlined bodies, the aerodynamic center is significantly *ahead* of the nose, and in this case, a rigid sting would have to extend at least to AC in order for stable towing. As a final note, since the Munk moment persists even in inviscid flow, AC moves infinitely far forward as viscosity effects diminish.

8.5 Role of Fins in Moving the Aerodynamic Center

Control surfaces or fixed fins are often attached to the stern of a slender vehicle to enhance directional stability. Fixed surfaces induce lift and drag on the body:

$$\begin{aligned} L = \frac{1}{2}\rho A_f U^2 C_l(\alpha) &\simeq \gamma_L \alpha \\ D = -\frac{1}{2}\rho A_f U^2 C_d &\simeq -\gamma_D \text{ (constant)} \end{aligned} \tag{115}$$

These forces act somewhere on the fin, and are signed again to match the vehicle frame, with $\gamma > 0$ and $\alpha > 0$. The summed forces on the body are thus:

$$\begin{aligned} X &= F_a - |D|\cos\alpha + |L|\sin\alpha \\ &\simeq -\gamma_a - \gamma_D + \gamma_L\alpha^2 \\ Z &= F_n + |L|\cos\alpha + |D|\sin\alpha \\ &\simeq \gamma_n\alpha + \gamma_L\alpha + \gamma_D\alpha. \end{aligned} \tag{116}$$

All of the forces are pushing the vehicle up. If we say that the fins act a distance l_f behind the temporary origin O, and that the moment carried by the fins themselves is very small (compared to the moment induced by Ll_f) the total moment is as follows:

$$M = (-\gamma_m + \gamma_n l_n)\alpha + (\gamma_L + \gamma_D)l_f\alpha. \tag{117}$$

The moment about O vanishes if

$$\gamma_m = \gamma_n l_n + l_f(\gamma_L + \gamma_D). \tag{118}$$

The Munk moment γ_m opposes the aggregate effects of vorticity lift γ_n and the fins' lift and drag $\gamma_L + \gamma_D$. With very large fins, this latter term is large, so that l_f might be very small; this is the case of AC moving aft toward the fins. A vehicle with excessively large fins will be difficult to turn and maneuver.

Equation 118 contains two length measurements, referenced to an arbitrary body point O. To solve it explicitly, let l_{fn} denote the (positive) distance that the fins are located behind F_n; this is likely a small number, since both effects usually act near the stern. We solve for l_f:

$$l_f = \frac{\gamma_m + \gamma_n l_{fn}}{\gamma_n + \gamma_L + \gamma_D}. \tag{119}$$

This is the distance that AC is located forward of the fins, and thus AC can be referenced to any other fixed point easily. Without fins, it can be recalled that

$$l_n = \frac{\gamma_m}{\gamma_n}.$$

Hence, the fins act directly in the denominator to shorten l_f. Note that if the fins are located forward of the vortex shedding force F_n, i.e., $l_{fn} < 0$, l_f is reduced, but since AC is referenced to the fins, there is no net gain in stability.

8.6 Aggregate Effects of Body and Fins

Since all of the terms discussed so far have the same dependence on α, it is possible to group them into a condensed form. Setting $\hat{F}_a$ and $\hat{F}_n$ to account for the fuselage and fins, we have

$$\begin{aligned} X &= \hat{F}_a \simeq -\frac{1}{2}\rho \hat{C}_a \hat{A}_o U^2 \\ Z &= \hat{F}_n \simeq \frac{1}{2}\rho \hat{C}'_n \hat{A}_o U^2 \alpha \\ M &= -\hat{F}_n x_{AC} \simeq -\frac{1}{2}\hat{C}'_n \hat{A}_o U^2 x_{AC} \alpha. \end{aligned} \tag{120}$$

8.7 Coefficients Z_w, M_w, Z_q, and M_q for a Slender Body

The angle of attack α is related to the cross-body velocity w as follows:

$$\begin{aligned} \alpha &= -\tan^{-1}\left(\frac{w}{u}\right) \\ &\simeq -\frac{w}{U} \text{ for } U >> w. \end{aligned} \tag{121}$$

We can then write several linear hydrodynamic coefficients easily:

$$\begin{aligned} Z_w &= -\frac{1}{2}\rho \hat{C}'_n \hat{A}_o U \\ M_w &= \frac{1}{2}\rho \hat{C}'_n \hat{A}_o U x_{AC}. \end{aligned} \tag{122}$$

The rotation of the vessel involves complex flow, which depends on both w and q, as well as their derivatives. To start, we consider the contribution of the fins only – slender body theory, introduced shortly, provides good results for the hull. The fin center of pressure is located a distance l_f aft of the body origin, and we assume that the vehicle is moving horizontally, with an instantaneous pitch angle of θ. The angle of attack seen by the fin is a combination of a part due to θ and a part linear with q:

$$\alpha_f \simeq -\theta + \frac{l_f q}{U} \tag{123}$$

and so lateral force and moment derivatives (for the fin alone) emerge as

$$\begin{aligned} Z_q &= -\frac{1}{2}\rho C'_l A_f U l_f \\ M_q &= -\frac{1}{2}\rho C'_l A_f U l_f^2. \end{aligned} \tag{124}$$

9 SLENDER-BODY THEORY

9.1 Introduction

Consider a slender body with $d << L$, that is mostly straight. The body could be asymmetric in cross-section, or even flexible, but we require that the lateral variations are small and

smooth along the length. The idea of the slender-body theory, under these assumptions, is to think of the body as a longitudinal stack of thin sections, each having an easily-computed added mass. The effects are integrated along the length to approximate lift force and moment. Slender-body theory is accurate for small ratios d/L, except near the ends of the body.
As one example, if the diameter of a body of revolution is $d(s)$, then we can compute $\delta m_a(x)$, where the nominal added mass value for a cylinder is

$$\delta m_a = \rho \frac{\pi}{4} d^2 \delta x. \tag{125}$$

The added mass is equal to the mass of the water displaced by the cylinder. The equation above turns out to be a good approximation for a number of two-dimensional shapes, including flat plates and ellipses, if d is taken as the width dimension presented to the flow. Many formulas for added mass of two-dimensional sections, as well as for simple three-dimensional bodies, can be found in the books by Newman and Blevins.

9.2 Kinematics Following the Fluid

The added mass forces and moments derive from accelerations that fluid particles experience when they encounter the body. We use the notion of a fluid derivative for this purpose: the operator d/dt indicates a derivative taken in the frame of the passing particle, not the vehicle. Hence, this usage has an indirect connection with the derivative described in our previous discussion of rigid-body dynamics.
For the purposes of explaining the theory, we will consider the two-dimensional heave/surge problem only. The local geometry is described by the location of the centerline; it has vertical location (in body coordinates) of $z_b(x,t)$, and local angle $\alpha(x,t)$. The time-dependence indicates that the configuration is free to change with time, i.e., the body is flexible. Note that the curvilinear coordinate s is nearly equal to the body-reference (linear) coordinate x. The velocity of a fluid particle *normal* to the body at x is $w_n(t,x)$:

$$w_n = \frac{\partial z_b}{\partial t} \cos\alpha - U \sin\alpha. \tag{126}$$

The first component is the time derivative in the body frame, and the second due to the deflection of the particle by the inclined body. If the body reference frame is rotated to the flow, that is, if $w \neq 0$, then $\partial z_b/\partial t$ will contain w. For small angles, $\sin\alpha \simeq \tan\alpha = \partial z_b/\partial x$, and we can write

$$w_n \simeq \frac{\partial z_b}{\partial t} - U \frac{\partial z_b}{\partial x}.$$

The fluid derivative operator in action is as follows:

$$w_n = \frac{dz_b}{dt} = \left(\frac{\partial}{\partial t} - U \frac{\partial}{\partial x} \right) z_b.$$

9.3 Derivative Following the Fluid

A more formal derivation for the fluid derivative operator is quite simple. Let $\mu(x,t)$ represent some property of a fluid particle.

$$\begin{aligned}\frac{d}{dt}\left[\mu(x,t)\right] &= \lim_{\delta t\to 0}\frac{1}{\delta t}\left[\mu(x+\delta x,t+\delta t)-\mu(x,t)\right]\\ &= \left[\frac{\partial\mu}{\partial t}-U\frac{\partial\mu}{\partial x}\right].\end{aligned}$$

The second equality can be verified using a Taylor series expansion of $\mu(x+\delta x,t+\delta t)$:

$$\mu(x+\delta x,t+\delta t)=\mu(x,t)+\frac{\partial\mu}{\partial t}\delta t+\frac{\partial\mu}{\partial x}\delta x+h.o.t.,$$

and noting that $\delta x=-U\delta t$. The fluid is convected downstream with respect to the body.

9.4 Differential Force on the Body

If the local transverse velocity is $w_n(x,t)$, then the differential inertial force on the body here is the derivative (following the fluid) of the momentum:

$$\delta F=-\frac{d}{dt}\left[m_a(x,t)w_n(x,t)\right]\delta x. \tag{127}$$

Note that we could here let the added mass vary with time also – this is the case of a changing cross-section! The lateral velocity of the point $z_b(x)$ in the body-reference frame is

$$\frac{\partial z_b}{\partial t}=w-xq, \tag{128}$$

such that

$$w_n=w-xq-U\alpha. \tag{129}$$

Taking the derivative, we have

$$\frac{\delta F}{\delta x}=-\left(\frac{\partial}{\partial t}-U\frac{\partial}{\partial x}\right)\left[m_a(x,t)(w(t)-xq(t)-U\alpha(x,t))\right].$$

We now restrict ourselves to a rigid body, so that neither m_a nor α may change with time.

$$\frac{\delta F}{\delta x}=m_a(x)(-\dot{w}+x\dot{q})+U\frac{\partial}{\partial x}\left[m_a(x)(w-xq-U\alpha)\right]. \tag{130}$$

9.5 Total Force on a Vessel

The net lift force on the body, computed with strip theory is

$$Z = \int_{x_T}^{x_N} \delta F dx \tag{131}$$

where x_T represents the coordinate of the tail, and x_N is the coordinate of the nose. Expanding, we have

$$\begin{aligned} Z &= \int_{x_T}^{x_N} m_a(x) \left[-\dot{w} + x\dot{q}\right] dx + U \int_{x_T}^{x_N} \frac{\partial}{\partial x} \left[m_a(x)(w - xq - U\alpha)\right] dx \\ &= -m_{33}\dot{w} - m_{35}\dot{q} + U m_a(x)(w - xq - U\alpha)|_{x=x_T}^{x=x_N}. \end{aligned}$$

We made use here of the added mass definitions

$$\begin{aligned} m_{33} &= \int_{x_T}^{x_N} m_a(x) dx \\ m_{35} &= -\int_{x_T}^{x_N} x m_a(x) dx. \end{aligned}$$

Additionally, for vessels with pointed noses and flat tails, the added mass m_a at the nose is zero, so that a simpler form occurs:

$$Z = -m_{33}\dot{w} - m_{35}\dot{q} - U m_a(x_T)(w - x_T q - U\alpha(x_T)). \tag{132}$$

In terms of the linear hydrodynamic derivatives, the strip theory thus provides

$$\begin{aligned} Z_{\dot{w}} &= -m_{33} \\ Z_{\dot{q}} &= -m_{35} \\ Z_w &= -U m_a(x_T) \\ Z_q &= U x_T m_a(x_T) \\ Z_{\alpha(x_T)} &= U^2 m_a(x_T). \end{aligned}$$

It is interesting to note that both Z_w and $Z_{\alpha(x_T)}$ depend on a nonzero base area. In general, however, potential flow estimates do not create lift (or drag) forces for a smooth body, so this should come as no surprise. The two terms are clearly related, since their difference depends only on how the body coordinate system is oriented to the flow. Another noteworthy fact is that the lift force depends only on α at the tail; α could take any value(s) along the body, with no effect on Z.

9.6 Total Moment on a Vessel

A similar procedure can be applied to the moment predictions from slender body theory (again for small α):

$$\begin{aligned} M &= -\int_{x_T}^{x_N} x\delta F dx \\ &= \int_{x_T}^{x_N} x\left(\frac{\partial}{\partial t} - U\frac{\partial}{\partial x}\right)\left[m_a(x)(w - xq - U\alpha)\right] dx \\ &= \int_{x_T}^{x_N} x m_a(x)(\dot{w} - x\dot{q}) dx - U\int_{x_T}^{x_N} x\frac{\partial}{\partial x}\left[m_a(x)(w - xq + U\alpha)\right] dx. \end{aligned}$$

Then we make the further definition

$$m_{55} = \int_{x_T}^{x_N} x^2 m_a(x) dx,$$

(note that $m_{35} = m_{53}$) and use integration by parts to obtain

$$\begin{aligned} M &= -m_{35}\dot{w} - m_{55}\dot{q} - U x m_a(x)(w - xq - U\alpha)|_{x=x_T}^{x=x_N} + \\ &\quad U\int_{x_T}^{x_N} m_a(x)(w - xq - U\alpha) dx. \end{aligned}$$

The integral above contains the product $m_a(x)\alpha(x)$, which must be calculated if α changes along the length. For simplicity, we now assume that α is in fact constant on the length, leading to

$$\begin{aligned} M &= -m_{35}\dot{w} - m_{55}\dot{q} + U x_T m_a(x_T)(w - x_T q - U\alpha) + \\ &\quad U m_{33} w + U m_{35} q - U^2 m_{33}\alpha. \end{aligned}$$

Finally, the linear hydrodynamic moment derivatives are

$$\begin{aligned} M_{\dot{w}} &= -m_{35} \\ M_{\dot{q}} &= -m_{55} \\ M_w &= U x_T m_a(x_T) + U m_{33} \\ M_q &= -U x_T^2 m_a(x_T) + U m_{35} \\ M_\alpha &= -U^2 x_T m_a(x_T) - U^2 m_{33}. \end{aligned}$$

The derivative M_w is closely-related to the Munk moment, whose linearization would provide $M_w = (m_{33} - m_{11})U$. The Munk moment (an exact result) may therefore be used to make a correction to the second term in the slender-body approximation above of M_w. As with the lift force, M_w and M_α are closely related, depending only on the orientation of the body frame to the flow.

9.7 Relation to Wing Lift

There is an important connection between the slender body theory terms involving added mass at the tail ($m_a(x_T)$), and low aspect-ratio wing theory. The lift force from the latter is of the form $L = -\frac{1}{2}\rho U A C_l' w$, where $A = cs$, the product of chord (long) and span (short). The lift coefficient slope is approximated by (Hoerner)

$$C_l' \approx \frac{1}{2}\pi(AR), \tag{133}$$

where (AR) is the aspect ratio. Inserting this approximation into the lift formula, we obtain

$$L = -\frac{\pi}{4}\rho s^2 U w. \tag{134}$$

Now we look at a slender body approximation of the same force: The added mass at the tail is $m_a(x_T) = \rho s^2 \pi/4$, and using the slender-body estimate for Z_w, we calculate for lift:

$$\begin{aligned} Z &= -m_a(x_T) U w \\ &= -\frac{\pi}{4}\rho s^2 U w. \end{aligned}$$

Slender-body theory is thus able to recover exactly the lift of a low-aspect ratio wing. Where does the slender-body predict the force will act? Recalling that $M_w = U m_{33} + U x_T m_a(x_T)$, and since $m_{33} = 0$ for a front-back symmetric wing, the estimated lift force acts at the trailing edge. This location will tend to stabilize the wing, in the sense that it acts to orient the wing parallel to the incoming flow.

9.8 Convention: Hydrodynamic Mass Matrix A

Hydrodynamic derivatives that depend on accelerations are often written as components of a mass matrix A. By listing the body-referenced velocities in the order $\vec{s} = [u, v, w, p, q, r]$, we write $(M + A)\dot{\vec{s}} = \vec{F}$, where M is the mass matrix of the *material* vessel and F is a generalized force. Therefore $A_{33} = -Z_{\dot{w}}$, $A_{5,3} = -M_{\dot{w}}$, and so on.

10 PRACTICAL LIFT CALCULATIONS

10.1 Characteristics of Lift-Producing Mechanisms

At a small angle of attack, a slender body experiences transverse force due to: helical body vortices, the blunt trailing end, and fins. The helical body vortices are stable and symmetric in this condition, and are convected continuously into the wake. The low pressure associated with the vortices provides the suction force, usually toward the stern of the vehicle. The blunt trailing end induces lift as a product of added mass effects, and can be accurately modeled with slender body theory. A blunt trailing edge also induces some drag, which itself is stabilizing. The fins can often be properly modeled using experimental data parametrized with aspect ratio and several other geometric quantities.

As the angle of attack becomes larger, the approximations in the fin and slender-body analysis will break down. The helical vortices can become bigger while remaining stable, but eventually will split randomly. Some of it convects downstream, and the rest peels away from the body; this shedding is nonsymmetric, and greatly increases drag by widening the wake. In the limit of a 90° angle of attack, vorticity sheds as if from a bluff-body, and there is little axial convection.

10.2 Jorgensen's Formulas

There are some heuristic and theoretical formulas for predicting transverse force and moment on a body at various angles of attack, and we now present one of them due to Jorgensen. The formulas provide a good systematic procedure for design, and are best suited to vehicles with a blunt trailing edge. We call the area of the stern the *base area.*
Let the body have length L, and reference area A_r. This area could be the frontal projected area, the planform area, or the wetted area. The body travels at speed U, and angle of attack α. The normal force, axial force, and moment coefficients are defined as follows:

$$\begin{aligned} C_N &= \frac{F_N}{\frac{1}{2}\rho U^2 A_r} \\ C_A &= \frac{F_A}{\frac{1}{2}\rho U^2 A_r} \\ C_M &= \frac{M_{x_m}}{\frac{1}{2}\rho U^2 A_r L}. \end{aligned} \tag{135}$$

The moment M_{x_m} is taken about a point x_m, measured back from the nose; this location is arbitrary, and appears explicitly in the formula for C_m. Jorgensen gives the coefficients as follows:

$$C_N = \frac{A_b}{A_r}\sin 2\alpha \cos\frac{\alpha}{2} + \frac{A_p}{A_r} C_{d_n} \sin^2\alpha \tag{136}$$

$$C_A = C_{A_o}\cos^2\alpha \tag{137}$$

$$C_M = -\frac{\nabla - A_b(L - x_m)}{A_r L}\sin 2\alpha\cos\frac{\alpha}{2} - C_{d_n}\frac{A_p}{A_r}\left(\frac{x_m - x_c}{L}\right)\sin^2\alpha. \tag{138}$$

We have listed only the formulas for the special case of circular cross-section, although the complete equations do account for more complex shapes. Further, we have assumed that $L >> D$, which is also not a constraint in the complete equations. The parameters used here are

- A_b: stern base area. $A_b = 0$ for a body that tapers to a point at the stern.
- C_{d_n}: crossflow drag coefficient; equivalent to that of an infinite circular cylinder. If "normal" Reynolds number $Re_n = U\sin\alpha D/\nu$,

  - $C_d \approx 1.2, Re_n < 3 \times 10^5$
  - $C_d \approx 0.3, 3 \times 10^5 < Re_n < 7 \times 10^5$
  - $C_d \approx 0.6, 7 \times 10^5 < Re_n$.

- A_p: planform area.
- C_{A_o}: axial drag at zero angle of attack, both frictional and form. $C_{A_o} \simeq 0.002-0.006$ for slender streamlined bodies, based on wetted surface area. It depends on $Re = UL/\nu$.
- ∇: body volume.
- x_c: distance from the nose backwards to the center of the planform area.

In the formula for normal force, we see that if $A_b = 0$, only drag forces act to create lift, through a $\sin^2\alpha$-dependence. Similarly, the axial force is simply the zero-α result, modified by $\cos^2\alpha$. In both cases, scaling of U^2 into the body principle directions is all that is required. There are several terms that match exactly the slender-body theory approximations for small α. These are the first term in the normal force (C_N), and the entire first term in the moment (C_M), whether or not $A_b = 0$. Finally, we note that the second term in C_M disappears if $x_m = x_c$, i.e., if the moment is referenced to the center of the planform area. The idea here is that the fore and aft components of crossflow drag cancel out.
The aerodynamic center (again referenced toward the stern, from the nose) can be found after the coefficients are computed:

$$x_{AC} = x_m + \frac{C_M}{C_N} L. \tag{139}$$

As written, the moment coefficient is negative if the moment destabilizes the body, while C_N is always positive. Thus, the moment seeks to move the AC forward on the body, but the effect is moderated by the lift force.

10.3 Hoerner's Data: Notation

An excellent reference for experimental data is the two-volume set by S. Hoerner. It contains a large amount of aerodynamic data from many different types of vehicles, wings, and other common engineering shapes. A few notations are used throughout the books, and are described here.
First, dynamic pressure is given as $q = \frac{1}{2}\rho U^2$, such that two typical body lift coefficients are:

$$C_Y = \frac{Y}{DLq}$$
$$C_{Y_d} = \frac{Y}{D^2 q}.$$

The first version uses the *rectangular planform area* as a reference, while the second uses the *square frontal area*. Hence, $C_Y = C_{Y_d} D/L$. Two moment coefficients are:

$$C_N = \frac{N}{LD^2q}$$
$$C_{N_1} = \frac{N_1}{LD^2q},$$

where N is the moment taken about the body mid-point, and N_1 is taken about the nose. Note that the reference area for moment is the *square* frontal projection, and the reference length is body length L. The following relation holds for these definitions

$$C_{N_1} = C_N + \frac{1}{2}C_{Y_d}.$$

The lift and moment coefficients are strongly dependent on angle of attack; Hoerner uses the notation

$$\begin{aligned} C_{nb} &= \frac{\partial C_N}{\partial b} \\ C_{n\cdot b} &= \frac{\partial C_{N_1}}{\partial b} \\ C_{yb} &= \frac{\partial C_y}{\partial b}, \end{aligned}$$

and so on, where b is the angle of attack, usually in degrees. It follows from above that $C_{n\cdot b} = C_{nb} + C_{ydb}/2$.

10.4 Slender-Body Theory vs. Experiment

In an experiment, the net moment is measured, comprising both the destabilizing part due to the potential flow, and the stabilizing part due to vortex shedding or a blunt tail. Comparison of the measurements and the theory allows us to place the action point of the suction force. This section gives the formula for this location in Hoerner's notation, and gives two further examples of how well the slender-body theory matches experiments.

For $L/D > 6$, the slender-body (pure added mass) estimates give $\tilde{C}_{nb} \simeq -0.015/deg$, acting to destabilize the vehicle. The value compares well with $-0.027/deg$ for a long cylinder and $-0.018/deg$ for a long ellipsoid; it also reduces to -0.009 for $L/D = 4$. Note that the negative sign here is consistent with Hoerner's convention that destabilizing moments have negative sign.

The experimental lift force is typically given by $C_{Yb} \simeq 0.003/deg$; this acts at a point on the latter half of the vehicle, stabilizing the angle. Because this coefficient scales roughly with wetted area, proportional to LD, it changes little with L/D. It can be compared with a low-aspect ratio wing, which achieves an equivalent lift of $\pi(AR)/2 = 0.0027$ for $(AR) = 10 \simeq D/L$.

The point at which the viscous forces act can then be estimated as the following distance aft of the nose:

$$\frac{x}{L} = \frac{C_{n\cdot b} - \tilde{C}_{nb}}{C_{ydb}} \tag{140}$$

The calculation uses experimental values of C_{ydb} and $C_{n\cdot b}$, the moment slope referenced to the nose. In the table following are values from Hoerner (p. 13-2, Figure 2) for a symmetric and a blunt-ended body.

	symmetric	blunt
L/D	6.7	6.7
$\tilde{C}_{nb}$	-0.012	-0.012
C_{yb}	0.0031	0.0037
C_{ydb}	0.021	0.025
$C_{n\cdot b}$	0.0012 (stable)	0.0031 (stable)
C_{nb}	-0.0093 (unstable)	-0.0094 (unstable)
x/L	0.63	0.60

In comparing the two body shapes, we see that the moment at the nose is much more stable (positive) for the body with a blunt trailing edge. At the body midpoint, however, both vehicles are equally unstable. The blunt-tailed geometry has a much larger lift force, but it acts too close to the midpoint to add any stability there.
The lift force dependence on the blunt tail is not difficult to see, using slender-body theory. Consider a body, with trailing edge radius r. The slender-body lift force associated with this end is simply the product of speed U and local added mass $m_a(x_T)$ (in our previous notation). It comes out to be

$$Z = \frac{1}{2}\rho U^2(\pi r^2)(2\alpha), \tag{141}$$

such that the first term in parentheses is an effective area, and the second is a lift coefficient. With respect to the area πr^2, the lift curve slope is therefore $2/rad$. Expressed in terms of the Hoerner reference area D^2, the equivalent lift coefficient is $C_{ydb} = 0.0044/deg$, where we made the assumption here that $2r/D = 0.4$ for the data in the table. This lift difference, due solely to the blunt end condition, is consistent with measurements.

10.5 Slender-Body Approximation for Fin Lift

Let us now consider two fins of span s each, acting at the tail end of the vehicle. This is the case if the vehicle body tapers to a point where the fins have their trailing edge. The slender-body approximation of lift as a result of blunt-end conditions is

$$Z = \pi s^2 \rho U^2 \alpha. \tag{142}$$

Letting the aspect ratio be $(AR) = (2s)^2/A_f$, where A_f is the total area of the fin pair, substitution will give a lift curve slope of

$$C_l' = \frac{\pi}{2}(AR).$$

This is known as Jones' formula, and is quite accurate for $(AR) \simeq 1$. It is inadequate for higher-aspect ratio wings however, overestimating the lift by about 30% when $(AR) = 2$, and worsening further as (AR) grows.

11 FINS AND LIFTING SURFACES

Vessels traveling at significant speed typically use rudders, elevators, and other streamlined control surfaces to maneuver. Their utility arises mainly from the high lift forces they can develop, with little drag penalty. Lift is always defined to act in a direction perpendicular to the flow, and drag in the same direction as the flow.

11.1 Origin of Lift

A lifting surface is nominally an extrusion of a streamlined cross-section: the cross-section has a rounded leading edge, sharp trailing edge, and a smooth surface. The theory of lifting surfaces centers on the Kutta condition, which requires that fluid particle streamlines do not wrap around the trailing edge of the surface, but instead rejoin with streamlines from the other side of the wing at the trailing edge. This fact is true for a non-stalled surface at any angle of attack.
Since the separation point on the front of the section rotates with the angle of attack, it is clear that the fluid must travel faster over one side of the surface than the other. The reduced Bernoulli pressure this induces can then be thought of as the lift-producing mechanism. More formally, lift arises from circulation Γ:

$$\Gamma = \oint \vec{V} \cdot d\vec{s}. \tag{143}$$

and then $L = -\rho U \Gamma$. Circulation is the integral of velocity around the cross-section, and a lifting surface requires circulation in order to meet the Kutta condition.

11.2 Three-Dimensional Effects: Finite Length

Since all practical lifting surfaces have finite length, the flow near the ends may be three-dimensional. Prandtl's inviscid theory provides some insight. Since bound circulation cannot end abruptly at the wing end, it continues on in the fluid, leading to so-called wing-tip vortices. This continuation causes induced velocities at the tips, and some induced drag. Another description for the wing-tip vortices is that the pressure difference across the surface simply causes flow around the end.
A critical parameter which governs the extent of three-dimensional effects is the aspect ratio:

$$AR = \frac{span}{chord} = \frac{span^2}{area}. \tag{144}$$

The second representation is useful for non-rectangular control surfaces. The effective span is taken to be the length between the free ends of a symmetric wing. If the wing is attached

to a wall, the effective span is twice the physical value, by reflection, and in this case the effective aspect ratio is therefore twice the physical value.
The aspect ratio is a strong determinant of wing performance: for a given angle of attack, a larger aspect ratio achieves a higher lift value, but also stalls earlier.
Lift is written as

$$L = \frac{1}{2}\rho U^2 A C_l, \tag{145}$$

where A is the single-side area of the surface. For angles of attack α below stall, the lift coefficient C_l is nearly linear with α: $C_l = C_l'\alpha$, where C_l' is called the lift coefficient slope, and has one empirical description

$$C_l' = \frac{1}{\frac{1}{2\pi\bar{\alpha}} + \frac{1}{\pi(AR)} + \frac{1}{2\pi(AR)^2}}, \tag{146}$$

where α is in radians, $\bar{\alpha} \simeq 0.90$, and AR is the effective aspect ratio. When $AR \to \infty$, the theoretical and maximum value for C_l' is 2π.
The lift generated on a surface is the result of a distributed pressure field; this fact creates both a net force and a net moment. A single equivalent force acts at a so-called center of action x_A, which depends mainly on the aspect ratio. For high AR, $x_A \simeq c/4$, measured back from the front of the wing. For low AR, $x_A \simeq c/2$.

11.3 Ring Fins

Ring fins are useful when space allows, since they are omnidirectional, and structurally more robust than cantilevered plane surfaces. The effective aspect ratio for a ring of diameter d is given as

$$AR = \frac{4d}{\pi c}. \tag{147}$$

The effective area of the ring is taken as

$$A_e = \frac{\pi}{2} dc, \tag{148}$$

and we thus have $L = \frac{1}{2}\rho U^2 A_e C_l'\alpha$, where one formula for C_l' is

$$C_l' = \frac{1}{0.63 + \frac{1}{\pi(AR)}}. \tag{149}$$

12 PROPELLERS AND PROPULSION

12.1 Introduction

We discuss in this section the nature of steady and unsteady propulsion. In many marine vessels and vehicles, an engine (diesel or gas turbine, say) or an electric motor drives the

propeller through a linkage of shafts, reducers, and bearings, and the effects of each part are important in the response of the net system. Large, commercial surface vessels spend the vast majority of their time operating in open-water and at constant speed. In this case, steady propulsion conditions are generally optimized for fuel efficiency. An approximation of the transient behavior of a system can be made using the quasi-static assumption. In the second section, we list several low-order models of thrusters, which have recently been used to model and simulate truly unsteady conditions.

12.2 Steady Propulsion of Vessels

The notation we will use is as given in Table 1, and there are two different flow conditions to consider. *Self-propelled* conditions refer to the propeller being installed and its propelling the vessel; there are no additional forces or moments on the vessel, such as would be caused by a towing bar or hawser. Furthermore, the flow around the hull interacts with the flow through the propeller. We use an *sp* subscript to indicate specifically self-propulsion conditions. Conversely, when the propeller is run in open water, i.e., not behind a hull, we use an o subscript; when the hull is towed with no propeller we use a t subscript. When subscripts are not used, generalization to either condition is implied. Finally, because of similitude (using diameter D in place of L when the propeller is involved), we do not distinguish between the magnitude of forces in model and full-scale vessels.

R_{sp}	N	hull resistance under self-propulsion
R_t	N	towed hull resistance (no propeller attached)
T	N	thrust of the propeller
n_e	Hz	rotational speed of the engine
n_m	Hz	maximum value of n_e
n_p	Hz	rotational speed of the propeller
λ		gear ratio
Q_e	Nm	engine torque
Q_p	Nm	propeller torque
η_g		gearbox efficiency
P_e	W	engine power
P_p	W	propeller shaft power
D	m	propeller diameter
U	m/s	vessel speed
U_p	m/s	water speed seen at the propeller
Q_m	Nm	maximum engine torque
f	kg/s	fuel rate (or energy rate in electric motor)
f_m	kg/s	maximum value of f

Table 1: Nomenclature

12.2.1 Basic Characteristics

In the steady state, force balance in self-propulsion requires that

$$R_{sp} = T_{sp}. \tag{150}$$

The gear ratio λ is usually large, indicating that the propeller turns much more slowly than the driving engine or motor. The following relations define the gearbox:

$$\begin{aligned} n_e &= \lambda n_p \\ Q_p &= \eta_g \lambda Q_e, \end{aligned} \tag{151}$$

and power follows as $P_p = \eta_g P_e$, for any flow condition. We call $J = U_p/n_p D$ the advance ratio of the prop when it is exposed to a water speed U_p; note that in the wake of the vessel, U_p may not be the same as the speed of the vessel U. A propeller operating *in open water* can be characterized by two nondimensional parameters which are both functions of J:

$$K_T = \frac{T_o}{\rho n_p^2 D^4} \text{ (thrust coefficient)} \tag{152}$$

$$K_Q = \frac{Q_{p_o}}{\rho n_p^2 D^5} \text{ (torque coefficient).} \tag{153}$$

The open-water propeller efficiency can be written then as

$$\eta_o = \frac{T_o U}{2\pi n_p Q_{p_o}} = \frac{J(U) K_T}{2\pi K_Q}. \tag{154}$$

This efficiency divides the useful thrust power by the shaft power. Thrust and torque coefficients are typically nearly linear over a range of J, and therefore fit the approximate form:

$$\begin{aligned} K_T(J) &= \beta_1 - \beta_2 J \\ K_Q(J) &= \gamma_1 - \gamma_2 J. \end{aligned} \tag{155}$$

As written, the four coefficients $[\beta_1, \beta_2, \gamma_1, \gamma_2]$ are usually positive, as shown in the figure. We next introduce three factors useful for scaling and parameterizing our mathematical models:

- $U_p = U(1 - w)$; w is referred to as the *wake fraction*. A typical wake fraction of 0.1, for example, indicates that the incoming velocity seen by the propeller is only 90% of the vessel's speed. The propeller is operating in a wake.

  In practical terms, the wake fraction comes about this way: Suppose the open water thrust of a propeller is known at a given U and n_p. Behind a vessel moving at speed U, and with the propeller spinning at the same n_p, the prop creates some extra thrust.

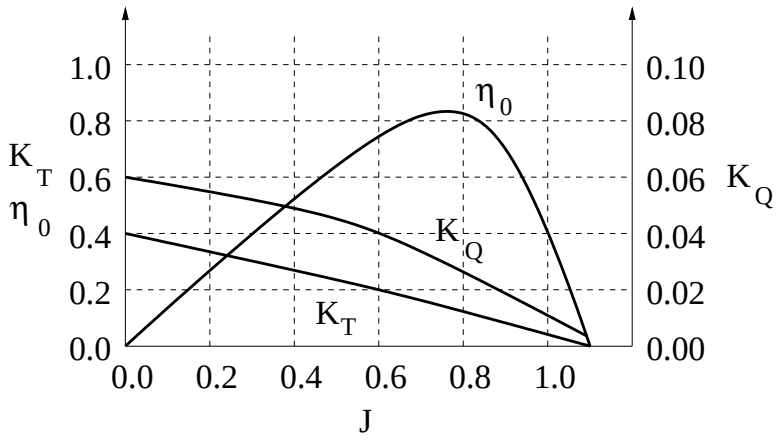


Figure 4: Typical thrust and torque coefficients.

w scales U at the prop and thus J; w is then chosen so that the open water thrust coefficient matches what is observed. The wake fraction can also be estimated by making direct velocity measurements behind the hull, with no propeller.

- $R_t = R_{sp}(1-t)$. Often, a propeller will increase the resistance of the vessel by creating low-pressure on its intake side (near the hull), which makes $R_{sp} > R_t$. In this case, t is a small positive number, with 0.2 as a typical value. t is called the *thrust deduction* even though it is used to model resistance of the hull; it is obviously specific to both the hull and the propeller(s), and how they interact.

  The thrust deduction is particularly useful, and can be estimated from published values, if only the towed resistance of a hull is known.

- $Q_{p_o} = \eta_R Q_{p_{sp}}$. The *rotative efficiency* η_R, which may be greater than one, translates self-propelled torque to open water torque, for the same incident velocity U_p, thrust T, and rotation rate n_p. η_R is meant to account for spatial variations in the wake of the vessel which are not captured by the wake fraction, as well as the turbulence induced by the hull. Note that in comparison with the wake fraction, rotative efficiency equalizes torque instead of thrust.

A common measure of efficiency, the quasi-propulsive efficiency, is based on the towed resistance, and the self-propelled torque.

$$\begin{aligned} \eta_{QP} &= \frac{R_t U}{2\pi n_p Q_{p_{sp}}} \\ &= \frac{T_o(1-t)U_p\eta_R}{2\pi n_p(1-w)Q_{p_o}} \\ &= \eta_o\eta_R\frac{(1-t)}{(1-w)}. \end{aligned} \tag{156}$$

T_o and Q_{p_o} are values for the inflow speed U_p, and thus that η_o is the open-water propeller efficiency at this speed. It follows that $T_o(U_p) = T_{sp}$, which was used to complete the above equation. The quasi-propulsive efficiency can be greater than one, since it relies on the towed resistance and in general $R_t > R_{sp}$. The ratio $(1-t)/(1-w)$ is often called the hull efficiency, and we see that a small thrust deduction t and a large wake fraction w are beneficial effects, but which are in competition. A high rotative efficiency and open water propeller efficiency (at U_p) obviously contribute to an efficient overall system.

12.2.2 Solution for Steady Conditions

The linear form of K_T and K_Q (Equation 156) allows a closed-form solution for the steady-operating conditions. Suppose that the towed resistance is of the form

$$R_t = \frac{1}{2}\rho C_r A_w U^2, \tag{157}$$

where C_r is the resistance coefficient (which will generally depend on Re and Fr), and A_w is the wetted area. Equating the self-propelled thrust and resistance then gives

$$\begin{aligned}
T_{sp} &= R_{sp} \\
T_o &= R_t/(1-t) \\
K_T(J(U_p))\rho n_p^2 D^4(1-t) &= \frac{1}{2}\rho C_r A_w U^2 \\
(\beta_1 - \beta_2 J(U_p))\rho n_p^2 D^4(1-t) &= \frac{1}{2}\rho C_r A_w \frac{U_p^2}{(1-w)^2} \\
\beta_1 - \beta_2 J(U_p) &= \underbrace{\frac{C_r A_w}{2D^2(1-t)(1-w)^2}}_{\delta} J(U_p)^2 \\
J(U_p) &= \frac{-\beta_2 + \sqrt{\beta_2^2 + 4\beta_1\delta}}{2\delta}.
\end{aligned} \tag{158}$$

The last equation predicts the steady-state advance ratio of the vessel, depending only on the propeller open characteristics, and on the hull. The vessel speed can be computed by recalling that $J(U) = U/n_p D$ and $U_p = U(1-w)$, but it is clear that we need now to find n_p. This requires a torque equation, which necessitates a model of the drive engine or motor.

12.2.3 Engine/Motor Models

The torque-speed maps of many engines and motors fit the form

$$Q_e = Q_m F(f/f_m, n_e/n_m), \tag{159}$$

where $F()$ is the characteristic function. For example, gas turbines roughly fit the curves shown in the figure (Rubis). More specifically, if $F()$ has the form

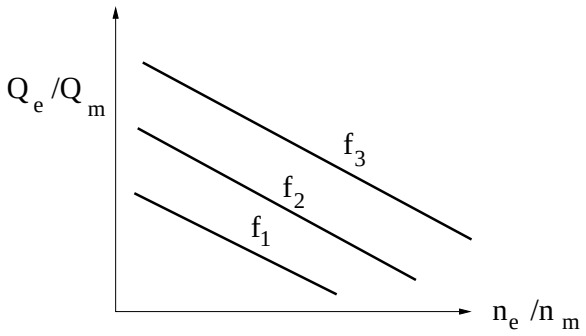


Figure 5: Typical gas turbine engine torque-speed characteristic for increasing fuel rates f_1, f_2, f_3.

$$\begin{aligned} F(f/f_f m, n_e/n_m) &= -\left(a\frac{f}{f_m}+b\right)\frac{n_e}{n_m}+\left(c\frac{f}{f_m}+d\right) \\ &= -\alpha_1\frac{n_p}{n_m/\lambda}+\alpha_2. \end{aligned} \tag{160}$$

then a closed-form solution for n_e (and thus n_p) can be found. The manipulations begin by equating the engine and propeller torque:

$$\begin{aligned} Q_{p_o}(J(U_p)) &= \eta_R Q_{p_s p}(J(U_p)) \\ \rho n_p^2 D^5 K_Q(J(U_p)) &= \eta_R\eta_g\lambda Q_e \\ \rho n_p^2 D^5(\gamma_1-\gamma_2 J(U_p)) &= \eta_R\eta_g\lambda Q_m F(f/f_m, n_e/n_m) \\ \frac{n_p^2}{(n_m/\lambda)^2} &= \underbrace{\frac{\eta_R\eta_g\lambda Q_m}{\rho D^5(\gamma_1-\gamma_2 J(U_p))(n_m/\lambda)^2}}_{\epsilon}\left(-\alpha_1\frac{n_p}{(n_m/\lambda)}+\alpha_2\right) \\ \frac{n_p}{(n_m/\lambda)} &= \frac{-\epsilon\alpha_1+\sqrt{\epsilon^2\alpha_1^2+4\epsilon\alpha_2}}{2}. \end{aligned} \tag{161}$$

Note that the fuel rate enters through both α_1 and α_2.
The dynamic response of the coupled propulsion and ship systems, under the assumption of quasi-static propeller conditions, is given by

$$\begin{aligned} (m+m_a)\dot{u} &= T_{sp}-R_{sp} \\ 2\pi I_p\dot{n}_p &= \eta_g\lambda Q_e - Q_{p_s p}. \end{aligned} \tag{162}$$

Making the necessary substitutions creates a nonlinear model with f as the input; this is left as a problem for the reader.

12.3 Unsteady Propulsion Models

When accurate positioning of the vehicle is critical, the quasi-static assumption used above does not suffice. Instead, the transient behavior of the propulsion system needs to be considered. The problem of unsteady propulsion is still in development, although there have been some very successful models in recent years. It should be pointed out that the models described below all pertain to open-water conditions and electric motors, since the positioning problem has been central to bluff vehicles with multiple electric thrusters.
We use the subscript m to denote a quantity in the motor, and p for the propeller.

12.3.1 One-State Model: Yoerger *et al.*

The torque equation at the propeller and the thrust relation are

$$I_p\dot{\omega}_p = \lambda Q_m - K_\omega \omega_p|\omega_p| \quad (163)$$
$$T = C_t\omega_p|\omega_p|. \quad (164)$$

where I_p is the total (material plus fluid) inertia reflected to the prop;the propeller spins at ω_p radians per second. The differential equation in ω_p pits the torque delivered by the motor against a quadratic-drag type loss which depends on rotation speed. The thrust is then given as a static map directly from the rotation speed.
This model requires the identification of three parameters: I_p, K_ω, and C_t. It is a first-order, nonlinear, low-pass filter from Q_m to T, whose bandwidth depends directly on Q_m.

12.3.2 Two-State Model: Healey *et al.*

The two-state model includes the velocity of a mass of water moving in the vicinity of the blades. It can accommodate a tunnel around the propeller, which is very common in thrusters for positioning. The torque equation, similarly to the above, is referenced to the motor and given as

$$I_m\dot{\omega}_m = -K_\omega\omega_m + K_v V - Q_p/\lambda. \quad (165)$$

Here, K_ω represents losses in the motor due to spinning (friction and resistive), and K_v is the gain on the input voltage (so that the current amplifier is included in K_v). The second dynamic equation is for the fluid velocity at the propeller:

$$\rho AL\gamma\dot{U}_p = -\rho A\Delta\beta(U_p - U)|U_p - U| + T. \quad (166)$$

Here A is the disc area of the tunnel, or the propeller disk diameter if no tunnel exists. L denotes the length of the tunnel, and γ is the effective added mass ratio. Together, $\rho AL\gamma$ is the added mass that is accelerated by the blades; this mass is always nonzero, even if there is no tunnel. The parameter $\Delta\beta$ is called the differential momentum flux coefficient across the propeller; it may be on the order of 0.2 for propellers with tunnels, and up to 2.0 for open propellers.

The thrust and torque of the propeller are approximated using wing theory, which invokes lift and drag coefficients, as well as an effective angle of attack and the propeller pitch. However, these formulae are static maps, and therefore introduce no new dynamics. As with the one-state model of Yoerger *et al.*, this version requires the identification of the various coefficients from experiments. This model has the advantage that it creates a thrust overshoot for a step input, which is in fact observed in experiments.

13 ELECTRIC MOTORS

Modern underwater vehicles and surface vessels are making increased use of electrical actuators, for all range of tasks including weaponry, control surfaces, and main propulsion. This section gives a very brief introduction to the most prevalent electrical actuators: The DC motor, the AC induction motor, and the AC synchronous motor. For the latter two technologies, we consider the case of three-phase power, which is generally preferred over single-phase because of much higher power density; three-phase motors also have simpler starting conditions. AC motors are generally simpler in construction and more robust than DC motors, but at the cost of increased control complexity.

This section provides working parametric models of these three motor types. As with gas turbines and diesel engines, the dynamic response of the actuator is quite fast compared to that of the system being controlled, say a submarine or surface vessel. Thus, we concentrate on portraying the *quasi-static* properties of the actuator – in particular, the torque/speed characteristics as a function of the control settings and electrical power applied.

The discussion below on these various motors is generally invertible (at least for DC and AC synchronous devices) to cover both motors (electrical power in, mechanical power out) and generators (mechanical power in, electrical power out). We will only cover motors in this section, however; a thorough treatment of generators can be found in the references listed. The book by Bradley (19??) has been drawn on heavily in the following.

13.1 Basic Relations

13.1.1 Concepts

First we need the notion of a magnetic flux, analagous to an electrical current, denoted Φ; a common unit is the Weber or Volt-second. The flux density

$$B = \Phi / A \tag{167}$$

is simply the flux per unit area, given in Teslas such that $1T = 1W/m^2$. Correponding to electrical field (Volts/m) is the magnetic field intensity H, in Amperes/meter:

$$H = \frac{B}{\mu_o \mu_r} = \frac{\Phi}{A \mu_o \mu_r}; \tag{168}$$

$\mu_o \approx 4\pi \times 10^{-}7 Henries/meter$ is the permeability of free space, and μ_r is a (dimensionless) relative permeability. The product $\mu_o \mu_r$ represents the real permeability of the material, and is thus the analog of electrical conductivity. A small area A or low relative permeability drives up the field intensity for a given flux Φ.

13.1.2 Faraday's Law

The voltage generated in a conductor experiencing a time rate of change in magnetic flux is given as

$$e = -\frac{d\Phi}{dt} \tag{169}$$

This voltage is commonly called the back-electromotive force or back-e.m.f., since it typically opposes the driving current; it is in fact a limiting factor in DC motors.

13.1.3 Ampere's Law

Current passing through a conductor in a closed loop generates a perpendicular magnetic field intensity given by

$$I = \oint H dl. \tag{170}$$

An important point is that N circular wraps of the same conductor carrying current I induce the field $H = \pi DNI$, where D is the diameter of the circle.

13.1.4 Force

Forces are generated from the orthogonal components of magnetic flux density B and current I:

$$F = I \times B. \tag{171}$$

The units of this force is N/m, and so represents a distributed force on the conductor.

13.2 DC Motors

The DC motor in its simplest form can be described by three relations:

$$\begin{aligned} e_a &= K\Phi\omega \\ V &= e_a + R_a i_a \\ T &= K\Phi i_a, \end{aligned}$$

where

- K is a constant of the motor
- Φ is the airgap magnetic flux per pole (Webers)
- ω is the rotational speed of the motor (rad/s)
- e_a is the back-e.m.f.
- V is the applied voltage
- R_a is the armature resistance on the rotor
- i_a is the current delivered to the armature on the rotor
- T is developed torque

The magnetic field in a typical motor is stationary (on the stator) and is created by permanent magnets or by coils, i.e., Faraday's law. Current is applied to the rotor armature through slip rings, and thus the force on each conductor in the armature is given by $\vec{F} = \vec{i_a} \times \vec{B}$. Back-e.m.f. is created because the conductors in the rotor rotate through the stationary field, causing a relative rate of change of flux. The armature voltage loop contains the back-e.m.f. plus the resistive losses in the windings. As expected, torque scales with the product of magnetic flux and current.
There are three main varieties of DC motors, all of which make use of the relations above. Speed control of the DC motor is primarily through the voltage V, either directly or through pulse-width modulation, but the stator flux could also be controlled in the shunt/independent configurations.

13.2.1 Permanent Field Magnets

Here, the magnetic field is created by permanent magnets arranged in the stator, imposing a steady Φ. The product $K\Phi$ is generally written as k_t, the torque constant of a DC motor, and has units of Nm/A. When SI units are used, k_t also describes back-e.m.f.. The three basic relations are thus rewritten

$$\begin{aligned} e_a &= k_t\omega \\ V &= e_a + R_a i_a \\ T &= k_t i_a, \end{aligned}$$

which leads via substitution to

$$\begin{aligned} \omega &= \frac{1}{k_t}\left[V - \frac{R_a T}{k_t}\right], \text{ or} \\ T &= \frac{k_t}{R_a}\left[V - k_t\omega\right]. \end{aligned}$$

This result indicates that the torque developed scales linearly with the applied voltage, but that it also scales negatively with the motor speed. Hence, at the speed $\omega = V/K_t$, no torque is created. Additionally, the maximum torque is created at zero speed.
Control of these motors is through the voltage V, or, more commonly, directly through current i_a, which gives torque directly.

13.2.2 Shunt or Independent Field Windings

The field created by the stator can be strengthened by replacing the permanent magnets with electromagnets. The field windings are commonly placed in series with the rotor circuit, in parallel with it (shunt connection), or, they may be powered from a completely separate circuit. The latter two cases are effectively equivalent, in the sense that current and hence the field strength can be modulated easily, through a variable resistance in the shunt case. We have

$$\omega = \frac{1}{K\Phi}\left[V - \frac{R_a T}{K\Phi}\right],$$

with the important property that the second term in brackets is small due to the increased field strength, compared with the permanent magnet case above. Thus, the motor speed is effectively independent of torque, which makes these motor types ideal for self-regulation applications. At very high torques and currents, however, the total available flux will be reduced because of field armature reactance; the speed starts to degrade as shown in the figure.

13.2.3 Series Windings

When the field windings are arranged in series with the rotor circuit, the flux is

$$\Phi = K_s i_a,$$

where K_s is a constant of the field winding. This additional connection requires

$$V = e_a + (R_a + R_s) i_a;$$

the field winding brings a new resistance R_s into the voltage loop. It follows through the substitutions that

$$\begin{aligned} T &= K K_s i_a^2 \rightarrow \\ I_a &= \sqrt{\frac{T}{K K_s}} \\ \omega &= \frac{V}{\sqrt{K K_s T}} - \frac{R_a + R_s}{K K_s}. \end{aligned}$$

The effects of resistance are usually quite small, so that the first term dominates, leading to a nonlinear torque/speed characteristic. The starting torque from this kind of motor is exceptionally high, and the series field winding finds wide application in railway locomotives. At the same time, it should be observed that under light loading, the series motor may well self-destruct since there is no intrinsic upper limit to speed!

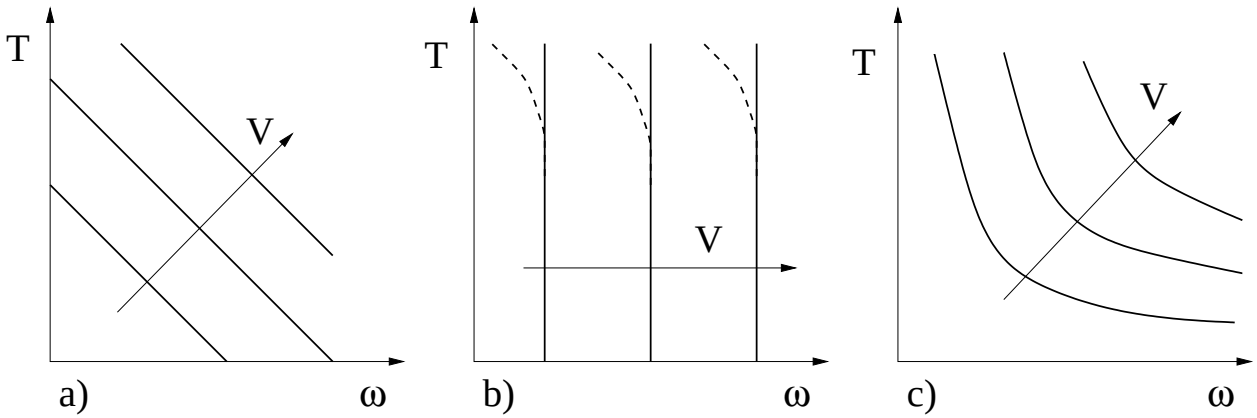


Figure 6: Torque-speed characteristics of three types of DC motors: a) permanent field magnets, b) shunt or independent field winding, c) series field winding.

Some variations on the series and shunt connections are common, and referred to as *compound* DC motors. These achieve other torque/speed curves, including increasing torque with increasing speed, which can offset the speed droop due to field armature reaction effects in the shunt motor.

13.3 Three-Phase Synchronous Motor

The rotor is either fitted with permanent magnets or supplied with DC current to create a static field on the rotor. The stator field windings are driven with three (balanced) phases of an AC supply, such that a moving field is created which rotates around the stator. The torque exerted on the rotor tries to align the two fields, and so the rotor follows the rotating stator field at the same speed. Note that if the rotor speed lags that of the stator field, there is no net torque; hence the name synchronous motor.
A simple model of the synchronous motor is straightforward. As with the DC motors, the voltage loop equation for a single phase on the stator gives

$$V = e_a + j i_a X,$$

where V, e_a, and i_a are now phasors (magnitude of V and e_a measured with respect to ground), $j = \sqrt{-1}$, and X is the reactance (armature and stator leakage) of the machine. Then, let ϕ denote the angle between i_a and V. Equating the electrical (all three phases) and mechanical power gives

$$3Vi_a \cos\phi = T\omega.$$

Next, let δ denote the angle between the phasors e_a and V. It follows from the voltage loop equation that

$$\begin{aligned} i_a \cos\phi X &= e_a \sin\delta \rightarrow \\ T &= \frac{p}{2\omega}\frac{3Ve_a \sin\delta}{X}, \text{ or} \\ T &= \frac{p}{2\omega}\frac{V_{ab}e_{a,ab}\sin\delta}{X} \end{aligned}$$

where

- p is the number of poles on the rotor; two poles means one north pole and one south pole, etc.
- ω is both the rotational speed of the rotor, and the rotational speed of the stator field
- V_{ab} is the line-to-line voltage, equal to $\sqrt{3}V$
- $e_{a,ab}$ is the line-to-line back-e.m.f., equal to $\sqrt{3}e_a$
- *delta* is the angle by which the rotor field lags the stator field (rad)
- X is the synchronous reactance

The torque scales with $\sin\delta$, and thus the rotor lags the stator field when the motor is powering; in a generator, the stator field lags the rotor. If the load torque exceeds the available torque, the synchronous motor can slip one or more poles, causing a large transient disturbance.
Speed control of the three-phase synchronous machine is generally through the frequency of the three-phase power supply, ω, with the assumption that adequate voltage and current are available.

13.4 Three-Phase Induction Motor

Like the synchronous machine, the induction machine has windings on the stator to create a rotating magnetic field at frequency ω. Letting the rotor speed be ω_r, we see immediately that if $\omega \neq \omega_r$, a potential field will be induced on any conductor on the rotor. In the case of a squirrel-cage rotor design, the rotor is made of conductor bars which are shorted out through rings at the ends, and hence the potential field will cause a real current flow. Torque is then generated through the familiar $F = I \times B$ relation. The fact of unequal field and rotor speeds in the induction motor is related to several unique effects, leading to torque-speed characteristic which differs significantly from both the DC the AC synchronous machines.

First we define the slip ratio

$$s = \frac{\omega - \omega_r}{\omega}; \tag{172}$$

a slip ratio of zero means $\omega = \omega_r$ (and hence zero torque because the magnetic field seen by the rotor is constant) while a slip ratio of one implies the rotor is stopped. Most induction motors are designed to operate at a small positive slip ratio, say 0.1-0.2, for reasons described below.
Next, since the magnetic flux lines pass through the rotor, the number of ampere-turns on stator and rotor is equivalent, that is, they form an ideal inductor:

$$N_r I_r = N_s I_s. \tag{173}$$

We consider per-unit quantities from here on, for which we set $N = N_r = N_s$ and hence $I_r = I_s$. If the stator flux at a particular location is $\Phi_s = \Phi_o \sin \omega t$, the associated voltage is $e_s = N d\Phi/dt = N\Phi_s^o \omega \cos \omega t$. On the rotor, the same flux applies, but it rotates more slowly: $\Phi_r = \Phi_o \sin \omega_r t = \Phi_o \sin s\omega t$. Hence the rotor voltage is $e_r = N d\Phi_r/dt = N\Phi_o s\omega \cos s\omega t$. Then the RMS voltage of the stator and rotor sides of the inductive coupling are related by

$$\frac{E_s}{E_r} = \frac{1}{s}. \tag{174}$$

The voltage seen at the stator scales inversely with the slip ratio, for a constant voltage at the rotor. In per-unit terms, the current in the rotor and stator are equivalent, and this then indicates that the rotor impedance, seen from the stator, also scales inversely with the slip ratio:

$$\begin{aligned} Z_{rs} &= \frac{1}{s}\left[R_r + jsX_r\right] \\ &= \frac{1}{s}R_r + jX_r. \end{aligned}$$

The factor of s in the rotor inductance occurs because the field seen by the rotor is actually rotating at $s\omega$.
Next, we construct the (one phase) Thevenin equivalent circuit of the stator: it has a voltage source V_t, and equivalent resistance R and inductance X. This is to be paired with the rotor resistance and inductance, *reflected to the stator*, giving the following current

$$I = \frac{V_t}{\sqrt{(R_r/s + R)^2 + (X_r + X)^2}}.$$

Finally, we need to express the torque/speed characteristic of the machine. The mechanical power is $P_m = (1-s)\omega T$, while the power delivered across the airgap is $P_{gap} = I^2 R_r/s$. The

actual power dissipated in the copper is related to the real rotor resistance: $P_{loss} = I^2 R_r$, and hence the mechanical power is $P_m = 3(P_{gap} - P_{loss}) = 3P_{gap}(1-s)$. It follows that the efficiency of the motor is simply $\eta = 1 - s$. Combining the mechanical power with the torque equation gives

$$\begin{aligned} T &= \frac{P_m}{(1-s)\omega} = \frac{3P_{gap}}{\omega} \\ &= \frac{3V_t^2 R_r}{s\omega\left[(R_r/s + R)^2 + (X_r + X)^2\right]}. \end{aligned}$$

Maximum torque is developed at a slight slippage, with decreased values at lower speeds.

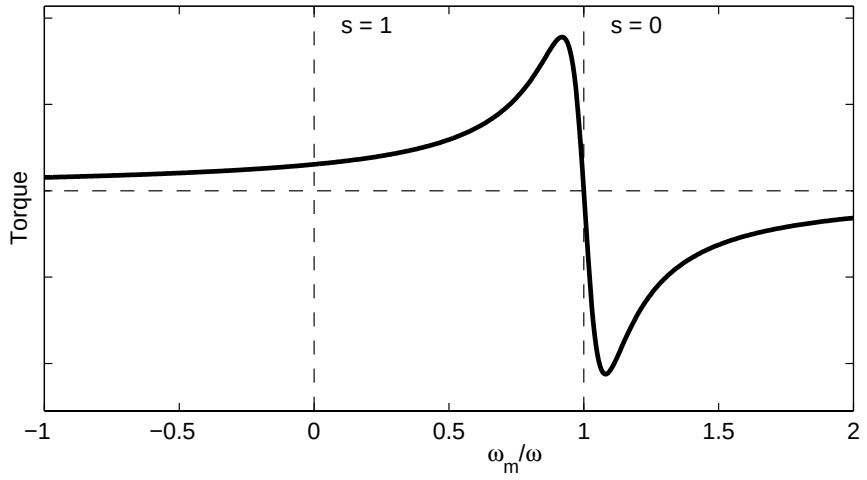


Figure 7: Torque-speed characteristics of a typical three-phase induction motor.

14 TOWING OF VEHICLES

Vehicles which are towed have some similarities to the vehicles that have been discussed so far. For example, towed vehicles are often streamlined, and usually need good directional stability. Some towed vehicles might have active lifting surfaces or thrusters for attitude control. On the other hand, if they are to be supported by a cable, towed vehicles may be quite heavy in water, and do not have to be self-propelled. The cable itself is an important factor in the behavior of the complete towed system, and in this section, we concentrate on cable mechanics more than vehicle characteristics, which can generally be handled with the same tools as other vehicles, i.e., slender-body theory, wing theory, linearization, etc.. Some basic guidelines for vehicle design are given at the end of this section.

Modern cables can easily exceed $5000m$ in length, even a heavy steel cable with $2cm$ diameter. The cables are generally circular in cross section, and may carry power conductors and

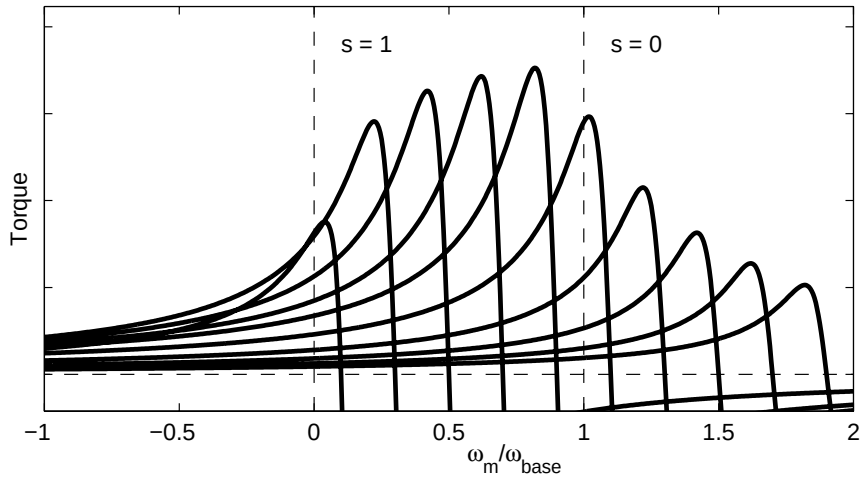


Figure 8: Effects of constant Volts/Hz speed control; voltage is constant above electrical frequency ω_{base}.

multiple communication channels (fiber optic). The extreme L/D ratio for these cables obviates any bending stiffness effects.

Cable systems come in a variety of configurations, and one main division may be made simply of the density of the cable. Light-tether systems are characterized by neutrally-buoyant (or nearly so) cables, with either a minimal vehicle at the end, as in a towed array, or a vehicle capable of maneuvering itself, such as a remotely-operated vehicle. The towed array is a relatively high-velocity system that nominally streams out horizontally behind the vessel. An ROV, on the other hand, operates at low speed, and must have large propulsors to control the tether if there are currents. Heavy systems, in contrast, employ a heavy cable and possibly a heavy weight; the rationale is that gravity will tend to keep the cable vertical and make the deployment robust against currents and towing speed. The heavy systems will generally transmit surface motions and tensions to the towed vehicle much more easily than light-tether systems.

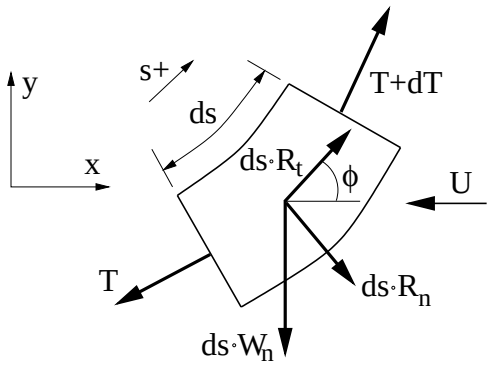
We will not discuss light systems specifically here, but rather look at heavy systems. Most of the analysis can be adapted to either case, however.

14.1 Statics

14.1.1 Force Balance

For the purposes of deriving the static configuration of a cable in a flow, we assume for the moment that that it is inextensible. Tension and hydrostatic pressure will elongate a cable, but the effect is usually a small percentage of the total length.

We employ the curvilinear axial coordinate s, which we take to be zero at the bottom end of the cable; upwards along the cable is the positive direction. The free-body diagram shown



has the following components:

- W_n: net in-water weight of the cable per unit length.
- $R_n(s)$: external normal force, per unit length.
- $R_t(s)$: external tangential force, per unit length.
- $T(s)$: local tension.
- $\phi(s)$: local inclination angle.

Force balance in the tangential and normal coordinates gives two coupled equations for T and ϕ:

$$\frac{dT}{ds} = W_n \sin\phi - R_t \tag{175}$$

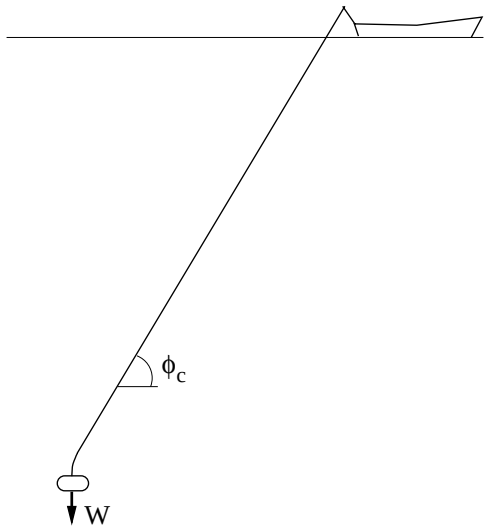
$$T\frac{d\phi}{ds} = W_n \cos\phi + R_n. \tag{176}$$

The external forces are primarily fluid drag; the tangential drag is controlled by a frictional drag coefficient C_t, and the normal drag scales with a crossflow drag coefficient C_n. In both cases, the fluid velocity vector, U horizontal toward the left, is to be projected onto the relevant axes, leading to

$$R_t = -\frac{1}{2}\rho C_t d U^2 \cos^2\phi \tag{177}$$

$$R_n = -\frac{1}{2}\rho C_n d U^2 \sin^2\phi. \tag{178}$$

Note that we simplified the drag laws from a usual form $v|v|$ to v^2, since as drawn, $0 \le \phi \le \pi/2$.



The equations for T and ϕ can be integrated along the cable coordinate s to find the cable's static configuration. Two boundary conditions are needed, and the common case is that a force balance on the vehicle, dominated by drag, weight, and the cable tension, provides both $T(0)$ and $\phi(0)$. For example, a very heavy but low-drag vehicle will impose $\phi(0) \simeq \pi/2$, with $T(0)$ equal to the in-water weight of the vehicle.
With regard to Cartesian coordinates x, y, the cable configuration follows

$$\frac{dx}{ds} = \cos\phi \tag{179}$$

$$\frac{dy}{ds} = \sin\phi. \tag{180}$$

The simultaneous integration of all four equations (T, ϕ, x, y) defines the cable configuration, and current dependency may be included, say U is a function of y.

14.1.2 Critical Angle

For very deep systems, the total weight of cable will generally exceed that the vehicle. This gives rise to a configuration in which the cable is straight for a majority of its length, but turns as necessary at the vehicle end, to meet the bottom boundary condition. In the straight part of the cable, normal weight and drag components are equalized. The uniform angle is called the critical angle ϕ_c, and can be approximated easily. Let the relative importance of weight be given as

$$\delta = \frac{W_n}{\rho C_n d U^2},$$

so that the condition $d\phi/ds = 0$ requires from the force balance

$$\delta \cos\phi_c - \frac{1}{2}\sin^2\phi_c = 0.$$

We are considering the case of $0 < \phi_c < \pi/2$. Substituting $\sin^2\phi_c = 1 - \cos^2\phi_c$, we solve a quadratic equation and keep only the positive solution:

$$\cos\phi_c = \sqrt{\delta^2 + 1} - 1. \tag{181}$$

In the case of a very heavy cable, δ is large, and the linear approximation of the square root $\sqrt{1+\epsilon} \approx 1 + \epsilon/2$ gives

$$\begin{aligned} \cos\phi_c &\simeq \frac{1}{2\delta} \longrightarrow \\ \phi_c &\simeq \frac{\pi}{2} - \frac{1}{2\delta}. \end{aligned} \tag{182}$$

For a very light cable, δ is small; the same approximation gives

$$\cos\phi_c \simeq 1 - \delta \longrightarrow \phi_c \simeq \sqrt{2\delta}.$$

The table below gives some results of the exact solution, and the approximations.

δ	exact	$\delta >> 1$	$\delta << 1$
0.1	0.44	-	0.45
0.2	0.61	-	0.63
0.5	0.91	0.57	1.00
1.0	1.14	1.07	1.41
2.0	1.33	1.32	-
5.0	1.47	1.47	-

14.2 Linearized Dynamics

14.2.1 Derivation

The most direct procedure for deriving useful linear dynamic equations for a planar cable problem is to consider the total tension and angle as made up of static parts summed with dynamic parts:

$$\begin{aligned} T(s,t) &= \bar{T}(s) + \tilde{T}(s,t) \\ \phi(s,t) &= \bar{\phi}(s) + \tilde{\phi}(s,t). \end{aligned}$$

We also write the axial deflection with respect to the static configuration as $p(s,t)$, and the lateral deflection $q(s,t)$. It follows that $\tilde{\phi} = \partial q/\partial s$. Now augment the two static configuration equations with inertial components:

$$\begin{aligned} m\frac{\partial^2 p}{\partial t^2} &= \frac{\partial \bar{T}}{\partial s} + \frac{\partial \tilde{T}}{\partial s} - W_n \sin(\bar{\phi}+\tilde{\phi}) - \frac{1}{2}\rho C_t d\left(U\cos\phi + \frac{\partial p}{\partial t}\right)^2 \\ (m+m_a)\frac{\partial^2 q}{\partial t^2} &= (\bar{T}+\tilde{T})\left(\frac{\partial \bar{\phi}}{\partial s} + \frac{\partial \tilde{\phi}}{\partial s}\right) - W_n\cos(\bar{\phi}+\tilde{\phi}) + \\ &\quad \frac{1}{2}\rho C_n d\left(U\sin\phi - \frac{\partial q}{\partial t}\right)^2. \end{aligned}$$

Here the material mass of the cable per unit length is m, and its transverse added mass is m_a. Note that avoiding the drag law form $v|v|$ again, we have implicitly assumed that $U\cos\phi > |\partial p/\partial t|$ and $U\sin\phi > |\partial q/\partial t|$. If it is not the case, say $U = 0$, then equivalent linearization can be used for the quadratic drag.
Now we perform the trigonometry substitutions in the weight terms, let $\phi \simeq \bar{\phi}$ for the calculation of drag, and substitute the constitutive (Hooke's) law

$$\frac{\partial \tilde{T}}{\partial s} = EA\frac{\partial^2 p}{\partial s^2}.$$

The static solution cancels out of both governing equations, and keeping only linear terms we obtain

$$\begin{aligned} m\frac{\partial^2 p}{\partial t^2} &= EA\frac{\partial^2 p}{\partial s^2} - W_n\cos\bar{\phi}\frac{\partial q}{\partial s} - \rho C_t dU\cos\bar{\phi}\frac{\partial p}{\partial t} \\ (m+m_a)\frac{\partial^2 q}{\partial t^2} &= \bar{T}\frac{\partial^2 q}{\partial s^2} + EA\frac{\partial p}{\partial s}\frac{\partial \bar{\phi}}{\partial s} + W_n\sin\bar{\phi}\frac{\partial q}{\partial s} - \rho C_n dU\sin\bar{\phi}\frac{\partial q}{\partial t}. \end{aligned}$$

The axial dynamics (p) couples with the lateral equation through the weight term $-W_n\cos\bar{\phi}\tilde{\phi}$. The lateral dynamics ($q$) couples with the axial through the term $\tilde{T}\partial\bar{\phi}/\partial s$. An additional weight term $W_n\sin\bar{\phi}\partial q/\partial s$ also appears. The uncoupled dynamics are both in the form of damped wave equations

$$\begin{aligned} m\frac{\partial^2 p}{\partial t^2} + b_t\frac{\partial p}{\partial t} &= EA\frac{\partial^2 p}{\partial s^2} \\ (m+m_a)\frac{\partial^2 q}{\partial t^2} + b_n\frac{\partial q}{\partial t} &= \bar{T}\frac{\partial^2 q}{\partial s^2} + W_n\sin\bar{\phi}\frac{\partial q}{\partial s}, \end{aligned}$$

where we made the substitution $b_t = \rho C_t dU\cos\bar{\phi}$ and $b_n = \rho C_n dU\sin\bar{\phi}$. To a linear approximation, the out-of-plane vibrations of a cable are also governed by the second equation above.

Because of light damping in the tangential direction, heavy cables easily transmit motions and tensions along their length, and can develop longitudinal resonant conditions (next section). In contrast, the lateral cable motions are heavily damped, such that disturbances only travel a few tens or hundreds of meters before they dissipate. The nature of the lateral response, in and out of the towing plane, is a very slow, damped nonlinear filter. High-frequency vessel motions in the horizontal plane are completely missed by the vehicle, while low-frequency motions occur sluggishly, and only after a significant delay time.

	axial	lateral
wave speed	$\sqrt{\frac{EA}{m}}$ FAST	$\sqrt{\frac{\bar{T}(s)}{m+m_a}}$ SLOW
natural frequency	$\frac{n\pi}{L}\sqrt{\frac{EA}{m}}$	$O\left(\frac{n\pi}{L}\sqrt{\frac{\bar{T}(L/2)}{m+m_a}}\right)$
mode shape	sine/cosine	Bessel function
damping	$C_t \simeq O(0.01)$	$C_n \simeq O(1)$
disturbances travel down cable?	YES	NO

14.2.2 Damped Axial Motion

Mode Shape. The axial direction is of particular interest, since it is lightly damped and forced by the heaving of vessels in seas. Consider a long cable governed by the damped wave equation

$$m\ddot{p} + b_t\dot{p} = EAp'' \tag{183}$$

We use over-dots to indicate time derivatives, and primes to indicate spatial derivatives. At the surface, we impose the motion

$$p(L,t) = P\cos\omega t, \tag{184}$$

while the towed vehicle, at the lower end, is an undamped mass responding to the local tension variations:

$$EA\frac{\partial p(0,t)}{\partial s} = M\ddot{p}(0,t). \tag{185}$$

These top and bottom behaviors comprise the boundary conditions for the wave equation. We let $p(s,t) = \tilde{p}(s)\cos\omega t$, so that

$$\tilde{p}'' + \left(\frac{m\omega^2 - i\omega b_t}{EA}\right)\tilde{p} = 0. \tag{186}$$

This admits the solution $\tilde{p}(s) = c_1 \cos ks + c_2 \sin ks$, where

$$k = \sqrt{\frac{m\omega^2 - i\omega b_t}{EA}}. \tag{187}$$

Note that k is complex when $b_t \neq 0$. The top and bottom boundary conditions give, respectively,

$$\begin{aligned} P &= c_1 \cos kL + c_2 \sin kL \\ 0 &= c_1 + \delta c_2, \end{aligned}$$

where $\delta = EAk/\omega^2 M$. These can be combined to give the solution

$$\tilde{p} = P \frac{\delta \cos ks - \sin ks}{\delta \cos kL - \sin kL}. \tag{188}$$

In the case that $M \to 0$, the scalar $\delta \to \infty$, simplifying the result to $\tilde{p} = P \cos ks / \cos kL$.

Dynamic Tension. It is possible to compute the dynamic tension via $\tilde{T} = EA\tilde{p}'$. We obtain

$$\tilde{T} = -EAPk \frac{\delta \sin ks + \cos ks}{\delta \cos kL - \sin kL}. \tag{189}$$

There are two dangerous situations:

- The maximum tension is $\bar{T} + |\tilde{T}|$ and must be less than the working load of the cable. This is normally problematic at the top of the cable, where the static tension is highest.
- If $|\tilde{T}| > \bar{T}$, the cable will unload completely and then reload with extremely high impulsive forces. This is known as snap loading; it occurs primarily at the vehicle, where $\bar{T}$ is low.

Natural Frequency. The natural frequency can be found by letting $b_t = 0$, and investigating the singularity of $\tilde{p}$, for which $\delta \cos kL = \sin kL$. In general, $kL << 1$, but we find that a first-order approximation yields $\omega = \sqrt{EA/LM}$, which is only a correct answer if $M >> mL$, i.e., the system is dominated by the vehicle mass. Some higher order terms need to be kept. We start with better approximations for sin() and cos():

$$\delta \left(1 - \frac{(kL)^2}{2}\right) = kL \left(1 - \frac{(kL)^2}{6}\right).$$

Employing the definition for δ, and recalling that $\omega^2 = k^2 EA/m$, we arrive at

$$\frac{mL}{M} \left(1 - \frac{(kL)^2}{2}\right) = (kL)^2 \left(1 - \frac{(kL)^2}{6}\right).$$

If we match up to second order in kL, then

$$\omega = \sqrt{\frac{EA/L}{M + mL/2}}.$$

This has the familiar form of the square root of a stiffness divided by a mass: the stiffness of the cable is EA/L, and the mass that is oscillating is $M + mL/2$. In very deep water, the effects of $mL/2$ dominate; if $\rho_c = m/A$ is the density of the cable, we have the approximation

$$\omega \simeq \frac{1}{L}\sqrt{\frac{2E}{\rho_c}}.$$

A few examples are given below for a steel cable with $E = 200 \times 10^9 Pa$, and $\rho_c = 7000 kg/m^3$. The natural frequencies near wave excitation at the surface vessel must be taken into account in any design or deployment. Even if a cable can withstand the effects of resonance, it may be undesirable to expose the vehicle to these motions. Some solutions in use today are: stable vessels (e.g., SWATH), heave compensation through an active crane, a clump weight below which a light cable is employed, and an S-shaped length of cable at the bottom formed with flotation balls.

$L = 500m$	$\omega_n = 15.0 rad/s$
$1000m$	$7.6 rad/s$
$2000m$	$3.7 rad/s$
$5000m$	$1.5 rad/s$

14.3 Cable Strumming

Cable strumming causes a host of problems, including obvious fatigue when the amplitudes and frequencies are high. The most noteworthy issue with towing is that the vibrations may cause the normal drag coefficient C_n to increase dramatically – from about 1.2 for a non-oscillating cable to as high as 3.5. This drag penalty decreases the critical angle of towing, so that larger lengths of cable are needed to reach a given depth, and the towed system lags further and further behind the surface vessel. The static tension will increase accordingly as well.

Strumming of cables is caused by the proximity of a preferred vortex formation frequency ω_S to the natural frequency of the structure ω_n. This latter frequency can be obtained as a zero of the lightly-damped Bessel function solution of the lateral dynamics equation above. The preferred frequency of vortex formation is given by the empirical relation $\omega_S = 2\pi SU/d$, where S is the Strouhal number, about 0.16-0.20 for a large range of Re. Strumming of amplitude $d/2$ or greater can occur for $0.6 < \omega_S/\omega_n < 2.0$. The book by Blevins is a good general reference.

14.4 Vehicle Design

The physical layout of a towed vehicle is amenable to the analysis tools of self-propelled vehicles, with the main exceptions that the towpoint presents a large mean force as well as some disturbances, and that the vehicle can be quite heavy in water. Here are basic guidelines to be considered:

1. The towpoint must be located above the vehicle center of in-water weight, for basic roll and pitch stability.

2. The towpoint should be forward of the aerodynamic center, for towing stability reasons.

3. The combined center of mass (material and added mass) should be longitudinally *between* the towpoint and the aerodynamic center, and nearer the towpoint. This will ensure that high-frequency disturbances do not induce excessive pitching.

4. The towpoint should be longitudinally forward of the center of in-air weight, so that the vehicle enters the water fins first, and self-stabilizes with $U > 0$.

5. The center of buoyancy should be behind the in-water center of weight, so that the vehicle pitches downward at small U, and hence the net lift force is downward, away from the surface.

Meeting all of these criteria simultaneously is no small feat, and the performance of the device is very sensitive to small perturbations in the geometry. For this reason, full-scale experiments are commonly used in the design process.

15 TRANSFER FUNCTIONS & STABILITY

The reader is referred to *Laplace Transforms* in the section *MATH FACTS* for preliminary material on the Laplace transform. Partial fractions are presented here, in the context of control systems, as the fundamental link between pole locations and stability.

15.1 Partial Fractions

Solving linear time-invariant systems by the Laplace Transform method will generally create a signal containing the (factored) form

$$Y(s) = \frac{K(s+z_1)(s+z_2)\cdots(s+z_m)}{(s+p_1)(s+p_2)\cdots(s+p_n)}. \tag{190}$$

Although for the moment we are discussing the signal $Y(s)$, later we will see that dynamic systems are described in the same format: in that case we call the impulse response $G(s)$ a transfer function. A system transfer function is identical to its impulse response, since $L(\delta(t)) = 1$.

The constants $-z_i$ are called the zeros of the transfer function or signal, and $-p_i$ are the poles. Viewed in the complex plane, it is clear that the magnitude of $Y(s)$ will go to zero at the zeros, and to infinity at the poles.
Partial fraction expansions alter the form of $Y(s)$ so that the simple transform pairs can be used to find the time-domain output signals. We must have $m < n$; if this is not the case, then we have to divide the numerator by the denominator as necessary to find a simple form.

15.2 Partial Fractions: Unique Poles

Under the condition $m < n$, it is a fact that $Y(s)$ is equivalent to

$$Y(s) = \frac{a_1}{s+p_1} + \frac{a_2}{s+p_2} + \cdots \frac{a_n}{s+p_n}, \tag{191}$$

in the special case that all of the poles are unique and real. The coefficient a_i is termed the *residual* associated with the i'th pole, and once all these are found it is a simple matter to go back to the transform table and look up the time-domain responses.
How to find a_i? A simple rule applies: multiply the right-hand sides of the two equations above by $(s+p_i)$, evaluate them at $s = -p_i$, and solve for a_i, the only one left.

15.3 Example: Partial Fractions with Unique Real Poles

$$G(s) = \frac{s(s+6)}{(s+4)(s-1)} e^{-2s}.$$

Since we have a pure delay and $m = n$, we can initially work with $G(s)/se^{-2s}$. We have

$$\frac{s+6}{(s+4)(s-1)} = \frac{a_1}{s+4} + \frac{a_2}{s-1}, \text{ giving}$$

$$\begin{aligned} a_1 &= \left[\tfrac{(s+6)(s+4)}{(s+4)(s-1)}\right]_{s=-4} = -\frac{2}{5} \\ a_2 &= \left[\tfrac{(s+6)(s-1)}{(s+4)(s-1)}\right]_{s=1} = \frac{7}{5} \end{aligned}$$

Thus

$$\begin{aligned} L^{-1}(G(s)/se^{-2s}) &= -\frac{2}{5}e^{-4t} + \frac{7}{5}e^{t} \longrightarrow \\ g(t) &= \delta(t-2) + \frac{8}{5}e^{-4(t-2)} + \frac{7}{5}e^{t-2}. \end{aligned}$$

The impulse response is needed to account for the step change at $t = 2$. Note that in this example, we were able to apply the derivative operator s *after* expanding the partial

fractions. For cases where a second derivative must be taken, i.e., $m \geq n+1$, special care should be used when accounting for the signal *slope* discontinuity at $t = 0$. The more traditional method, exemplified by Ogata, may prove easier to work through.
The case of repeated real roots may be handled elegantly, but this condition rarely occurs in applications.

15.4 Partial Fractions: Complex-Conjugate Poles

A complex-conjugate pair of poles should be kept together, with the following procedure: employ the form

$$Y(s) = \frac{b_1 s + b_2}{(s+p_1)(s+p_2)} + \frac{a_3}{s+p_3} + \cdots, \tag{192}$$

where $p_1 = p_2^*$ (complex conjugate). As before, multiply through by $(s+p_1)(s+p_2)$, and then evaluate at $s = -p_1$.

15.5 Example: Partial Fractions with Complex Poles

$$G(s) = \frac{s+1}{s(s+j)(s-j)} = \frac{b_1 s + b_2}{(s+j)(s-j)} + \frac{a_3}{s} :$$

$$\begin{aligned}
\left[\frac{s+1}{s}\right]_{s=-j} &= [b_1 s + b_2]_{s=-j} \longrightarrow \\
1 + j &= -b_1 j + b_2 \longrightarrow \\
b_1 &= -1 \\
b_2 &= 1; \text{ also} \\
\left[\frac{s+1}{(s+j)(s-j)}\right]_{s=0} &= a_3 = 1.
\end{aligned}$$

Working out the inverse transforms from the table of pairs, we have simply (noting that $\zeta = 0$)

$$g(t) = -\cos t + \sin t + 1(t).$$

15.6 Stability in Linear Systems

In linear systems, *exponential stability* occurs when all the real exponents of e are strictly negative. The signals decay within an exponential envelope. If one exponent is 0, the response never decays or grows in amplitude; this is called *marginal stability*. If at least one real exponent is positive, then one element of the response grows without bound, and the system is *unstable*.

15.7 Stability $\Longleftrightarrow$ Poles in LHP

In the context of partial fraction expansions, the relationship between stability and pole locations is especially clear. The unit step function $1(t)$ has a pole at zero, the exponential e^{-at} has a pole at $-a$, and so on. All of the other pairs exhibit the same property: *A system is stable if and only if all of the poles occur in the left half of the complex plane.* Marginally stable parts correlate with a zero real part, and unstable parts to a positive real part.

15.8 General Stability

There are two definitions, which apply to systems with input $\vec{u}(t)$ and output $\vec{y}(t)$.

1. **Exponential**. If $\vec{u}(t) = \vec{0}$ and $\vec{y}(0) = \vec{y}_o$, then $|y_i(t)| < \alpha e^{-\gamma t}$, for finite α and $\gamma > 0$. The output asymptotically approaches zero, within a decaying exponential envelope.

2. **Bounded-Input Bounded-Output (BIBO)**. If $\vec{y}(0) = \vec{0}$, and $|f_i(t)| < \gamma, \gamma > 0$ and finite, then $|y_i(t)| < \alpha, \alpha > 0$ and finite.

In linear time-invariant systems, the two definitions are identical. Exponential stability is easy to check for linear systems, but for nonlinear systems, BIBO stability is usually easier to achieve.

16 CONTROL FUNDAMENTALS

16.1 Introduction

16.1.1 Plants, Inputs, and Outputs

Controller design is about creating dynamic systems that behave in useful ways. Many target systems are physical; we employ controllers to steer ships, fly jets, position electric motors and hydraulic actuators, and distill alcohol. Controllers are also applied in macroeconomics and many other important, non-physical systems. It is the fundamental concept of controller design that a set of input variables acts through a given "plant" to create an output. Feedback control then uses sensed plant outputs to apply corrective inputs:

Plant	Inputs	Outputs	Sensors
Jet aircraft	elevator, rudder, etc.	altitude, hdg	altimeter, GPS
Marine vessel	rudder angle	heading	gyrocompass
Hydraulic robot	valve position	tip position	joint angle
U.S. economy	fed interest rate, etc.	prosperity	inflation, M1
Nuclear reactor	cooling, neutron flux	power level	temp., pressure

16.1.2 The Need for Modeling

Effective control system design usually benefits from an accurate model of the plant, although it must be noted that many industrial controllers can be tuned up satisfactorily with no knowledge of the plant. Ziegler and Nichols, for example, developed a general recipe which we detail later. In any event, plant models simply do not match real-world systems exactly; we can only hope to capture the basic components in the form of differential or integro-differential equations.
Beyond prediction of plant behavior based on physics, the process of *system identification* generates a plant model from data. The process is often problematic, however, since the measured response could be corrupted by sensor noise or physical disturbances in the system which cause it to behave in unpredictable ways. At some frequency high enough, most systems exhibit effects that are difficult to model or reproduce, and this is a limit to controller performance.

16.1.3 Nonlinear Control

The bulk of this subject is taught using the tools of linear systems analysis. The main reason for this restriction is that nonlinear systems are difficult to model, difficult to design controllers for, and difficult overall! Within the paradigm of linear systems, there are many sets of powerful tools available. The reader interested in nonlinear control is referred to the book by Slotine and Li (1991).

16.2 Representing Linear Systems

Except for the most heuristic methods of tuning up simple systems, control system design depends on a model of the plant. The transfer function description of linear systems has already been described in the discussion of the Laplace transform. The state-space form is an entirely equivalent *time-domain* representation that makes a clean extension to systems with multiple inputs and multiple outputs, and opens the way to standard tools from linear algebra.

16.2.1 Standard State-Space Form

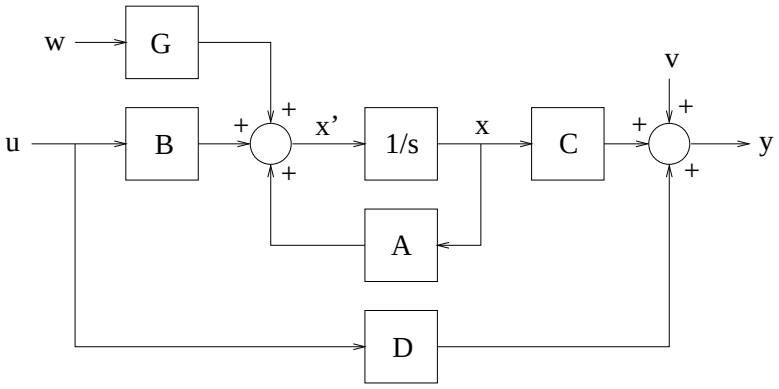
We write a linear system in a state-space form as follows

$$\begin{aligned} \dot{x} &= Ax + Bu + Gw \\ y &= Cx + Du + v \end{aligned} \tag{193}$$

where

- x is a state vector, with as many elements as there are orders in the governing differential equations.
- A is a matrix mapping x to its derivative; A captures the natural dynamics of the system without external inputs.

- B is an input gain matrix for the control input u.
- G is a gain matrix for unknown disturbance w; w drives the state just like the control u.
- y is the observation vector, comprised mainly of a linear combination of states Cx (where C is a matrix).
- Du is a direct map from input to output (usually zero for physical systems).
- v is an unknown sensor noise which corrupts the measurement.



16.2.2 Converting a State-Space Model into a Transfer Function

There are a number of canonical state-space forms available, which can create the same transfer function. In the case of no disturbances or noise, the transfer function (or transfer matrix) can be written as

$$G(s) = \frac{y(s)}{u(s)} = C(sI - A)^{-1}B + D, \tag{194}$$

where I is the identity matrix with the same size as A. A similar equation holds for $y(s)/w(s)$, and clearly $y(s)/v(s) = I$.

16.2.3 Converting a Transfer Function into a State-Space Model

It may be possible to write the corresponding differential equation along one row of the state vector, and then cascade derivatives. For example, consider the following system:

$$\begin{aligned} my''(t) + by'(t) + ky(t) &= u'(t) + u(t) \text{ (mass-spring-dashpot)} \\ G(s) &= \frac{s+1}{ms^2 + bs + k} \end{aligned}$$

Setting $\vec{x} = [y', y]^T$, we obtain the system

$$\begin{aligned} \frac{d\vec{x}}{dt} &= \begin{bmatrix} -b/m & -k/m \\ 1 & 0 \end{bmatrix} \vec{x} + \begin{bmatrix} 1/m \\ 0 \end{bmatrix} u \\ y &= [1 \;\; 1]\, \vec{x} \end{aligned}$$

Note specifically that $dx_2/dt = x_1$, leading to an entry of 1 in the off-diagonal of the second row in A. Entries in the C-matrix are easy to write in this case because of linearity; the system response to u' is the same as the derivative of the system response to u.

16.3 PID Controllers

The most common type of industrial controller is the proportional-integral-derivative (PID) design. If u is the output from the controller, and e is the error signal it receives, this control law has the form

$$\begin{aligned} u(t) &= k_p e(t) + k_i \int_0^t e(\tau) d\tau + k_d e'(t), \\ C(s) = \frac{U(s)}{E(s)} &= k_p + \frac{k_i}{s} + k_d s \\ &= k_p \left[1 + \frac{1}{\tau_i s} + \tau_d s \right], \end{aligned} \tag{195}$$

where the last line is written using the conventions of one overall gain k_p, plus a time characteristic to the integral part (τ_i) and and time characteristic to the derivative part (τ_d). In words, the proportional part of this control law will create a control action that scales linearly with the error – we often think of this as a spring-like action. The integrator is accumulating the error signal over time, and so the control action from this part will continue to grow as long as an error exists. Finally, the derivative action scales with the derivative of the error. The controller will retard motion toward zero error, which helps to reduce overshoot.
The common variations are: P, PD, PI, PID.

16.4 Example: PID Control

Consider the case of a mass (m) sliding on a frictionless table. It has a perfect thruster that generates force $u(t)$, but is also subject to an unknown disturbance $d(t)$. If the linear position of the mass is $y(t)$, and it is perfectly measured, we have the plant

$$my''(t) = u(t) + d(t).$$

Suppose that the desired condition is simply $y(t) = 0$, with initial conditions $y(0) = y_o$ and $y'(0) = 0$.

16.4.1 Proportional Only

A proportional controller alone invokes the control law $u(t) = -k_p y(t)$, so that the closed-loop dynamics follow

$$my''(t) = -k_p y(t) + d(t).$$

In the absence of $d(t)$, we see that $y(t) = y_o \cos\sqrt{\frac{k_p}{m}}t$, a marginally stable response that is undesirable.

16.4.2 Proportional-Derivative Only

Let $u(t) = -k_p y(t) - k_d y'(t)$, and it follows that

$$my''(t) = -k_p y(t) - k_d y'(t) + d(t).$$

The system now resembles a second-order mass-spring-dashpot system where k_p plays the part of the spring, and k_d the part of the dashpot. With an excessively large value for k_d, the system would be overdamped and very slow to respond to any command. In most applications, a small amount of overshoot is employed because the response time is shorter. The k_d value for critical damping in this example is $2\sqrt{mk_p}$, and so the rule is $k_d < 2\sqrt{mk_p}$. The result, easily found using the Laplace transform, is

$$y(t) = y_o e^{\frac{-k_d}{2m}t}\left[\cos\omega_d t + \frac{k_d}{2m\omega_d}\sin\omega_d t\right],$$

where $\omega_d = \sqrt{4mk_p - k_d^2}/2m$. This response is exponentially stable as desired. Note that if the mass had a very large amount of natural damping, a *negative* k_d could be used to cancel some of its effect and speed up the system response.

Now consider what happens if $d(t)$ has a constant bias d_o: it balances exactly the proportional control part, eventually settling out at $y(t = \infty) = d_o/k_p$. To achieve good rejection of d_o with a PD controller, we would need to set k_p very large. However, very large values of k_p will also drive the resonant frequency ω_d up, which is unacceptable.

16.4.3 Proportional-Integral-Derivative

Now let $u(t) = -k_p y(t) - k_i \int_0^t y(\tau)d\tau - k_d y'(t)$: we have

$$my''(t) = -k_p y(t) - k_i \int_0^t y(\tau)d\tau - k_d y'(t) + d(t).$$

The control system has now created a third-order closed-loop response. If $d(t) = d_o$, a time derivative leads to

$$my'''(t) + k_p y'(t) + k_i y(t) + k_d y''(t) = 0,$$

so that $y(t = \infty) = 0$, as desired, provided the roots are stable.

16.5 Heuristic Tuning

For many practical systems, tuning of a PID controller may proceed without any system model. This is especially pertinent for plants which are open-loop stable, and can be safely tested with varying controllers. One useful approach is due to Ziegler and Nichols (e.g., Bélanger,1995), which transforms the basic characteristics of a step response (e.g., the input is $1(t)$) into a reasonable PID design. The idea is to approximate the response curve by a first-order lag (gain k and time constant τ) and a pure delay T:

$$G(s) \simeq \frac{ke^{-Ts}}{\tau s + 1} \tag{196}$$

The following rules apply *only* if the plant contains no dominating, lightly-damped complex poles, and has no poles at the origin:

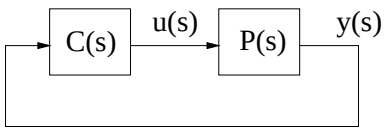
P	$k_p = 1.0\tau/T$		
PI	$k_p = 0.9\tau/T$	$k_i = 0.27\tau/T^2$	
PID	$k_p = 1.2\tau/T$	$k_i = 0.60\tau/T^2$	$k_d = 0.60\tau$

Note that if no pure time delay exists ($T = 0$), this recipe suggests the proportional gain can become arbitrarily high! Any characteristic other than a true first-order lag would therefore be expected to cause a measurable delay.

16.6 Block Diagrams of Systems

16.6.1 Fundamental Feedback Loop

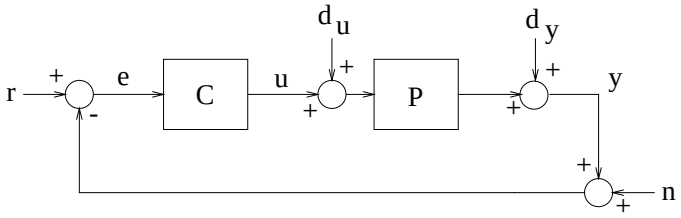
The topology of a feedback system can be represented graphically by considering each dynamical system element to reside within a box, having an input line and an output line. For example, the plant used above (a simple mass) has transfer function $P(s) = 1/ms^2$, which relates the input, force $u(s)$, into the output, position $y(s)$. In turn, the PD-controller has transfer function $C(s) = k_p + k_d s$; its input is the error signal $E(s) = -y(s)$, and its output is force $u(s)$. The feedback loop in block diagram form is shown below.



16.6.2 Block Diagrams: General Case

The simple feedback system above is augmented in practice by three external inputs. The first is a process disturbance, which can be taken to act at the input of the physical plant, or at the output. In the former case, it is additive with the control action, and so has

some physical meaning. In the second case, the disturbance has the same units as the plant output.
Another external input is the *reference command* or *setpoint*, used to create a more general error signal $e(s) = r(s) - y(s)$. Note that the feedback loop, in trying to force $e(s)$ to zero, will necessarily make $y(s)$ approximate $r(s)$.
The final input is sensor noise, which usually corrupts the feedback signal $y(s)$, causing some error in the evaluation of $e(s)$, and so on. Sensors with very poor noise properties can ruin the performance of a control system, no matter how perfectly understood are the other components.



16.6.3 Primary Transfer Functions

Some algebra shows that

$$\begin{aligned}
\frac{e}{r} &= \frac{1}{1+PC} = S \\
\frac{y}{r} &= \frac{PC}{1+PC} = T \\
\frac{u}{r} &= \frac{C}{1+CP} = U.
\end{aligned}$$

$e/r = S$ relates the reference input and noise to the error, and is known as the *sensitivity function*. We would generally like S to be small at low frequencies, so that the tracking error there is small. $y/r = T$ is called the *complementary sensitivity function*. Note that $S + T = 1$, implying that these two functions must always trade off; they cannot both be small or large at the same time. Other systems we encounter again later are the *(forward) loop transfer function* PC, the loop transfer function broken between C and P: CP, and

$$\begin{aligned}
\frac{e}{d_u} &= \frac{-P}{1+PC} \\
\frac{y}{d_u} &= \frac{P}{1+PC} \\
\frac{u}{d_u} &= \frac{-CP}{1+CP} \\
\frac{e}{d_y} &= \frac{-1}{1+PC} = -S
\end{aligned}$$

$$\begin{aligned}
\frac{y}{d_y} &= \frac{1}{1+PC} = S \\
\frac{u}{d_y} &= \frac{-C}{1+CP} = -U \\
\frac{e}{n} &= \frac{-1}{1+PC} = -S \\
\frac{y}{n} &= \frac{-PC}{1+PC} = -T \\
\frac{u}{n} &= \frac{-C}{1+CP} = -U.
\end{aligned}$$

If the disturbance is taken at the plant output, then the three functions S, T, and U (control action) completely describe the system. This will in fact be the procedure when we address loopshaping.

17 MODAL ANALYSIS

17.1 Introduction

The evolution of states in a linear system occurs through independent modes, which can be driven by external inputs, and observed through plant output. This section provides the basis for modal analysis of systems. Throughout, we use the state-space description of a system with $D = 0$:

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

17.2 Matrix Exponential

17.2.1 Definition

In the instance of an unforced response to initial conditions, consider the system

$$\dot{\vec{x}} = A\vec{x}, \qquad \vec{x}(t=0) = \vec{\chi}.$$

In the scalar case, the response is $x(t) = \chi e^{at}$, giving a decaying exponential if $a < 0$. The same notation holds for the case of a vector $\vec{x}$, and matrix A:

$$\begin{aligned}
\vec{x}(t) &= e^{At}\vec{\chi}, \text{ where} \\
e^{At} &= I + At + \frac{(At)^2}{2!} + \cdots
\end{aligned}$$

e^{At} is usually called the matrix exponential.

17.2.2 Modal Canonical Form

Introductory material on the eigenvalue problem and modal decomposition can be found in the *MATH FACTS* section. This modal decomposition of A leads to a very useful state-space representation. Namely, since $A = V\Lambda V^{-1}$, a transformation of state variables can be made, $\vec{x} = V\vec{z}$, leading to

$$\begin{aligned} \dot{\vec{z}} &= \Lambda\vec{z} + V^{-1}B\vec{u} \\ \vec{y} &= CV\vec{z}. \end{aligned} \tag{197}$$

This is called the modal canonical form, since the states are simply the modal amplitudes. These states are uncoupled in Λ, but may be coupled through the input ($V^{-1}B$) and output (CV) mappings. The modal form is numerically robust for computations.

17.2.3 Modal Decomposition of Response

Now we are ready to look at the matrix exponential e^{At} in terms of its constituent modes. Employing the above form for A, we find that

$$\begin{aligned} e^{At} &= I + At + \frac{(At)^2}{2!} + \cdots \\ &= V\left(I + \Lambda t + \frac{(\Lambda t)^2}{2!} + \cdots\right)W^T \\ &= Ve^{\Lambda t}W^T \\ &= \sum_{i=1}^{n} e^{\lambda_i t}\vec{v}_i\vec{w}_i^T. \end{aligned}$$

In terms of the response to an initial condition $\vec{\chi}$, we have

$$\vec{x}(t) = \sum_{i=1}^{n} e^{\lambda_i t}\vec{v}_i(\vec{w}_i^T\vec{\chi}).$$

The product $\vec{w}_i^T\vec{\chi}$ is a scalar, the projection of the initial conditions onto the i'th mode. If $\vec{\chi}$ is perpendicular to $\vec{w}_i$, then the product is zero and the i'th mode does not respond. Otherwise, the i'th mode does participate in the response. The projection of the i'th mode onto the *states* $\vec{x}$ is through the right eigenvector $\vec{v}_i$.
For stability of the system, the eigenvalues of A, that is, λ_i, must have negative real parts; they are in fact the poles of the equivalent transfer function description.

17.3 Forced Response and Controllability

Now consider the system with an external input $\vec{u}$:

$$\dot{\vec{x}} = A\vec{x} + B\vec{u}, \qquad \vec{x}(t=0) = \vec{\chi}.$$

Taking the Laplace transform of the system, taking into account the initial condition for the derivative, we have

$$\begin{aligned} s\vec{x}(s) - \vec{\chi} &= A\vec{x}(s) + B\vec{u}(s) \longrightarrow \\ \vec{x}(s) &= (sI - A)^{-1}\vec{\chi} + (sI - A)^{-1}B\vec{u}(s). \end{aligned}$$

Thus $(sI - A)^{-1}$ can be recognized as the Laplace transform of the matrix exponential e^{At}. In the time domain, the second term then has the form of a convolution of the matrix exponential and the net input $B\vec{u}$:

$$\begin{aligned} \vec{x}(t) &= \int_0^t e^{A(t-\tau)}B\vec{u}(\tau)d\tau \\ &= \sum_{i=1}^{n}\int_0^t e^{\lambda_i(t-\tau)}\vec{v}_i\vec{w}_i^T B\vec{u}(\tau)d\tau. \end{aligned}$$

Suppose now that there are m inputs, such that $B = [\vec{b}_1, \vec{b}_2, \cdots, \vec{b}_m]$. Then some rearrangement will give

$$\vec{x}(t) = \sum_{i=1}^{n}\vec{v}_i\sum_{k=1}^{m}(\vec{w}_i^T\vec{b}_k)\int_0^t e^{\lambda_i(t-\tau)}\vec{u}_k(\tau)d\tau.$$

The product $\vec{w}_i^T\vec{b}_k$, a scalar, represents the projection of the k'th control channel onto the i'th mode. We say that the i'th mode is controllable from the k'th input if the product is nonzero. If a given mode i has $\vec{w}_i^T\vec{b}_k = 0$ for all input channels k, then the mode is uncontrollable.
In normal applications, controllability for the entire system is checked using the following test: Construct the so-called controllability matrix:

$$M_c = [B, AB, \cdots, A^{n-1}B]. \tag{198}$$

This matrix has size $n \times (nm)$, where m is the number of input channels. If M_c has rank n, then the system is controllable, i.e., all modes are controllable.

17.4 Plant Output and Observability

We now turn to a related question: can the complete state vector of the system be observed given only the output measurements $\vec{y}$, and the known control $\vec{u}$? The response due to the external input is easy to compute deterministically, through the convolution integral. Consider the part due to initial conditions $\vec{\chi}$. We found above

$$\vec{x}(t) = \sum_{i=1}^{n} e^{\lambda_i t} \vec{v}_i \vec{w}_i^T \vec{\chi}.$$

The observation is $\vec{y} = C\vec{x}$ (r channels of output), and writing

$$C = \begin{bmatrix} \vec{c}_1^T \\ \cdot \\ \vec{c}_r^T \end{bmatrix}.$$

the k'th channel of the output is

$$y_k(t) = \sum_{i=1}^{n} (\vec{c}_k^T \vec{v}_i) e^{\lambda_i t} (\vec{w}_i^T \vec{\chi}).$$

The i'th mode is observable in the k'th output if the product $\vec{c}_k^T \vec{v}_i \neq 0$. We say that a system is observable if every mode can be seen in at least one output channel. The usual test for system observability requires computation of the observability matrix:

$$M_o = \left[C^T, A^T C^T, \cdots, (A^T)^{n-1} C^T \right]. \tag{199}$$

This matrix has size $n \times (rn)$; the system is observable if M_o has rank n.

18 CONTROL SYSTEMS – LOOPSHAPING

18.1 Introduction

This section formalizes the notion of loopshaping for linear control system design. The loopshaping approach is inherently two-fold. First, we shape the open-loop transfer function (or matrix) $P(s)C(s)$, to meet performance and robustness specifications. Once this is done, then the compensator must be computed, from from knowing the nominal product $P(s)C(s)$, and the nominal plant $P(s)$.
Most of the analysis here is given for single-input, single-output systems, but the link to multivariable control is not too difficult. In particular, absolute values of transfer functions are replaced with the maximum singular values of transfer matrices. Design based on singular values is the idea of L_2-control, or LQG/LTR, to be presented in the next lectures.

18.2 Roots of Stability – Nyquist Criterion

We consider the SISO feedback system with reference trajectory $r(s)$ and plant output $y(s)$, as given previously. The tracking error signal is defined as $e(s) = r(s) - y(s)$, thus forming the negative feedback loop. The sensitivity function is written as

$$S(s) = \frac{e(s)}{r(s)} = \frac{1}{1 + P(s)C(s)},$$

where $P(s)$ represents the plant transfer function, and $C(s)$ the compensator. The closed-loop *characteristic equation*, whose roots are the poles of the closed-loop system, is $1 + P(s)C(s) = 0$, equivalent to $\underline{P}(s)\underline{C}(s) + \overline{P}(s)\overline{C}(s) = 0$, where the underline and overline denote the denominator and numerator, respectively. The Nyquist criterion allows us to assess the stability properties of a system based on $P(s)C(s)$ only. This method for design involves plotting the complex loci of $P(s)C(s)$ for the range $\omega = [-\infty, \infty]$. There is no explicit calculation of the closed-loop poles, and in this sense the design approach is quite different from the root-locus method (see Ogata).

18.2.1 Mapping Theorem

We impose a reasonable assumption from the outset: The number of poles in $P(s)C(s)$ exceeds the number of zeros. It is a reasonable constraint because otherwise the loop transfer function could pass signals with infinitely high frequency. In the case of a PID controller (two zeros) and a second-order zero-less plant, this constraint can be easily met by adding a high-frequency rolloff to the compensator, the equivalent of low-pass filtering the error signal.

Let $F(s) = 1 + P(s)C(s)$. The heart of the Nyquist analysis is the mapping theorem, which answers the following question: How do paths in the s-plane map into paths in the F-plane? We limit ourselves to *closed, clockwise*(CW) paths in the s-plane, and the remarkable result of the mapping theorem is

Every zero of $F(s)$ enclosed in the s-plane generates exactly one CW encirclement of the origin in the $F(s)$-plane. Conversely, every pole of $F(s)$ enclosed in the s-plane generates exactly one CCW encirclement of the origin in the $F(s)$-plane. Since CW and CCW encirclements of the origin may cancel, the relation is often written $Z - P = CW$.

The trick now is to make the trajectory in the s-plane enclose all unstable poles, i.e., the path encloses the entire right-half plane, moving up the imaginary axis, and then proceeding to the right at an arbitrarily large radius, back to the negative imaginary axis.

Since the zeros of $F(s)$ are in fact the poles of the closed-loop transfer function, e.g., $S(s)$, stability requires that there are *no* zeros of $F(s)$ in the right-half s-plane. This leads to a slightly shorter form of the above relation:

$$P = CCW. \tag{200}$$

In words, stability requires that the number of unstable poles in $F(s)$ is equal to the number of CCW encirclements of the origin, as s sweeps around the entire right-half s-plane.

18.2.2 Nyquist Criterion

The Nyquist criterion now follows from one translation. Namely, encirclements of the origin by $F(s)$ are equivalent to encirclements of the point $(-1 + 0j)$ by $F(s) - 1$, or $P(s)C(s)$.

Then the stability criterion can be cast in terms of the *unstable poles of* $P(s)C(s)$, *instead of those of* $F(s)$:

$$P = CCW \longleftrightarrow \text{ closed-loop stability} \tag{201}$$

This is in fact the complete Nyquist criterion for stability. It is a necessary and sufficient condition that the number of unstable poles in the loop transfer function $P(s)C(s)$ must be matched by an equal number of CCW encirclements of the critical point $(-1 + 0j)$.
There are several details to keep in mind when making Nyquist plots:

- If neither the plant nor the controller have unstable modes, then the loci of $P(s)C(s)$ must not encircle the critical point at all.

- Because the path taken in the s-plane includes negative frequencies (i.e., the negative imaginary axis), the loci of $P(s)C(s)$ occur as complex conjugates – the plot is symmetric about the real axis.

- The requirement that the number of poles in $P(s)C(s)$ exceeds the number of zeros means that at high frequencies, $P(s)C(s)$ always decays such that the loci go to the origin.

- For the multivariable (MIMO) case, the procedure of looking at individual Nyquist plots for each element of a transfer matrix is unreliable and outdated. Referring to the multivariable definition of $S(s)$, we should count the encirclements for the function $[det(I + P(s)C(s)) - 1]$ instead of $P(s)C(s)$. The use of gain and phase margin in design is similar to the SISO case.

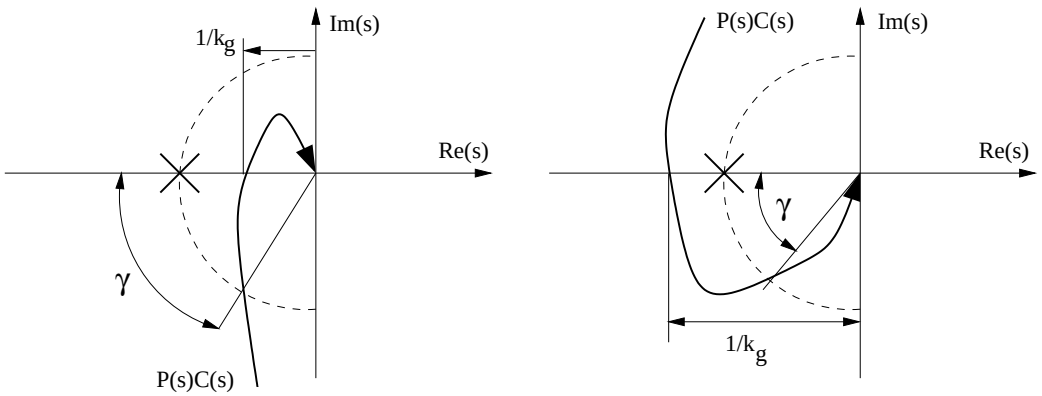
18.2.3 Robustness on the Nyquist Plot

The question of robustness in the presence of modelling errors is central to control system design. There are two natural measures of robustness for the Nyquist plot, each having a very clear graphical representation. The loci need to stay away from the critical point; how close the loci come to it can be expressed in terms of magnitude and angle.

- When the angle of $P(s)C(s)$ is $-180°$, the magnitude $|P(s)C(s)|$ should not be near one.

- When the magnitude $|P(s)C(s)| = 1$, its angle should not be $-180°$.

These notions lead to definition of the *gain margin* k_g and *phase margin* γ for a design. As the figure shows, the definition of k_g is different for stable and unstable $P(s)C(s)$. Rules of thumb are as follows. For a stable plant, $k_g \geq 2$ and $\gamma \geq 30°$; for an unstable plant, $k_g \leq 0.5$ and $\gamma \geq 30°$.



18.3 Design for Nominal Performance

Performance requirements of a feedback controller, using the nominal plant model, can be cast in terms of the Nyquist plot. We restrict the discussion to the scalar case in the following sections.

Since the sensitivity function maps reference input $r(s)$ to tracking error $e(s)$, we know that $|S(s)|$ should be small at low frequencies. For example, if one-percent tracking is to be maintained for all frequencies below $\omega = \lambda$, then $|S(s)| < 0.01, \forall \omega < \lambda$. This can be formalized by writing

$$|W_1(s)S(s)| < 1, \tag{202}$$

where $W_1(s)$ is a stable weighting function of frequency. To force $S(s)$ to be small at low ω, $W_1(s)$ should be large in the same range. The requirement $|W_1(s)S(s)| < 1$ is equivalent to $|W_1(s)| < |1 + P(s)C(s)|$, and this latter condition can be interpreted as: The loci of $P(s)C(s)$ must stay outside the disk of radius $W_1(s)$, which is to be centered on the critical point $(-1+0j)$. The disk is to be quite large, possibly infinitely large, at the lower frequencies.

18.4 Design for Robustness

It is a ubiquitous observation that models of plants degrade with increasing frequency. For example, the DC gain and slow, lightly-damped modes or zeros are easy to observe, but higher-frequency components in the response may be hard to capture or even excite repeatably. Higher-frequency behavior may have more nonlinear properties as well.

The effects of modeling uncertainty can be considered to enter the nominal feedback system as a disturbance at the plant output, d_y. One of the most useful descriptions of model uncertainty is the multiplicative uncertainty:

$$\tilde{P}(s) = (1 + \Delta(s)W_2(s))P(s). \tag{203}$$

Here, $P(s)$ represents the nominal plant model used in the design of the control loop, and $\tilde{P}(s)$ is the actual, perturbed plant. The perturbation is of the multiplicative type, $\Delta(s)W_2(s)P(s)$, where $\Delta(s)$ is an *unknown but stable* function of frequency for which $|\Delta(s)| \leq 1$. The weighting function $W_2(s)$ scales $\Delta(s)$ with frequency; $W_2(s)$ should be growing with increasing frequency, since the uncertainty grows. However, $W_2(s)$ should not grow any faster than necessary, since it will turn out to be at the cost of nominal performance. In the scalar case, the weight can be estimated as follows: since $\tilde{P}/P - 1 = \Delta W_2$, it will suffice to let $|\tilde{P}/P - 1| < |W_2|$.

Example: Let $\tilde{P} = k/(s-1)$, where k is in the range 2–5. We need to create a nominal model $P = k_0/(s-1)$, with the smallest possible value of W_2, which will not vary with frequency in this case. Two equations can be written using the above estimate, for the two extreme values of k, yielding $k_0 = 7/2$, and $W_2 = 3/7$.

For constructing the Nyquist plot, we observe that
$\tilde{P}(s)C(s) = (1 + \Delta(s)W_2(s))P(s)C(s)$. The path of the perturbed plant could be anywhere on a disk of radius $|W_2(s)P(s)C(s)|$, centered on the nominal loci $P(s)C(s)$. The robustness condition is that this disk should not intersect the critical point. This can be written as

$$\begin{aligned} |1 + PC| &> |W_2PC| \longleftrightarrow \\ 1 &> \frac{|W_2PC|}{|1 + PC|} \longleftrightarrow \\ 1 &> |W_2T|, \end{aligned} \tag{204}$$

where T is the complementary sensitivity function. The last inequality is thus a condition for robust stability in the presence of multiplicative uncertainty parametrized with W_2.

18.5 Robust Performance

The condition for good performance with plant uncertainty is a combination of the above two conditions. Graphically, the disk at the critical point, with radius $|W_1|$, should not intersect the disk of radius $|W_2PC|$, centered on the nominal locus PC. This is met if

$$|W_1S| + |W_2T| < 1. \tag{205}$$

The robust performance requirement is related to the magnitude $|PC|$ at different frequencies, as follows:

1. At low frequency, $|W_1S| \simeq |W_1/PC|$, since $|PC|$ is large. This leads directly to the performance condition $|PC| > |W_1|$ in this range.

2. At high frequency, $W_2T| \simeq |W_2PC|$, since $|PC|$ is small. We must therefore have $|PC| < 1/|W_2|$, for robustness.

18.6 Implications of Bode's Integral

The loop transfer function PC cannot roll off too rapidly in the crossover region. The simple reason is that a steep slope induces a large phase loss, which in turn degrades the phase margin. To see this requires a short foray into Bode's integral. For a transfer function $H(s)$, the crucial relation is

$$angle(H(j\omega_0)) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{d}{d\nu}(ln|H(j\omega)) \cdot ln(coth(|\nu|/2))d\nu, \tag{206}$$

where $\nu = ln(\omega/\omega_0)$. The integral is hence taken over the log of a frequency normalized with ω_0. It is not hard to see how the integral controls the angle: the function $ln(coth(|\nu|/2))$ is nonzero only near $\nu = 0$, implying that the angle depends only on the local slope $d(ln|H|)/d\nu$. Thus, if the slope is large, the angle is large.

Example: Suppose $H(s) = \omega_0^n/s^n$, i.e., it is a simple function with n poles at the origin, and no zeros; ω_0 is a fixed constant. It follows that $|H| = \omega_0^n/\omega^n$, and $ln|H| = -nln(\omega/\omega_0)$, so that $d(ln|H|)/d\nu = -n$. Then we have just

$$angle(H) = -\frac{n}{\pi}\int_{-\infty}^{\infty} ln(coth(|\nu|/2))d\nu = -\frac{n\pi}{2}. \tag{207}$$

This integral is trivial to look up or compute. Each pole at the origin clearly induces 90° of phase loss. In the general case, each pole not at the origin induces 90° of phase loss for frequencies above the pole. Each zero at the origin adds 90° phase lead, while zeros not at the origin at 90° of phase lead for frequencies above the zero. In the immediate neighborhood of these poles and zeros, the phase may vary significantly with frequency.
The Nyquist loci are clearly susceptible to these variations is phase, and the phase margin can be easily lost if the slope of PC at crossover (where the magnitude is unity) is too steep. The slope can safely be first-order ($-20dB/decade$, equivalent to a single pole), and may be second-order
($-40dB/decade$) if an adequate phase angle can be maintained near crossover.

18.7 The Recipe for Loopshaping

In the above analysis, we have extensively described what the open loop transfer function PC should look like, to meet robustness and performance specifications. We have said very little about how to get the compensator C, the critical component. For clarity, let the designed loop transfer function be renamed, $L = PC$. We will use concepts from optimal linear control for the MIMO case, but in the scalar case, it suffices to just pick

$$C = L/P. \tag{208}$$

This extraordinarily simple step involves a plant inversion.
The overall idea is to first shape L as a stable transfer function meeting the requirements of stability and robustness, and then divide through by the plant.

- When the plant is stable and has stable zeros (minimum-phase), the division can be made directly.

- One caveat for the well-behaved plant is that lightly-damped poles or zeros should not be cancelled verbatim by the compensator, because the closed-loop response will be sensitive to any slight change in the resonant frequency. The usual procedure is to widen the notch or pole in the compensator, through a higher damping ratio.

- Non-minimum phase or unstable behavior in the plant can usually be handled by performing the loopshaping for the closest stable model, and then explicitly considering the effects of adding the unstable parts. In the case of unstable zeros, we find that they impose an unavoidable frequency limit for the crossover. In general, the troublesome zeros must be *faster* than the closed-loop frequency response.

  In the case of unstable poles, the converse is true: The feedback system must be faster than the corresponding frequency of the unstable mode.

When a control system involves multiple inputs and outputs, the ideas from scalar loopshaping can be adapted using the singular value. We list below some basic properties of the singular value decomposition, which is analogous to an eigenvector, or modal, analysis. Useful properties and relations for the singular value are found in the section *MATH FACTS*. The condition for MIMO robust performance can be written in many ways, including a direct extension of our scalar condition

$$\overline{\sigma}(W_1 S) + \overline{\sigma}(W_2 T) < 1. \tag{209}$$

The open-loop transfer matrix L should be shaped accordingly. In the following sections, we use the properties of optimal state estimation and control to perform the plant inversion for MIMO systems.

19 LINEAR QUADRATIC REGULATOR

19.1 Introduction

The simple form of loopshaping in scalar systems does not extend directly to multivariable (MIMO) plants, which are characterized by transfer matrices instead of transfer functions. The notion of optimality is closely tied to MIMO control system design. Optimal controllers, i.e., controllers that are the *best* possible, according to some figure of merit, turn out to generate only stabilizing controllers for MIMO plants. In this sense, optimal control solutions provide an automated design procedure – we have only to decide what figure of merit to use. The linear quadratic regulator (LQR) is a well-known design technique that provides practical feedback gains.

19.2 Full-State Feedback

For the derivation of the linear quadratic regulator, we assume the plant to be written in state-space form $\dot{x} = Ax + Bu$, and that *all* of the n states x are available for the controller. The feedback gain is a matrix K, implemented as $u = -K(x - x_{desired})$. The system dynamics are then written as:

$$\dot{x} = (A - BK)x + BKx_{desired}. \tag{210}$$

$x_{desired}$ represents the vector of desired states, and serves as the external input to the closed-loop system. The "A-matrix" of the closed loop system is $(A - BK)$, and the "B-matrix" of the closed-loop system is BK. The closed-loop system has exactly as many outputs as inputs: n. The column dimension of B equals the number of channels available in u, and must match the row dimension of K. Pole-placement is the process of placing the poles of $(A - BK)$ in stable, suitably-damped locations in the complex plane.

19.3 The Maximum Principle

First we develop a general procedure for solving optimal control problems, using the calculus of variations. We begin with the statement of the problem for a fixed end time t_f:

$$\begin{aligned} \text{choose } u(t) \text{ to minimize } J &= \psi(x(t_f)) + \int_{t_o}^{t_f} L(x(t), u(t), t)dt && (211) \\ \text{subject to } \dot{x} &= f(x(t), u(t), t) && (212) \\ x(t_o) &= x_o && (213) \end{aligned}$$

where $\psi(x(t_f), t_f)$ is the *terminal cost*; the total cost J is a sum of the terminal cost and an integral along the way. We assume that $L(x(t), u(t), t)$ is nonnegative. The first step is to augment the cost using the costate vector $\lambda(t)$.

$$\bar{J} = \psi(x(t_f)) + \int_{t_o}^{t_f} (L + \lambda^T(f - \dot{x}))dt \tag{214}$$

Clearly, $\lambda(t)$ can be anything we choose, since it multiplies $f - \dot{x} = 0$. Along the optimum trajectory, variations in J and hence $\bar{J}$ should vanish. This follows from the fact that J is chosen to be continuous in x, u, and t. We write the variation as

$$\delta\bar{J} = \psi_x \delta x(t_f) + \int_{t_o}^{t_f} \left[L_x \delta x + L_u \delta u + \lambda^T f_x \delta x + \lambda^T f_u \delta u - \lambda^T \delta\dot{x} \right] dt, \tag{215}$$

where subscripts denote partial derivatives. The last term above can be evaluated using integration by parts as

$$-\int_{t_o}^{t_f} \lambda^T \delta\dot{x} dt = -\lambda^T(t_f)\delta x(t_f) + \lambda^T(t_o)\delta x(t_o) + \int_{t_o}^{t_f} \dot{\lambda}^T \delta x dt, \tag{216}$$

and thus

$$\begin{aligned}\delta\bar{J} &= \psi_x(x(t_f))\delta x(t_f) + \int_{t_o}^{t_f}(L_u + \lambda^T f_u)\delta u dt + \\ &\int_{t_o}^{t_f}(L_x + \lambda^T f_x + \dot{\lambda}^T)\delta x dt - \lambda^T(t_f)\delta x(t_f) + \lambda^T(t_o)\delta x(t_o).\end{aligned} \tag{217}$$

Now the last term is zero, since we cannot vary the initial condition of the state by changing something later in time - it is a fixed value. This way of writing $\bar{J}$ makes it clear that there are three components of the variation that must independently be zero (since we can vary any of x, u, or $x(t_f)$):

$$L_u + \lambda^T f_u = 0 \tag{218}$$
$$L_x + \lambda^T f_x + \dot{\lambda}^T = 0 \tag{219}$$
$$\psi_x(x(t_f)) - \lambda^T(t_f) = 0. \tag{220}$$

The second and third requirements are met by explicitly setting

$$\dot{\lambda}^T = -L_x - \lambda^T f_x \tag{221}$$
$$\lambda_T(t_f) = \psi_x(x(t_f)). \tag{222}$$

The evolution of λ is given in *reverse time*, from a final state to the initial. Hence we see the primary difficulty of solving optimal control problems: the state propogates forward in time, while the costate propogates backward. The state and costate are coordinated through the above equations.

19.4 Gradient Method Solution for the General Case

Numerical solutions to the general problem are iterative, and the simplest approach is the gradient method. It is outlined as follows:

1. For a given x_o, pick a control history $u(t)$.
2. Propogate $\dot{x} = f(x, u, t)$ forward in time to create a state trajectory.
3. Evaluate $\psi_x(x(t_f))$, and the propogate the costate backward in time from t_f to t_o, using Equation 221.
4. At each time step, choose $\delta u = -K(L_u + \lambda^T f_u)$, where K is a positive scalar or a positive definite matrix in the case of multiple input channels.
5. Let $u = u + \delta u$.
6. Go back to step 2 and repeat loop until solution has converged.

The first three steps are consistent in the sense that x is computed directly from $x(t_o$ and u, and λ is computed directly from x and $x(t_f)$. All of $\delta\bar{J}$ in Equation 217 except the integral with δu is therefore eliminated explicitly. The choice of δu in step 4 then achieves $\delta\bar{J} < 0$, unless $\delta u = 0$, in which case the problem is solved.
The gradient method is quite popular in applications and for complex problems may be the only way to effectively develop a solution. The major difficulties are computational expense, and the requirement of having a reasonable control trajectory to begin.

19.5 LQR Solution

In the case of the Linear Quadratic Regulator (with zero terminal cost), we set $\psi = 0$, and

$$L = \frac{1}{2}x^T Q x + \frac{1}{2}u^T R u, \tag{223}$$

where the requirement that $L \geq 0$ implies that both Q and R are positive definite. In the case of linear plant dynamics also, we have

$$\begin{aligned} L_x &= x^T Q & (224)\\ L_u &= u^T R & (225)\\ f_x &= A & (226)\\ f_u &= B, & (227) \end{aligned}$$

so that

$$\begin{aligned} \dot{x} &= Ax + Bu & (228)\\ x(t_o) &= x_o & (229)\\ \dot{\lambda} = -Qx - A^T\lambda & & (230)\\ \lambda(t_f) &= 0 & (231)\\ Ru + B^T\lambda &= 0. & (232) \end{aligned}$$

Since the systems are clearly linear, we try a connection $\lambda = Px$. Inserting this into the $\dot{\lambda}$ equation, and then using the $\dot{x}$ equation, and a substitution for u, we obtain

$$PAx + A^T Px + Qx - PBR^{-1}B^T Px + \dot{P} = 0. \tag{233}$$

This has to hold for all x, so in fact it is a matrix equation, the *matrix Riccati equation*. The steady-state solution is given satisfies

$$PA + A^T P + Q - PBR^{-1}B^T P = 0. \tag{234}$$

19.6 Optimal Full-State Feedback

This equation is the *matrix algebraic Riccati equation* (MARE), whose solution P is needed to compute the optimal feedback gain K. The MARE is easily solved by standard numerical tools in linear algebra.
The equation $Ru + B^T\lambda = 0$ gives the feedback law:

$$u = -R^{-1}B^T Px. \tag{235}$$

19.7 Properties and Use of the LQR

Static Gain. The LQR generates a static gain matrix K, which is not a dynamical system. Hence, the order of the closed-loop system is the same as that of the plant.

Robustness. The LQR achieves infinite gain margin: $k_g = \infty$, implying that the loci of (PC) (scalar case) or $(det(I+PC)-1)$ (MIMO case) approach the origin along the imaginary axis. The LQR also guarantees phase margin $\gamma \geq 60$ degrees. This is in good agreement with the practical guidelines for control system design.

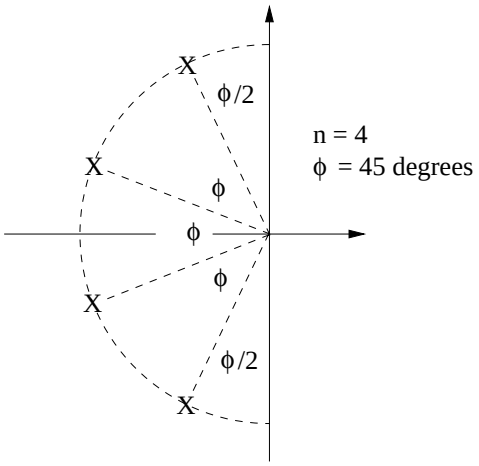
Output Variables. In many cases, it is not the states x which are to be minimized, but the output variables y. In this case, we set the weighting matrix $Q = C^T Q' C$, since $y = Cx$, and the auxiliary matrix Q' weights the plant output.

Behavior of Closed-Loop Poles: Expensive Control. When $R >> C^T Q' C$, the cost function is dominated by the control effort u, and so the controller minimizes the control action itself. In the case of a completely stable plant, the gain will indeed go to zero, so that the closed-loop poles approach the open-loop plant poles in a manner consistent with the scalar root locus.

The optimal control *must* always stabilize the closed-loop system, however, so there should be some account made for unstable plant poles. The expensive control solution puts stable closed-loop poles at the *mirror* images of the unstable plant poles.

Behavior of Closed-Loop Poles: Cheap Control. When $R << C^T Q' C$, the cost function is dominated by the output errors y, and there is no penalty for using large u. There are two groups of closed-loop poles. First, poles are placed at stable plant zeros, and at the mirror images of the unstable plant zeros. This part is akin to the high-gain limiting case of the root locus. The remaining poles assume a Butterworth pattern, whose radius increases to infinity as R becomes smaller and smaller.

The Butterworth pattern refers to an arc in the stable left-half plane, as shown in the figure. The angular separation of n closed-loop poles on the arc is constant, and equal to $180°/n$. An angle $90°/n$ separates the most lightly-damped poles from the imaginary axis.



19.8 Proof of the Gain and Phase Margins

The above stated properties of the LQR control are not difficult to prove. We need an intermediate tool (which is very useful in its own right) to prove it: the Liapunov function. Consider a scalar, continuous, positive definite function of the state $V(x)$. If this function is always decreasing, except at the origin, then the origin is asymptotically stable. There is an intuitive feel to the Liapunov function; it is like the energy of a physical system. If it can be shown that energy is leaving such a system, then it will "settle." It suffices to find one Liapunov function to show that a system is stable.
Now, in the case of the LQR control, pick
$V(x) = \frac{1}{2}x^T P x$.
It can be shown that P is positive definite; suppose that instead of the design gain B, we have actual gain B', giving (with constant P based on the design system)

$$\begin{aligned} \dot{V} &= 2x^T P\dot{x} && (236)\\ &= 2x^T P(Ax - B'R^{-1}B^T Kx) && (237)\\ &= x^T(-Q + PBR^{-1}B^T P)x - 2x^T KB'R^{-1}B^T Px. && (238) \end{aligned}$$

where we used the Riccati equation in a substitution. Since we require $\dot{V} < 0$ for $x \neq 0$, we must have

$$Q + PBG - 2P(B - B')G > 0, \tag{239}$$

where $G = R^{-1}B^T P$. Let $B' = BN$, where N is a diagonal matrix of random, unknown gains on the control channels. The above equation becomes

$$Q - PB(I - 2N)R^{-1}B^T P > 0. \tag{240}$$

This is satisfied if, for all channels, $N_{i,i} > 1/2$. Thus, we see that the LQR provides for one-half gain reduction, and infinite gain amplification, in all channels.

The phase margin is seen (roughly) by allowing N to be a diagonal matrix of transfer functions $N_{i,i} = e^{j\phi_i}$. In this case, we require that the *real part* of $N_{i,i}$ is greater than one-half. This is true for all $|\phi_i| < 60°$, and hence sixty degrees of phase margin is provided in all channels.

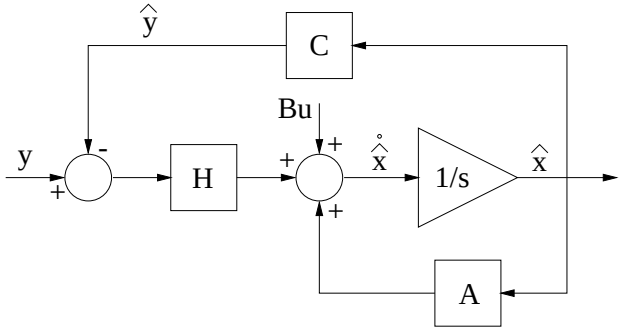
20 KALMAN FILTER

20.1 Introduction

In the previous section, we derived the linear quadratic regulator as an optimal solution for the full-state feedback control problem. The inherent assumption was that each state was known perfectly. In real applications, the measurements are subject to disturbances, and may not allow reconstruction of all the states. This *state estimation* is the task of a model-based estimator having the form:

$$\dot{\hat{x}} = A\hat{x} + Bu + H(y - C\hat{x}) \tag{241}$$

The vector $\hat{x}$ represents the state estimate, whose evolution is governed by the nominal A and B matrices of the plant, and a correction term with the estimator gain matrix H. H operates on the estimation error mapped to the plant output y, since $C\hat{x} = \hat{y}$. Given statistical properties of real plant disturbances and sensor noise, the *Kalman Filter* designs an optimal H.



20.2 Problem Statement

We consider the state-space plant model given by:

$$\begin{aligned} \dot{x} &= Ax + Bu + W_1 \\ y &= Cx + W_2. \end{aligned} \tag{242}$$

There are n states, m inputs, and l outputs, so that A has dimension $n \times n$, B is $n \times m$, and C is $l \times n$. The plant subject to two random input signals, W_1 and W_2. W_1 represents

disturbances to the plant, since it drives $\dot{x}$ directly; W_2 denotes sensor noise, which corrupts the measurement y.
An important assumption of the Kalman Filter is that W_1 and W_2 are each vectors of unbiased, independent white noise, and that all the $n + l$ channels are uncorrelated. Hence, if $E(\cdot)$ denotes the expected value,

$$\begin{aligned} E(W_1(t_1)W_1(t_2)^T) &= V_1\delta(t_1 - t_2) && (243)\\ E(W_2(t_1)W_2(t_2)^T) &= V_2\delta(t_1 - t_2) && (244)\\ E(W_1(t)W_2(t)^T) &= 0_{n\times l}. && (245)\end{aligned}$$

Here $\delta(t)$ represents the impulse (or delta) function. V_1 is an $n \times n$ diagonal matrix of intensities, and V_2 is an $l \times l$ diagonal matrix of intensities.
The estimation error may be defined as $e = x - \hat{x}$. It can then be verified that

$$\begin{aligned} \dot{e} &= [Ax + Bu + W_1] - [A\hat{x} + Bu + H(y - C\hat{x})] && (246)\\ &= (A - HC)e + (W_1 - HW_2). \end{aligned}$$

The eigenvalues of the matrix $A - HC$ thus determine the stability properties of the estimation error dynamics. The second term above, $W_1 + HW_2$ is considered an external input.
The Kalman filter design provides H that minimizes the scalar cost function

$$J = E(e^T W e), \tag{247}$$

where W is an unspecified symmetric, positive definite weighting matrix. A related matrix, the symmetric *error covariance*, is defined as

$$\Sigma = E(ee^T). \tag{248}$$

There are two main steps for deriving the optimal gain H.

20.3 Step 1: An Equation for $\dot{\Sigma}$

The evolution of Σ follows from some algebra and the convolution form of $e(t)$. We begin with

$$\begin{aligned} \dot{\Sigma} &= E(\dot{e}e^T + e\dot{e}^T) && (249)\\ &= E[(A - HC)ee^T + (W_1 - HW_2)e^T + ee^T(A^T - C^T H^T) + \\ &\quad e(W_1^T - W_2^T H^T)]. \end{aligned}$$

The last term above can be expanded, using the property that

$$e(t) = e^{(A-HC)t}e(0) + \int_0^t e^{(A-HC)(t-\tau)}(W_1(\tau) - HW_2(\tau))\, d\tau.$$

We have

$$\begin{aligned} E(e(W_1^T - W_2^T H^T)) &= e^{(A-HC)t}E(e(0)(W_1^T - W_2^T H^T)) + \\ &\quad \int_0^t e^{(A-HC)(t-\tau)}E((W_1(\tau) - HW_2(\tau)) \\ &\quad\quad (W_1^T(t) - W_2^T(t)H^T))\, d\tau \\ &= \int_0^t e^{(A-HC)(t-\tau)}E((W_1(\tau) - HW_2(\tau)) \\ &\quad\quad (W_1^T(t) - W_2^T(t)H^T))\, d\tau \\ &= \int_0^t e^{(A-HC)(t-\tau)}(V_1\delta(t-\tau) + HV_2H^T\delta(t-\tau))\, d\tau \\ &= \frac{1}{2}V_1 + \frac{1}{2}HV_2H^T. \end{aligned}$$

To get from the first right-hand side to the second, we note that the initial condition $e(0)$ is uncorrelated with $W_1^T - W_2^T H^T$. The fact that W_1 and HW_2 are uncorrelated leads to the third line, and the final result follows from

$$\int_0^t \delta(t-\tau)d\tau = \frac{1}{2},$$

i.e., the written integral includes only *half* of the impulse.
The final expression for $E(e(W_1^T - W_2^T H^T))$ is symmetric, and therefore appears in Equation 249 twice, leading to

$$\dot{\Sigma} = (A - HC)\Sigma + \Sigma(A^T - C^T H^T) + V_1 + HV_2H^T. \tag{250}$$

This equation governs propagation of the error covariance. It is independent of the initial condition $e(0)$, and depends on the (as yet) unknown estimator gain matrix H.

20.4 Step 2: H as a Function of Σ

We now make the connection between $\Sigma = E(ee^T)$ (a matrix) and $J = E(e^T We)$ (a scalar). The *trace* of a matrix is the sum of its diagonal elements, and it can be verified that

$$J = E(e^T We) = trace(\Sigma W). \tag{251}$$

We now introduce an auxiliary cost function defined as

$$J' = trace(\Sigma W + \Lambda F), \tag{252}$$

where F is an $n \times n$ matrix of zeros, and Λ is an $n \times n$ matrix of unknown Lagrange multipliers. Note that since F is zero, $J' = J$, so minimizing J' solves the same problem. Lagrange multipliers provide an ingenious mechanism for drawing constraints into the optimization; the constraint we invoke is the evolution of Σ, Equation 250:

$$J' = trace\left(\Sigma W + \Lambda(-\dot{\Sigma} + A\Sigma - HC\Sigma + \Sigma A^T - \Sigma C^T H^T + V_1 + HV_2H^T)\right) \tag{253}$$

If J' is an optimal cost, it follows that $\partial J'/\partial H = 0$, i.e., the correct choice of H achieves an extremal value. We need the following lemmas, which give the derivative of a trace with respect to a constituent matrix:

$$\begin{aligned} \frac{\partial}{\partial H} trace(-\Lambda HC\Sigma) &= -\Lambda^T \Sigma C^T \\ \frac{\partial}{\partial H} trace(-\Lambda \Sigma C^T H^T) &= -\Lambda \Sigma C^T \\ \frac{\partial}{\partial H} trace(\Lambda HV_2 H^T) &= \Lambda^T HV_2 + \Lambda HV_2. \end{aligned}$$

Proofs of the first two are given at the end of this section; the last lemma uses the chain rule, and the previous two lemmas. Next, we enforce $\Lambda = \Lambda^T$, since the values are arbitrary. Then the condition $\partial J'/\partial H = 0$ leads to

$$\begin{aligned} 0 &= 2\Lambda(-\Sigma C^T + HV_2), \text{ satisfied if} \\ H &= \Sigma C^T V_2^{-1}. \end{aligned} \tag{254}$$

Hence the estimator gain matrix H can be written as a function of Σ. Inserting this back into Equation 250, we obtain

$$\dot{\Sigma} = A\Sigma + \Sigma A^T + V_1 - \Sigma C^T V_2^{-1} C\Sigma. \tag{255}$$

Equations 254 and 255 represent the practical solution to the Kalman filtering problem, which minimizes the squared-norm of the estimation error. The evolution of Σ is always stable, and depends only on the constant matrices $[A, C, V_1, V_2]$. Notice also that the result is independent of the weighting matrix W, which might as well be the identity.

20.5 Properties of the Solution

The solution involves a matrix Riccati equation, like the LQR, suggesting a duality with the LQR problem. This is in fact the case, and the same analysis and numerical tools can be applied to both methodologies.

The steady-state solution for Σ is valid for time-invariant systems, leading to a more common MARE form of Equation 255:

$$0 = A\Sigma + \Sigma A^T + V_1 - \Sigma C^T V_2^{-1} C\Sigma. \tag{256}$$

Duality of Linear Quadratic Regulator and Kalman Filter

Linear Quadratic Regulator	Kalman Filter
$\dot{x} = Ax + Bu$	$\dot{x} = Ax + Bu + W_1$
	$y = Cx + W_2$
$u = -Kx$	$\dot{\hat{x}} = A\hat{x} + Bu + H(y - C\hat{x})$
$2J = \int_0^\infty (x^T Q x + u^T R u) dt$	$J = E(e^T W e)$
$Q \geq 0, R > 0$	$V_1 \geq 0$, $V_2 > 0$
$K = R^{-1} B^T P$	$H = \Sigma C^T V_2^{-1}$
$PA + A^T P + Q - PBR^{-1}B^T P = 0$	$\Sigma A^T + A\Sigma + V_1 - \Sigma C^T V_2^{-1} C\Sigma = 0$

The Kalman Filter is guaranteed to create a stable nominal dynamics $A - HC$, as long as the plant is fully state-observable. This is dual to the stability guarantee of the LQR loop, when the plant is state-controllable. Furthermore, like the LQR, the KF loop achieves 60° phase margin, and infinite gain margin, for all the channels together or independently.
The qualitative dependence of the estimator gain $H = \Sigma C^T V_2^{-1}$ on the other parameters can be easily seen. Recall that V_1 is the intensity matrix of the plant disturbance, V_2 is the intensity of the sensor noise, and Σ is the error covariance matrix.

- A large uncertainty Σ creates large H, placing emphasis on the corrective action of the filter.
- A small disturbance V_1, and large sensor noise V_2 creates a small H, weighting the model dynamics $A\hat{x} + Bu$ more.
- A large disturbance V_1, and small sensor noise V_2 creates a large H, so that the filter's correction is dominant.

The limiting closed-loop poles of the Kalman filter are similar, and dual to those of the LQR:

- $V_2 << V_1$: good sensors, large disturbance, $H >> 1$, *dual to cheap-control problem.* Some closed-loop poles go to the stable plant zeros, or the mirror image of unstable plant zeros. The remaining poles follow a Butterworth pattern whose radius increases with increasing V_1/V_2.
- $V_2 >> V_1$: poor sensors, small disturbance, H small, *dual to expensive-control problem.* Closed-loop poles go to the stable plant poles, and the mirror images of the unstable plant poles.

20.6 Combination of LQR and KF

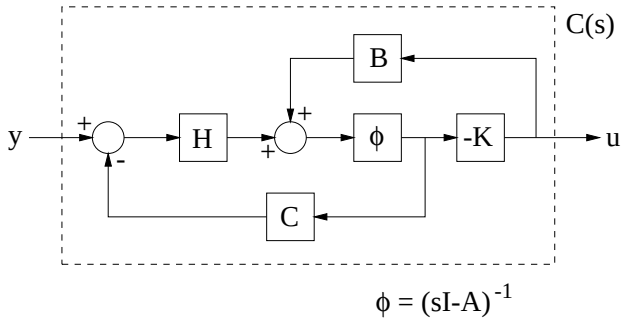
An optimal output feedback controller is created through the use of a Kalman filter coupled with an LQR full-state feedback gain. This combination is usually known as the Linear Quadratic Gaussian design, or LQG. For the plant given as

$$\begin{aligned} \dot{x} &= Ax + Bu + W_1 \\ y &= Cx + W_2, \end{aligned}$$

we put the Kalman Filter and controller gain G together as follows:

$$\dot{\hat{x}} = A\hat{x} + Bu + H(y - C\hat{x}) \quad (257)$$

$$u = -K\hat{x}. \quad (258)$$



There are two central points to this construction:

1. **Separation Principle**: The eigenvalues of the nominal closed-loop system are made of up the eigenvalues of $(A - HC)$ and the eigenvalues of $(A - BK)$, separately. See proof below.

2. **Output Tracking:** This compensator is a stand-alone system that, as written, tries to drive its input y to zero. It can be hooked up to receive tracking error $e(s) = r(s) - y(s)$ as an input instead, so that it is not limited to the regulation problem alone. In this case, $\hat{x}$ no longer represents an estimated *state*, but rather an estimated state tracking error. We use the output error as a control input in the next section, on loopshaping via loop transfer recovery.

20.7 Proofs of the Intermediate Results

20.7.1 Proof that $E(e^T We) = trace(\Sigma W)$

$$\begin{aligned} E(e^T We) &= E\left(\sum_{i=1}^{n} e_i \left(\sum_{j=1}^{n} W_{ij} e_j\right)\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \Sigma_{ij} W_{ji}, \end{aligned}$$

the transpose of W being valid since it is symmetric. Now consider the diagonal elements of the product ΣW:

$$\begin{aligned}
\Sigma W &= \begin{bmatrix} \Sigma_{11}W_{11} + \Sigma_{12}W_{21} + \cdots & \cdot & \cdot \\ \cdot & \Sigma_{21}W_{12} + \Sigma_{22}W_{22} + \cdots & \cdot \\ \cdot & \cdot & \cdots \end{bmatrix} \rightarrow \\
trace(\Sigma W) &= \sum_{i=1}^{n}\sum_{j=1}^{n} \Sigma_{ij}W_{ji}. \quad \square
\end{aligned}$$

20.7.2 Proof that $\frac{\partial}{\partial H} trace(-\Lambda H C \Sigma) = -\Lambda^T \Sigma C^T$

$$\begin{aligned}
trace(AHB) &= trace\left[\sum_{j=1}^{n} A_{ij} \sum_{k=1}^{l} H_{jk}B_{kl}\right], \text{ the } il'\text{th element} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij} \sum_{k=1}^{l} H_{jk}B_{ki},
\end{aligned}$$

where the second form is a sum over i of the ii'th elements. Now

$$\begin{aligned}
\frac{\partial}{\partial H_{j_o k_o}} trace(AHB) &= \sum_{i=1}^{n} A_{ij_o}B_{k_o i} \\
&= (BA)_{k_o j_o} \\
&= (BA)^T_{j_o k_o} \longrightarrow \\
\frac{\partial}{\partial H} trace(AHB) &= (BA)^T \\
&= A^T B^T. \quad \square
\end{aligned}$$

20.7.3 Proof that $\frac{\partial}{\partial H} trace(-\Lambda \Sigma C^T H^T) = -\Lambda \Sigma C^T$

$$\begin{aligned}
trace(AH^T) &= trace\left[\sum_{j=1}^{n} A_{ij}H^T_{jl}\right], \text{ the } il'\text{th element} \\
&= trace\left[\sum_{j=1}^{n} A_{ij}H_{lj}\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}H_{ij},
\end{aligned}$$

where the last form is a sum over the ii'th elements. It follows that

$$\begin{aligned} \frac{\partial}{\partial H_{i_o j_o}} trace(AH^T) &= A_{i_o j_o} \longrightarrow \\ \frac{\partial}{\partial H} trace(AH^T) &= A. \quad \square \end{aligned}$$

20.7.4 Proof of the Separation Principle

Without external inputs W_1 and W_2, the closed-loop system evolves according to

$$\frac{d}{dt} \left\{ \begin{array}{c} x \\ \hat{x} \end{array} \right\} = \left[\begin{array}{cc} A & -BK \\ HC & A - BK - HC \end{array} \right] \left\{ \begin{array}{c} x \\ \hat{x} \end{array} \right\}.$$

Using the definition of estimation error $e = x - \hat{x}$, we can write the above in another form:

$$\frac{d}{dt} \left\{ \begin{array}{c} x \\ e \end{array} \right\} = \left[\begin{array}{cc} A - BK & BK \\ 0 & A - HC \end{array} \right] \left\{ \begin{array}{c} x \\ e \end{array} \right\}.$$

If A' represents this compound A-matrix, then its eigenvalues are the roots of $det(sI - A') = 0$. However, the determinant of an upper triangular block matrix is the product of the determinants of each block on the diagonal: $det(sI - A') = det(sI - (A - BK))det(sI - (A - HC))$, and hence the separation of eigenvalues follows.

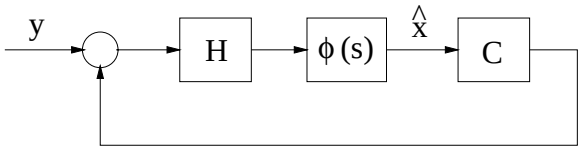
21 LOOP TRANSFER RECOVERY

21.1 Introduction

The Linear Quadratic Regulator(LQR) and Kalman Filter (KF) provide practical solutions to the full-state feedback and state estimation problems, respectively. If the sensor noise and disturbance properties of the plant are indeed well-known, then an LQG design approach, that is, combining the LQR and KF into an output feedback compensator, may yield good results. The LQR tuning matrices Q and R would be picked heuristically to give a reasonable closed-loop response.

There are two reasons to avoid this kind of direct LQG design procedure, however. First, although the LQR and KF each possess good robustness properties, there do exist plants for which there is *no* robustness guarantee for an LQG compensator. Even if one could steer clear of such pathological cases, a second problem is that this design technique has no clear equivalent in frequency space. It cannot be directly mapped to the intuitive ideas of loopshaping and the Nyquist plot, which are at the root of feedback control.

We now reconsider just the feedback loop of the Kalman filter. The KF has open-loop transfer function $L(s) = C\phi(s)H$, where $\phi(s) = (sI - A)^{-1}$. This follows from the estimator evolution equation



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

and the figure. Note that we have not included the factor Bu as part of the figure, since it does not affect the error dynamics of the filter.
As noted previously, the KF loop has good robustness properties, specifically to perturbations at the output $\hat{y}$, and further is amenable to output tracking. In short, the KF loop is an ideal candidate for a loopshaping design. Supposing that we have an estimator gain H which creates an attractive loop function $L(s)$, we would like to find the compensator $C(s)$ that establishes

$$\begin{aligned} P(s)C(s) &\approx C\phi(s)H, \text{ or} \\ C\phi(s)BC(s) &\approx C\phi(s)H. \end{aligned} \tag{259}$$

It will turn out that the LQR can be set up so that the an LQG-type compensator achieves exactly this result. The procedure is termed Loop Transfer Recovery (LTR), and has two main parts. First, one carries out a KF design for H, so that the Kalman filter loop itself has good performance and robustness properties. In this regard, the KF loop has sensitivity function $S(s) = (I + C\phi(s)H)^{-1}$ and complementary sensitivity $T(s) = (I + C\phi(s)H)^{-1}C\phi(s)H$. The condition $\overline{\sigma}(W_1(s)S(s)) + \overline{\sigma}(W_2(s)T(s)) < 1$ is sufficient for robust performance with multiplicative plant uncertainty at the output. Secondly, we pick suitable parameters of the LQR design, so that the LQG compensator satisfies the approximation of Equation 259.
LTR is useful as a SISO control technique, but has a much larger role in multivariable control.

21.2 A Special Property of the LQR Solution

Letting $Q = C^TC$ and $R = \rho I$, where I is the identity matrix, we will show (roughly) that

$$\lim_{\rho \to 0}(\sqrt{\rho}K) = WC,$$

where K is the LQR gain matrix, and W is an orthonormal matrix, for which $W^TW = I$.
First recall the gain and Riccati equations for the LQR:

$$\begin{aligned} K &= R^{-1}B^TP \\ 0 &= Q + PA + A^TP - PBR^{-1}B^TP. \end{aligned}$$

Now $Q = C^T C = C^T W^T W C = (WC)^T WC$. The Riccati equation becomes

$$0 = \rho(WC)^T WC + \rho PA + \rho A^T P - PBB^T P = 0.$$

In the limit as $\rho \to 0$, it must be the case that $P \to 0$ also, and so in this limit

$$\begin{aligned} \rho(WC)^T WC &\approx PBB^T P \\ &= (B^T P)^T B^T P \\ &= (R^{-1}B^T P)^T RR(R^{-1}B^T P) \\ &= \rho^2 K^T K \longrightarrow \\ WC &\approx \sqrt{\rho} K. \qquad \square \end{aligned}$$

Note that another orthonormal matrix W' could be used in separating K^T from K in the last line. This matrix may be absorbed into W through a matrix inverse, however, and so does not need to be written. The result of the last line establishes that the plant must be square: the number of inputs (i.e., rows of K) is equal to the number of outputs (i.e., rows of C).
Finally, we note that the above property is true only for LQR designs with minimum-phase plants, i.e., those with only stable zeros (Kwakernaak and Sivan).

21.3 The Loop Transfer Recovery Result

The theorem is stated as: If $\lim_{\rho \to 0}(\sqrt{\rho} K) = WC$ (the above result), with W an orthonormal matrix, then the limiting LQG controller $C(s)$ satisfies

$$\lim_{\rho \to 0} P(s)C(s) = C\phi(s)H.$$

The LTR method is limited by two conditions:

- The plant has an equal number of inputs and outputs.
- The design plant has no unstable zeros. The LTR method can be in fact be applied in the presence of unstable plant zeros, but the recovery is not to the Kalman filter loop transfer function. Instead, the recovered function will exhibit reasonable limitations inherent to unstable zeros. See Athans for more details and references on this topic.

The proof of the LTR result depends on some easy lemmas, given at the end of this section. First, we develop $C(s)$, with the definitions $\phi(s) = (sI - A)^{-1}$ and $X(s) = (\phi^{-1}(s) + HC)^{-1} = (sI - A + HC)^{-1}$.

$$\begin{aligned}
C(s) &= K(sI - A + BK + HC)^{-1}H \\
&= K(X^{-1}(s) + BK)^{-1}H, \text{ then use Lemma 2} \rightarrow \\
&= K(X(s) - X(s)B(I + KX(s)B)^{-1}KX(s))H \\
&= KX(s)H - KX(s)B(I + KX(s)B)^{-1}KX(s)H \\
&= (I - KX(s)B(I + KX(s)B)^{-1})KX(s)H, \text{ then use Lemma 3} \rightarrow \\
&= (I + KX(s)B)^{-1}KX(s)H \\
&= (\sqrt{\rho}I + \sqrt{\rho}KX(s)B)^{-1}\sqrt{\rho}KX(s)H.
\end{aligned}$$

Next we invoke the result from the LQR design, with $\rho \rightarrow 0$, to eliminate $\sqrt{\rho}K$:

$$\begin{aligned}
\lim_{\rho \rightarrow 0} C(s) &= (WCX(s)B)^{-1}WCX(s)H \\
&= (CX(s)B)^{-1}CX(s)H.
\end{aligned}$$

In the last expression, we used the assumption that W is square and invertible, both properties of orthonormal matrices. Now we look at the product $CX(s)$:

$$\begin{aligned}
CX(s) &= C(SI - A + HC)^{-1} \\
&= C(\phi^{-1}(s) + HC)^{-1}, \text{ then use Lemma 2} \rightarrow \\
&= C(\phi(s) - \phi(s)H(I + C\phi(s)H)^{-1}C\phi(s)) \\
&= (I - C\phi(s)H(I + C\phi(s)H)^{-1})C\phi(s), \text{ then use Lemma 3} \rightarrow \\
&= (I + C\phi(s)H)^{-1}C\phi(s).
\end{aligned}$$

This result, reintroduced into the limiting compensator, gives

$$\begin{aligned}
\lim_{\rho \rightarrow 0} C(s) &= ((I + C\phi(s)H)^{-1}C\phi(s)B)^{-1}(I + C\phi(s)H)^{-1}C\phi(s)H \\
&= (C\phi(s)B)^{-1}C\phi(s)H \\
&= P^{-1}(s)C\phi(s)H.
\end{aligned}$$

Finally it follows that $\lim_{\rho \rightarrow 0} P(s)C(s) = C\phi(s)H$, as desired.

21.4 Usage of the Loop Transfer Recovery

The idea of LTR is to "recover" a Kalman filter loop transfer function $L(s) = C\phi(s)H$, by using the limiting cheap-control LQR design, with $Q = C^TC$ and $R = \rho I$. The LQR design step is thus trivial.
Some specific techniques are useful.

- *Scale the plant outputs* (and references), so that one unit of error in one channel is as undesirable as one unit of error in another channel. For example, in depth and pitch control of a large submarine, one meter of depth error cannot be compared directly with one radian of pitch error.
- *Scale the plant inputs* in the same way. One Newton of propeller thrust cannot be compared with one radian of rudder angle.
- *Design for crossover frequency.* The bandwidth of the controller is roughly equal to the frequency at which the (recovered) loop transfer function crosses over $0dB$. Often, the bandwidth of is a more intuitive design parameter than is, for example, the high-frequency multiplicative weighting W_2. Quantitative uncertainty models are usually at the cost of a lengthy identification effort.
- *Integrators* should be part of the KF loop transfer function, if no steady-state error is to be allowed. Since the Kalman filter loop has only as many poles as the plant, the plant input channels must be augmented with the necessary additional poles (at the origin). Then, once the KF design is completed, and the compensator $C(s)$ is constructed, the integrators are moved from the plant over to the input side of the compensator. The tracking errors will accrue as desired.

21.5 Three Lemmas

Lemma 1: Matrix Inversion

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

Proof:

$$\begin{aligned}
(A+BCD)(A+BCD)^{-1} &= I \\
A(A+BCD)^{-1} &= I - BCD(A+BCD)^{-1} \\
(A+BCD)^{-1} &= A^{-1} - A^{-1}BCD(A+BCD)^{-1} \\
&= A^{-1} - A^{-1}BCD(I+A^{-1}BCD)^{-1}A^{-1} \\
&= A^{-1} - A^{-1}B(D^{-1}C^{-1}+A^{-1}B)^{-1}A^{-1} \\
&= A^{-1} - A^{-1}B(C^{-1}+DA^{-1}B)^{-1}DA^{-1}. \qquad \square
\end{aligned}$$

Lemma 2: Short Form of Lemma 1

$$(X^{-1}+BD)^{-1} = X - XB(I+DXB)^{-1}DX$$

Proof: substitute $A = X^{-1}$ and $C = I$ into Lemma 1.

Lemma 3

$$I - A(I + A)^{-1} = (I + A)^{-1}$$

Proof:

$$\begin{aligned} I - A(I + A)^{-1} &= (I + A)(I + A)^{-1} - A(I + A)^{-1} \\ &= (I + A - A)(I + A)^{-1} \\ &= (I + A)^{-1}. \qquad \square \end{aligned}$$

22 APPENDIX 1: MATH FACTS

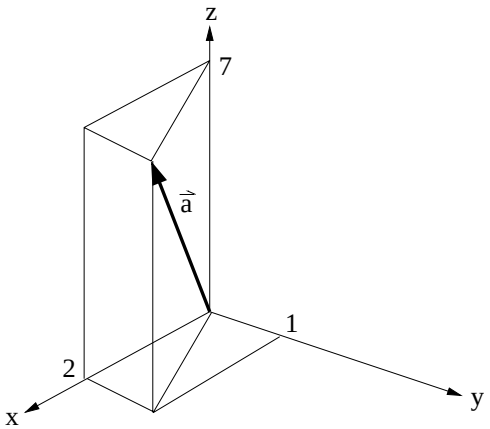
22.1 Vectors

22.1.1 Definition

A vector has a dual definition: It is a segment of a a line with direction, or it consists of its projection on a reference system $0xyz$, usually orthogonal and right handed. The first form is independent of any reference system, whereas the second (in terms of its components) depends directly on the coordinate system. Here we use the second notation, i.e., $\underline{\mathrm{x}}$ is meant as a column vector, whose components are found as projections of an (invariant) directed segment on a specific reference system.
We use the overhead arrow to denote a column vector, i.e., a *linear segment with a direction*. For example, in three-space, we write a vector in terms of its components with respect to a reference system as

$$\vec{a} = \left\{ \begin{array}{c} 2 \\ 1 \\ 7 \end{array} \right\}.$$

The elements of a vector have a graphical interpretation, which is particularly easy to see in two or three dimensions.



1. Vector addition:

$$\vec{a} + \vec{b} = \vec{c}$$

$$\left\{ \begin{array}{c} 2 \\ 1 \\ 7 \end{array} \right\} + \left\{ \begin{array}{c} 3 \\ 3 \\ 2 \end{array} \right\} = \left\{ \begin{array}{c} 5 \\ 4 \\ 9 \end{array} \right\}.$$

   Graphically, addition is stringing the vectors together head to tail.

2. Scalar multiplication:

$$-2 \times \left\{ \begin{array}{c} 2 \\ 1 \\ 7 \end{array} \right\} = \left\{ \begin{array}{c} -4 \\ -2 \\ -14 \end{array} \right\}.$$

22.1.2 Vector Magnitude

The total length of a vector of dimension m, its Euclidean norm, is given by

$$||\vec{x}|| = \sqrt{\sum_{i=1}^{m} x_i^2}.$$

This scalar is commonly used to normalize a vector to length one.

22.1.3 Vector Dot or Inner Product

The dot product of two vectors is a scalar equal to the sum of the products of the corresponding components:

$$\vec{x} \cdot \vec{y} = \vec{x}^T \vec{y} = \sum_{i=1}^{m} x_i y_i.$$

The dot product also satisfies

$$\vec{x} \cdot \vec{y} = ||\vec{x}||||\vec{y}|| \cos\theta,$$

where θ is the angle between the vectors.

22.1.4 Vector Cross Product

The cross product of two three-dimensional vectors $\vec{x}$ and $\vec{y}$ is another vector $\vec{z}$, $\vec{x} \times \vec{y} = \vec{z}$, whose

1. direction is normal to the plane formed by the other two vectors,
2. direction is given by the right-hand rule, rotating from $\vec{x}$ to $\vec{y}$,
3. magnitude is the area of the parallelogram formed by the two vectors – the cross product of two parallel vectors is zero – and
4. (signed) magnitude is equal to $||\vec{x}||||\vec{y}|| \sin\theta$, where θ is the angle between the two vectors, measured from $\vec{x}$ to $\vec{y}$.

In terms of their components,

$$\vec{x} \times \vec{y} = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = \begin{Bmatrix} (x_2 y_3 - x_3 y_2)\hat{i} \\ (x_3 y_1 - x_1 y_3)\hat{j} \\ (x_1 y_2 - x_2 y_1)\hat{k} \end{Bmatrix}.$$

22.2 Matrices

22.2.1 Definition

A matrix, or array, is equivalent to a set of column vectors of the same dimension, arranged side by side, say

$$A = [\vec{a} \ \ \vec{b}] = \begin{bmatrix} 2 & 3 \\ 1 & 3 \\ 7 & 2 \end{bmatrix}.$$

This matrix has three rows ($m = 3$) and two columns ($n = 2$); a vector is a special case of a matrix with one column. Matrices, like vectors, permit addition and scalar multiplication. We usually use an upper-case symbol to denote a matrix.

22.2.2 Multiplying a Vector by a Matrix

If A_{ij} denotes the element of matrix A in the i'th row and the j'th column, then the multiplication $\vec{c} = A\vec{v}$ is constructed as:

$$c_i = A_{i1}v_1 + A_{i2}v_2 + \cdots + A_{in}v_n = \sum_{j=1}^{n} A_{ij}v_j,$$

where n is the number of columns in A. $\vec{c}$ will have as many rows as A has rows (m). Note that this multiplication is defined only if $\vec{v}$ has as many rows as A has columns; they have

consistent *inner dimension* n. The product $\vec{v}A$ would be well-posed only if A had one row, and the proper number of columns. There is another important interpretation of this vector multiplication: Let the subscript : indicate all rows, so that each $A_{:j}$ is the j'th column vector. Then

$$\vec{c} = A\vec{v} = A_{:1}v_1 + A_{:2}v_2 + \cdots + A_{:n}v_n.$$

We are multiplying column vectors of A by the scalar elements of $\vec{v}$.

22.2.3 Multiplying a Matrix by a Matrix

The multiplication $C = AB$ is equivalent to a side-by-side arrangement of column vectors $C_{:j} = AB_{:j}$, so that

$$C = AB = [AB_{:1} \;\; AB_{:2} \;\; \cdots \;\; AB_{:k}],$$

where k is the number of columns in matrix B. The same inner dimension condition applies as noted above: the number of columns in A must equal the number of rows in B. Matrix multiplication is:

1. Associative. $(AB)C = A(BC)$.
2. Distributive. $A(B + C) = AB + AC$, $(B + C)A = BA + CA$.
3. NOT Commutative. $AB \neq BA$, except in special cases.

22.2.4 Common Matrices

Identity. The identity matrix is usually denoted I, and comprises a square matrix with ones on the diagonal, and zeros elsewhere, e.g.,

$$I_{3\times 3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The identity always satisfies $AI_{n\times n} = I_{m\times m}A = A$.

Diagonal Matrices. A diagonal matrix is square, and has all zeros off the diagonal. For instance, the following is a diagonal matrix:

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 3 \end{bmatrix}.$$

The product of a diagonal matrix with another diagonal matrix is diagonal, and in this case the operation is commutative.

22.2.5 Transpose

The transpose of a vector or matrix, indicated by a T superscript results from simply swapping the row-column indices of each entry; it is equivalent to "flipping" the vector or matrix around the diagonal line. For example,

$$\vec{a} = \begin{Bmatrix} 1 \\ 2 \\ 3 \end{Bmatrix} \longrightarrow \vec{a}^{\,T} = \{1 \;\; 2 \;\; 3\}$$

$$A = \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 8 & 9 \end{bmatrix} \longrightarrow A^T = \begin{bmatrix} 1 & 4 & 8 \\ 2 & 5 & 9 \end{bmatrix}.$$

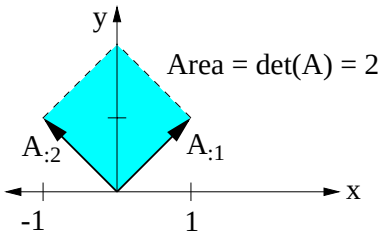
A very useful property of the transpose is

$$(AB)^T = B^T A^T.$$

22.2.6 Determinant

The determinant of a square matrix A is a scalar equal to *the volume* of the parallelepiped enclosed by the constituent vectors. The two-dimensional case is particularly easy to remember, and illustrates the principle of volume:

$$\begin{aligned} det(A) &= A_{11}A_{22} - A_{21}A_{12} \\ det\left(\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\right) &= 1 + 1 = 2. \end{aligned}$$



In higher dimensions, the determinant is more complicated to compute. The general formula allows one to pick a row k, perhaps the one containing the most zeros, and apply

$$det(A) = \sum_{j=1}^{j=n} A_{kj}(-1)^{k+j}\Delta_{kj},$$

where Δ_{kj} is the determinant of the sub-matrix formed by neglecting the k'th row and the j'th column. The formula is symmetric, in the sense that one could also target the k'th column:

$$det(A) = \sum_{j=1}^{j=n} A_{jk}(-1)^{k+j}\Delta_{jk}.$$

If the determinant of a matrix is zero, then the matrix is said to be singular – there is no volume, and this results from the fact that the constituent vectors do not span the matrix dimension. For instance, in two dimensions, a singular matrix has the vectors colinear; in three dimensions, a singular matrix has all its vectors lying in a (two-dimensional) plane. Note also that $det(A) = det(A^T)$. If $det(A) \neq 0$, then the matrix is said to be nonsingular.

22.2.7 Inverse

The inverse of a square matrix A, denoted A^{-1}, satisfies $AA^{-1} = A^{-1}A = I$. Its computation requires the determinant above, and the following definition of the $n \times n$ *adjoint* matrix:

$$adj(A) = \begin{bmatrix} (-1)^{1+1}\Delta_{11} & \cdots & (-1)^{1+n}\Delta_{1n} \\ \cdots & \cdots & \cdots \\ (-1)^{n+1}\Delta_{n1} & \cdots & (-1)^{n+n}\Delta_{nn}. \end{bmatrix}^T.$$

Once this computation is made, the inverse follows from

$$A^{-1} = \frac{adj(A)}{det(A)}.$$

If A is singular, i.e., $det(A) = 0$, then the inverse does not exist. The inverse finds common application in solving systems of linear equations such as

$$A\vec{x} = \vec{b} \longrightarrow \vec{x} = A^{-1}\vec{b}.$$

22.2.8 Trace

The trace of a square matrix is the sum of the diagonals:

$$tr(A) = \sum_{i=1}^{n} A_{ii}.$$

22.2.9 Eigenvalues and Eigenvectors

A typical eigenvalue problem is stated as

$$A\vec{x} = \lambda\vec{x},$$

where A is an $n \times n$ matrix, $\vec{x}$ is a column vector with n elements, and λ is a scalar. We ask for what nonzero vectors $\vec{x}$ (right eigenvectors), and scalars λ (eigenvalues) will the equation be satisfied. Since the above is equivalent to $(A - \lambda I)\vec{x} = \vec{0}$, it is clear that $det(A - \lambda I) = 0$. This observation leads to the solutions for λ; here is an example for the two-dimensional case:

$$\begin{aligned} A &= \begin{bmatrix} 4 & -5 \\ 2 & -3 \end{bmatrix} \longrightarrow \\ A - \lambda I &= \begin{bmatrix} 4-\lambda & -5 \\ 2 & -3-\lambda \end{bmatrix} \longrightarrow \\ det(A - \lambda I) &= (4-\lambda)(-3-\lambda) + 10 \\ &= \lambda^2 - \lambda - 2 \\ &= (\lambda + 1)(\lambda - 2). \end{aligned}$$

Thus, A has two eigenvalues, $\lambda_1 = -1$ and $\lambda_2 = 2$. Each is associated with a *right eigenvector* $\vec{x}$. In this example,

$$\begin{aligned} (A - \lambda_1 I)\vec{x}_1 &= \vec{0} \longrightarrow \\ \begin{bmatrix} 5 & -5 \\ 2 & -2 \end{bmatrix} \vec{x}_1 &= \vec{0} \longrightarrow \\ \vec{x}_1 &= \left\{\sqrt{2}/2, \ \ \sqrt{2}/2\right\}^T \end{aligned}$$

$$\begin{aligned} (A - \lambda_2 I)\vec{x}_2 &= \vec{0} \longrightarrow \\ \begin{bmatrix} 2 & -5 \\ 2 & -5 \end{bmatrix} \vec{x}_2 &= \vec{0} \longrightarrow \\ \vec{x}_2 &= \left\{5\sqrt{29}/29, \ \ 2\sqrt{29}/29\right\}^T. \end{aligned}$$

Eigenvectors are defined only within an arbitrary constant, i.e., if $\vec{x}$ is an eigenvector then $c\vec{x}$ is also an eigenvector for any $c \neq 0$. They are often normalized to have unity magnitude, and positive first element (as above). The condition that $rank(A - \lambda_i I) = rank(A) - 1$ indicates that there is only one eigenvector for the eigenvalue λ_i; more precisely, a unique direction for the eigenvector, since the magnitude can be arbitrary. If the left-hand side rank is less than this, then there are multiple eigenvectors that go with λ_i.

The above discussion relates only the right eigenvectors, generated from the equation $A\vec{x} = \lambda\vec{x}$. Left eigenvectors, defined as $\vec{y}^T A = \lambda\vec{y}^T$, are also useful for many problems, and can be defined simply as the right eigenvectors of A^T. A and A^T share the same eigenvalues λ, since they share the same determinant. Example:

$$(A^T - \lambda_1 I)\vec{y}_1 = \vec{0} \longrightarrow$$

$$\begin{aligned}
\begin{bmatrix} 5 & 2 \\ -5 & -2 \end{bmatrix} \vec{y}_1 &= \vec{0} \longrightarrow \\
\vec{y}_1 &= \left\{ 2\sqrt{29}/29, \ -5\sqrt{29}/29 \right\}^T \\
(A^T - \lambda_2 I)\vec{y}_2 &= \vec{0} \longrightarrow \\
\begin{bmatrix} 2 & 2 \\ -5 & -5 \end{bmatrix} \vec{y}_2 &= \vec{0} \longrightarrow \\
\vec{y}_2 &= \left\{ \sqrt{2}/2, \ -\sqrt{2}/2 \right\}^T .
\end{aligned}$$

22.2.10 Modal Decomposition

For simplicity, we consider matrices that have unique eigenvectors for each eigenvalue. The right and left eigenvectors corresponding to a particular eigenvalue λ can be defined to have unity dot product, that is $\vec{x}_i^T \vec{y}_i = 1$, with the normalization noted above. The dot products of a left eigenvector with the right eigenvectors corresponding to *different eigenvalues* are zero. Thus, if the set of right and left eigenvectors, V and W, respectively, is

$$\begin{aligned}
V &= [\vec{x}_1 \cdots \vec{x}_n] \text{, and} \\
W &= [\vec{y}_1 \cdots \vec{y}_n] ,
\end{aligned}$$

then we have

$$\begin{aligned}
W^T V &= I \text{, or} \\
W^T &= V^{-1}.
\end{aligned}$$

Next, construct a diagonal matrix containing the eigenvalues:

$$\Lambda = \begin{bmatrix} \lambda_1 & & 0 \\ & \cdot & \\ 0 & & \lambda_n \end{bmatrix};$$

it follows that

$$\begin{aligned}
AV &= V\Lambda \longrightarrow \\
A &= V\Lambda W^T \\
&= \sum_{i=1}^{n} \lambda_i \vec{v}_i \vec{w}_i^T .
\end{aligned}$$

Hence A can be written as a sum of modal components.[2]

[2]By carrying out successive multiplications, it can be shown that A^k has its eigenvalues at λ_i^k, and keeps the same eigenvectors as A.

22.2.11 Singular Value

Let G be an $m \times n$ real or complex matrix. The singular value decomposition (SVD) computes three matrices satisfying

$$G = U\Sigma V^*,$$

where U is $m \times m$, Σ is $m \times n$, and V is $n \times n$. The star notation indicates a complex-conjugate transpose (the Hermitian of the matrix). The matrix Σ has the form

$$\Sigma = \begin{bmatrix} \Sigma_1 & 0 \\ 0 & 0 \end{bmatrix}, \text{ where } \Sigma_1 = \begin{bmatrix} \sigma_1 & 0 & 0 \\ 0 & \cdot & 0 \\ 0 & 0 & \sigma_p \end{bmatrix},$$

and $p = min(m, n)$. Each nonzero entry on the diagonal of matrix Σ_1 is a real, positive singular value, ordered such that $\sigma_1 > \sigma_2 > \cdots \sigma_p$. Each σ_i^2 is an eigenvalue of $G^H G$ (or GG^H). The notation is common that $\sigma_1 = \overline{\sigma}$, the maximum singular value, and $\sigma_p = \underline{\sigma}$, the minimum singular value. The auxiliary matrices U and V are unitary, i.e., they satisfy $X^* = X^{-1}$. They are defined explicitly: U is the matrix of right eigenvectors of GG^H, and V is the matrix of right eigenvectors of $G^H G$. Like eigenvalues, the singular values of G are related to projections. σ_i represents the Euclidean size of the matrix G along the i'th singular vector:

$$\begin{aligned} \overline{\sigma} &= max_{||x||=1}||Gx|| \\ \underline{\sigma} &= min_{||x||=1}||Gx||. \end{aligned}$$

Other properties of the singular value include:

- $\overline{\sigma}(AB) \leq \overline{\sigma}(A)\overline{\sigma}(B)$.
- $\overline{\sigma}(A) = \sqrt{\lambda_{max}(A^*A)}$.
- $\underline{\sigma}(A) = \sqrt{\lambda_{min}(A^*A)}$.
- $\underline{\sigma}(A) = 1/\overline{\sigma}(A^{-1})$.
- $\overline{\sigma}(A) = 1/\underline{\sigma}(A^{-1})$.

22.3 Laplace Transform

22.3.1 Definition

The Laplace transform projects time-domain signals into a complex frequency-domain equivalent. The signal $y(t)$ has transform $Y(s)$ defined as follows:

$$Y(s) = L(y(t)) = \int_0^\infty y(\tau)e^{-s\tau}d\tau,$$

where s is a complex variable, properly constrained within a region so that the integral converges. $Y(s)$ is a complex function as a result. Note that the Laplace transform is linear, and so it is distributive: $L(x(t) + y(t)) = L(x(t)) + L(y(t))$. The following table gives a list of some useful transform pairs and other properties, for reference.

The last two properties are of special importance: for control system design, the differentiation of a signal is equivalent to multiplication of its Laplace transform by s; integration of a signal is equivalent to division by s. The other terms that arise will cancel if $y(0) = 0$, or if $y(0)$ is finite.

22.3.2 Convergence

We note first that the value of s affects the convergence of the integral. For instance, if $y(t) = e^t$, then the integral converges only for $Re(s) > 1$, since the integrand is e^{1-s} in this case. Although the integral converges within a well-defined region in the complex plane, the function $Y(s)$ is defined for all s through analytic continuation. This result from complex analysis holds that if two complex functions are equal on some arc (or line) in the complex plane, then they are equivalent everywhere. It should be noted however, that the Laplace transform is defined only within the region of convergence.

22.3.3 Convolution Theorem

One of the main points of the Laplace transform is the ease of dealing with dynamic systems. As with the Fourier transform, the convolution of two signals in the time domain corresponds with the multiplication of signals in the frequency domain. Consider a system whose impulse response is $g(t)$, being driven by an input signal $x(t)$; the output is $y(t) = g(t) * x(t)$. The *Convolution Theorem* is

$$y(t) = \int_0^t g(t-\tau)x(\tau)d\tau \Longleftrightarrow Y(s) = G(s)X(s).$$

Here's the proof given by Siebert:

$$\begin{aligned}
Y(s) &= \int_0^\infty y(t)e^{-st}dt \\
&= \int_0^\infty \left[\int_0^t g(t-\tau)\ x(\tau)\ d\tau\right] e^{-st}\ dt \\
&= \int_0^\infty \left[\int_0^\infty g(t-\tau)\ h(t-\tau)\ x(\tau)\ d\tau\right] e^{-st}dt \\
&= \int_0^\infty x(\tau)\left[\int_0^\infty g(t-\tau)\ h(t-\tau)\ e^{-st}\ dt\right] d\tau
\end{aligned}$$

$$
\begin{aligned}
y(t) &\longleftrightarrow Y(s) \\
\text{(Impulse)} \quad \delta(t) &\longleftrightarrow 1 \\
\text{(Unit Step)} \quad 1(t) &\longleftrightarrow \frac{1}{s} \\
\text{(Unit Ramp)} \quad t &\longleftrightarrow \frac{1}{s^2} \\
e^{-\alpha t} &\longleftrightarrow \frac{1}{s+\alpha} \\
\sin \omega t &\longleftrightarrow \frac{\omega}{s^2+\omega^2} \\
\cos \omega t &\longleftrightarrow \frac{s}{s^2+\omega^2} \\
e^{-\alpha t} \sin \omega t &\longleftrightarrow \frac{\omega}{(s+\alpha)^2+\omega^2} \\
e^{-\alpha t} \cos \omega t &\longleftrightarrow \frac{s+\alpha}{(s+\alpha)^2+\omega^2} \\
\frac{1}{b-a}(e^{-at}-e^{-bt}) &\longleftrightarrow \frac{1}{(s+a)(s+b)} \\
\frac{1}{ab}\left[1+\frac{1}{a-b}(be^{-at}-ae^{-bt})\right] &\longleftrightarrow \frac{1}{s(s+a)(s+b)} \\
\frac{\omega_n}{\sqrt{1-\zeta^2}}\, e^{-\zeta\omega_n t} \sin \omega_n\sqrt{1-\zeta^2}t &\longleftrightarrow \frac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2} \\
1-\frac{1}{\sqrt{1-\zeta^2}}\, e^{-\zeta\omega_n t} \sin\left(\omega_n\sqrt{1-\zeta^2}t+\phi\right) &\longleftrightarrow \frac{\omega_n^2}{s(s^2+2\zeta\omega_n s+\omega_n^2)} \\
\left(\phi=\tan^{-1}\frac{\sqrt{1-\zeta^2}}{\zeta}\right) & \\
\text{(Pure Delay)} \quad y(t-\tau)1(t-\tau) &\longleftrightarrow Y(s)e^{-s\tau} \\
\text{(Time Derivative)} \quad \frac{dy(t)}{dt} &\longleftrightarrow sY(s)-y(0) \\
\text{(Time Integral)} \quad \int_0^t y(\tau)d\tau &\longleftrightarrow \frac{Y(s)}{s}+\frac{\int_{0-}^{0+} y(t)dt}{s}
\end{aligned}
$$

$$\begin{aligned} &= \int_0^\infty x(\tau)\, G(s) e^{-s\tau}\, d\tau \\ &= G(s)X(s), \end{aligned}$$

where $h(t)$ is the unit step function. When $g(t)$ is the impulse response of a dynamic system, then $y(t)$ represents the output of this system when it is driven by the external signal $x(t)$.

22.3.4 Solution of Differential Equations by Laplace Transform

The Convolution Theorem allows one to solve (linear time-invariant) differential equations in the following way:

1. Transform the system impulse response $g(t)$ into $G(s)$, and the input signal $x(t)$ into $X(s)$, using the transform pairs.
2. Perform the multiplication in the Laplace domain to find $Y(s)$.
3. Ignoring the effects of pure time delays, break $Y(s)$ into partial fractions with no powers of s greater than 2 in the denominator.
4. Generate the time-domain response from the simple transform pairs. Apply time delay as necessary.

Specific examples of this procedure are given in a later section on transfer functions.

22.4 Background for the Mapping Theorem

The mapping theorem uses concepts from complex analysis, specifically Cauchy's Theorem and the Residue Theorem. References for this section include Ogata and Hildebrand.
First, consider a function of the complex variable $s = u + iv$: $f(s)$. We say $f(s)$ is analytic in a region S if it has finite derivative and takes only one value at each point s in S. Therefore discontinuous or multi-valued functions, e.g., $\sqrt{s}$, are not analytic functions. Polynomials in s are analytic, as are many functions that can be written as a Taylor or Laurent series. An especially important class for control system design is the rational function, a polynomial in s divided by another polynomial in s. Rational functions are consequently zero for certain values of s, the roots of the numerator, and undefined for other values of s, the roots of the numerator, also called the poles.
The integral of interest here is

$$\int f(s)ds$$

taken on a path in the s-plane. A closed path in the complex s-plane leads to a closed path in the $f(s)$ plane, but more than one point in the s plane can map to a single $f(s)$-plane point, so the number of complete loops may not be the same.
The usual rules of integration apply in complex analysis, so that, insofar as the antiderivative of $f(s)$, denoted $F(s)$ exists, and $f(s)$ is analytic on the path, we have

$$\int_{s_1}^{s_2} f(s)ds = F(s_2) - F(s_1).$$

It appears that this integral is zero for a closed path, since $s_1 = s_2$. Indeed, Cauchy's theorem states that it is, provided that $f(s)$ is analytic on the path, and *everywhere within the region enclosed*. This latter condition results from the following observation. Consider the function $f(s) = \alpha s^n$; on a circular path of radius r, we have $s = re^{i\theta}$, and thus

$$\begin{aligned} \int f(s)ds &= i\alpha r^{n+1} \int_0^{2\pi} e^{i(n+1)\theta} d\theta \\ &= i\alpha r^{n+1} \int_0^{2\pi} [\cos(n+1)\theta + i\sin(n+1)\theta] d\theta. \end{aligned}$$

The second integral is clearly zero for all n, whereas the first is zero except in the case of $n = -1$, for which we obtain

$$\int f(s)ds = i\alpha 2\pi.$$

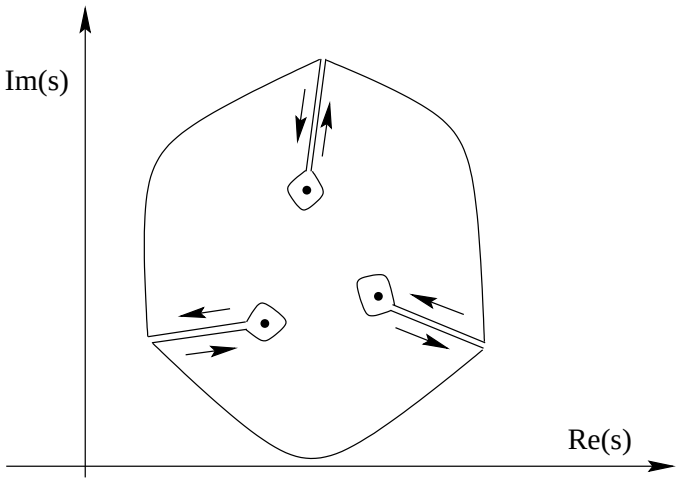
This result does not depend on r, and so applies to a vanishingly small circle around the point $s = 0$. It can be shown also that the result holds for any closed path around the *simple pole at $s = 0$*, which characterizes the function. The residue theorem is an extension to an arbitrary number of simple poles which are enclosed by a path:

$$\int f(s)ds = i2\pi \sum \alpha_i,$$

The constants α_i are the residues associated with the poles, i.e.,

$$f(s) = \frac{\alpha_1}{s - p_1} + \frac{\alpha_2}{s - p_2} + \ldots$$

We show in another section how any strictly proper rational function (that is, the polynomial order in the numerator is less than that of the denominator) in s can be put into this form. The connection between Cauchy's theorem and the residue theorem is indicated in the figure below. Here, the integral inside the closed path is zero because it excludes the three simple poles. Without the cuts however, the integral over the outermost curve would be equal to the summation of residues for the poles within (multiplied by $i2\pi$). Note that it is only the terms with $(s - a_i)^{-1}$, i.e., simple poles at a_i, that generate nonzero components.



Looking toward the mapping theorem now, consider the function

$$f(s) = \frac{(s-z_1)^{m_1}(s-z_2)^{m_2}\dots}{(s-p_1)^{n_1}(s-p_2)^{n_2}\dots}.$$

Working through some algebra, it is straightforward that

$$f'(s)/f(s) = \frac{m_1}{s-z_1} + \frac{m_2}{s-z_2} + \dots - \frac{n_1}{s-p_1} - \frac{n_2}{s-p_2}\dots$$

resulting in

$$\int f'(s)/f(s)ds = i2\pi(Z-P),$$

where $Z = m_1 + m_2 + \dots$ is the number of zeros in f, and $P = n_1 + n_2 + \dots$ is the number of zeros. The mapping theorem comes from now looking in detail at the integral above:

$$\begin{aligned} f(s) &= |f(s)|e^{i\theta(s)} \\ f'(s) &= \frac{d|f(s)|}{ds}e^{i\theta(s)} + i|f(s)|e^{i\theta(s)}\frac{d\theta(s)}{ds} \\ f'(s)/f(s) &= \frac{1}{|f(s)|}\frac{d|f(s)|}{ds} + i\frac{d\theta(s)}{ds} \\ &= \frac{d\log|f(s)|}{ds} + i\frac{d\theta(s)}{ds}, \end{aligned}$$

Considering the integral of $f'(s)/f(s)$ over a closed contour in the s-plane, we take advantage of the fact that exact differentials $d\log|f(s)|$ and $d\theta(s)$ have been found. Both terms pertain to the $f(s)$ plane, not the $f'(s)/f(s)$ plane. The first integral is zero,

$$\int_{s_1}^{s_2} d\log|f| = 0,$$

since $log|f(s)|$ has to be the same at the start and end of the closed contour.
Taking $\theta(s)$ counterclockwise through angle $n2\pi$ results in the second term

$$\int_0^{n2\pi} id\theta = in2\pi.$$

As noted above, a single circuit in the s-plane may or may not map to a single circuit in the $f(s)$-plane, so n depends directly on the function $f(s)$, and is not necessarily one. Assembling the results, we have

$$i2\pi(Z-P) = in2\pi \longrightarrow Z - P = n,$$

which is the mapping theorem. In words, it states that the number of zeros minus the number of poles enclosed in a contour in the s-plane is equal to the number of counterclockwise encirclements of the origin, in the $f(s)$ plane.

23 APPENDIX 2: ADDED MASS VIA LAGRANGIAN DYNAMICS

The development of rigid body inertial dynamics presented in a previous section depends on the rates of change of vectors expressed in a moving frame, specifically that of the vehicle. An alternative approach is to use the *lagrangian*, wherein the dynamic behavior follows directly from consideration of the kinetic co-energy of the vehicle; the end result is exactly the same. Since the body dynamics were already developed, we here develop the lagrangian technique, using the analogous example of added mass terms. Among other effects, the equations elicit the origins of the Munk moment.

23.1 Kinetic Energy of the Fluid

The added mass matrix for a body in six degrees of freedom is expressed as the matrix M_a, whose negative is equal to:

$$-M_a = \begin{bmatrix} X_{\dot{u}} & X_{\dot{v}} & X_{\dot{w}} & X_{\dot{p}} & X_{\dot{q}} & X_{\dot{r}} \\ Y_{\dot{u}} & Y_{\dot{v}} & Y_{\dot{w}} & Y_{\dot{p}} & Y_{\dot{q}} & Y_{\dot{r}} \\ Z_{\dot{u}} & Z_{\dot{v}} & Z_{\dot{w}} & Z_{\dot{p}} & Z_{\dot{q}} & Z_{\dot{r}} \\ K_{\dot{u}} & K_{\dot{v}} & K_{\dot{w}} & K_{\dot{p}} & K_{\dot{q}} & K_{\dot{r}} \\ M_{\dot{u}} & M_{\dot{v}} & M_{\dot{w}} & M_{\dot{p}} & M_{\dot{q}} & M_{\dot{r}} \\ N_{\dot{u}} & N_{\dot{v}} & N_{\dot{w}} & N_{\dot{p}} & N_{\dot{q}} & N_{\dot{r}} \end{bmatrix}, \tag{260}$$

where (X, Y, Z) denotes the force, (K, M, N) the moment, (u, v, w) denotes the velocity and (p, q, r) the angular velocity. The sense of M_a is that the fluid forces due to added mass are given by

$$\begin{Bmatrix} X_{am} \\ Y_{am} \\ Z_{am} \\ K_{am} \\ M_{am} \\ N_{am} \end{Bmatrix} = -M_a \frac{d}{dt} \begin{Bmatrix} u \\ v \\ w \\ p \\ q \\ r \end{Bmatrix}. \tag{261}$$

The added mass matrix M_a is completely analagous to the actual mass matrix of the vehicle,

$$M = \begin{bmatrix} m & 0 & 0 & 0 & z_G & -y_G \\ 0 & m & 0 & -z_G & 0 & x_G \\ 0 & 0 & m & y_G & -x_G & 0 \\ 0 & -z_G & y_G & I_{xx} & I_{xy} & I_{xz} \\ z_G & 0 & -x_G & I_{xy} & I_{yy} & I_{yz} \\ -y_G & x_G & 0 & I_{xz} & I_{yz} & I_{zz} \end{bmatrix}, \tag{262}$$

where $[x_G, y_G, z_G]$ are the (vessel frame) coordinates of the center of gravity. The mass matrix is symmetric, nonsingular, and positive definite. These properties are also true for the added mass matrix M_a, although symmetry fails when there is a constant forward speed.
The kinetic co-energy of the fluid E_k is found as:

$$E_k = -\frac{1}{2}\, q^{\,T} M_a\, q \tag{263}$$

where $q^{\,T} = (u, v, w, p, q, r)$. We expand to find in the non-symmetric case:

$$\begin{aligned} -2E_k \;=\; & X_{\dot{u}}u^2 + X_{\dot{v}}uv + X_{\dot{w}}uw + X_{\dot{p}}up + X_{\dot{q}}uq + X_{\dot{r}}ur + \\ & Y_{\dot{u}}uv + Y_{\dot{v}}v^2 + Y_{\dot{w}}vw + Y_{\dot{p}}vp + Y_{\dot{q}}vq + Y_{\dot{r}}vr + \\ & Z_{\dot{u}}uw + Z_{\dot{v}}vw + Z_{\dot{w}}w^2 + Z_{\dot{p}}wp + Z_{\dot{q}}wq + Z_{\dot{r}}wr + \\ & K_{\dot{u}}up + K_{\dot{v}}vp + K_{\dot{w}}wp + K_{\dot{p}}p^2 + K_{\dot{q}}pq + K_{\dot{r}}pr + \\ & M_{\dot{u}}uq + M_{\dot{v}}vq + M_{\dot{w}}wq + M_{\dot{p}}pq + M_{\dot{q}}q^2 + M_{\dot{r}}qr + \\ & N_{\dot{u}}ur + N_{\dot{v}}vr + N_{\dot{w}}wr + N_{\dot{p}}pr + N_{\dot{q}}qr + N_{\dot{r}}r^2. \end{aligned} \tag{264}$$

For the purposes of this discussion, we will assume from here on that M_a is symmetric, that is $M_a = M_a^T$, or, for example, $Y_{\dot{u}} = X_{\dot{v}}$. In general, this implies that the i'th force due to the j'th acceleration is equal to the j'th force due to the i'th acceleration, for i and j = [1,2,3,4,5,6]. Refering to the above equation, these terms occur as pairs, so the asymmetric case could be reconstructed easily from what follows. By restricting the added mass to be symmetric, then, we find:

$$\begin{aligned} -2E_k &= X_{\dot{u}}u^2 + Y_{\dot{v}}v^2 + Z_{\dot{w}}w^2 + K_{\dot{p}}p^2 + M_{\dot{q}}q^2 + N_{\dot{r}}r^2 + \\ &\quad 2X_{\dot{v}}uv + 2X_{\dot{w}}uw + 2X_{\dot{p}}up + 2X_{\dot{q}}uq + 2X_{\dot{r}}ur + \\ &\quad 2Y_{\dot{w}}vw + 2Y_{\dot{p}}vp + 2Y_{\dot{q}}vq + 2Y_{\dot{r}}vr + \\ &\quad 2Z_{\dot{p}}wp + 2Z_{\dot{q}}wq + 2Z_{\dot{r}}wr + \\ &\quad 2K_{\dot{q}}pq + 2K_{\dot{r}}pr + \\ &\quad 2M_{\dot{r}}qr. \end{aligned} \tag{265}$$

23.2 Kirchhoff's Relations

To derive the fluid inertia terms in the body-referenced equations of motion, we use Kirchhoff's relations with the co-energy E_k; see the derivation below, or Milne-Thomson (1960). These relations state that if $\vec{v} = [u, v, w]$ denotes the velocity vector and $\vec{\omega} = [p, q, r]$ the angular velocity vector, then the inertia terms expressed in axes affixed to a moving vehicle are

$$\vec{F} = -\frac{\partial}{\partial t}\left(\frac{\partial E_k}{\partial \vec{v}}\right) - \vec{\omega} \times \frac{\partial E_k}{\partial \vec{v}}, \tag{266}$$

$$\vec{Q} = -\frac{\partial}{\partial t}\left(\frac{\partial E_k}{\partial \vec{\omega}}\right) - \vec{\omega} \times \frac{\partial E_k}{\partial \vec{\omega}} - \vec{v} \times \frac{\partial E_k}{\partial \vec{v}}. \tag{267}$$

where $\vec{F} = [X, Y, Z]$ express the force vector, $\vec{Q} = [K, M, N]$ the moment vector, and $\times$ denotes the cross product.

23.3 Fluid Inertia Terms

Applying Kirchhoff's relations to the expression for the kinetic co-energy with a symmetric added mass matrix, we derive the following terms containing the fluid inertia forces:

$$\begin{aligned} X &= +X_{\dot{u}}\dot{u} + X_{\dot{v}}(\dot{v} - ur) + X_{\dot{w}}(\dot{w} + uq) + X_{\dot{p}}\dot{p} + X_{\dot{q}}\dot{q} + X_{\dot{r}}\dot{r} \\ &\quad -Y_{\dot{v}}vr + Y_{\dot{w}}(vq - wr) - Y_{\dot{p}}pr - Y_{\dot{q}}qr - Y_{\dot{r}}r^2 \\ &\quad +Z_{\dot{w}}wq + Z_{\dot{p}}pq + Z_{\dot{q}}q^2 + Z_{\dot{r}}qr. \end{aligned} \tag{268}$$

The Y and Z forces can be obtained, through rotational symmetry, in the form:

$$\begin{aligned} Y &= +X_{\dot{v}}(\dot{u} + ur) - X_{\dot{w}})up \\ &\quad +Y_{\dot{v}}(\dot{v} + vr) + Y_{\dot{w}}(\dot{w} - vp + wr) + Y_{\dot{p}}(\dot{p} + pr) + Y_{\dot{q}}(\dot{q} + qr) + Y_{\dot{r}}(\dot{r} + r^2) \\ &\quad -Z_{\dot{w}}wp - Z_{\dot{p}}p^2 - Z_{\dot{q}}pq - Z_{\dot{r}}pr \end{aligned} \tag{269}$$

$$\begin{aligned} Z &= -X_{\dot{u}}uq + X_{\dot{v}}(up - vq) + X_{\dot{w}}(\dot{u} - wq) - X_{\dot{p}}pq - X_{\dot{q}}q^2 - X_{\dot{r}}qr \\ &\quad +Y_{\dot{v}}vp + Y_{\dot{w}}(\dot{v} + wp) + Y_{\dot{p}}p^2 + Y_{\dot{q}}pq + Y_{\dot{r}}pr \\ &\quad +Z_{\dot{w}}\dot{w} + Z_{\dot{p}}\dot{p} + Z_{\dot{q}}\dot{q} + Z_{\dot{r}}\dot{r} \end{aligned} \tag{270}$$

The apparent imbalance of coefficients comes from symmetry, which allows us to use only the 21 upper-right elements of the added mass matrix in Equation 260, e.g., $M_{\dot{v}} = Y_{\dot{q}}$. The moments K, M, and N are obtained in a similar manner as:

$$\begin{aligned} K &= -X_{\dot{v}}wu + X_{\dot{w}}uv + X_{\dot{r}}uq + X_{\dot{p}}\dot{u} - X_{\dot{q}}ur \\ &\quad -Y_{\dot{v}}vw + Y_{\dot{w}}(v^2 - w^2) + Y_{\dot{p}}(\dot{v} - wp) - Y_{\dot{q}}(vr + wq) + Y_{\dot{r}}(vq - wr) \\ &\quad +Z_{\dot{w}}vw + Z_{\dot{p}}(\dot{w} + vp) + Z_{\dot{q}}(vq - wr) + Z_{\dot{r}}(wq + vr) \\ &\quad +K_{\dot{p}}\dot{p} + K_{\dot{q}}(\dot{q} - rp) + K_{\dot{r}}(\dot{r} + pq) \\ &\quad +M_{\dot{r}}(q^2 - r^2) - M_{\dot{q}}qr \\ &\quad +N_{\dot{r}}qr \end{aligned} \tag{271}$$

$$\begin{aligned} M &= +X_{\dot{u}}uw + X_{\dot{v}}vw + X_{\dot{w}}(w^2 - u^2) + X_{\dot{p}}(ur + wp) + X_{\dot{q}}(\dot{u} + wq) + X_{\dot{r}}(wr - up) \\ &\quad -Y_{\dot{w}}uv + Y_{\dot{p}}vr + Y_{\dot{q}}\dot{v} - Y_{\dot{r}}vp \\ &\quad -Z_{\dot{w}}uw + Z_{\dot{p}}(wr - up) + Z_{\dot{q}}(\dot{w} - uq) - Z_{\dot{r}}(ur + wp) \\ &\quad +K_{\dot{p}}pr + K_{\dot{q}}(\dot{p} + qr) + K_{\dot{r}}(r^2 - p^2) \\ &\quad +M_{\dot{q}}\dot{q} - M_{\dot{r}}pq + M_{\dot{r}}\dot{r} \\ &\quad -N_{\dot{r}}pr \end{aligned} \tag{272}$$

$$\begin{aligned} N &= -X_{\dot{u}}uv + X_{\dot{v}}(u^2 - v^2) - X_{\dot{w}}vw - X_{\dot{p}}(uq + vp) + X_{\dot{q}}(up - vq) + X_{\dot{r}}(\dot{u} - vr) \\ &\quad +Y_{\dot{v}}uv + Y_{\dot{w}}uw + Y_{\dot{p}}(up - vq) + Y_{\dot{q}}(uq + vp) + Y_{\dot{r}}(\dot{v} + ur) \\ &\quad -Z_{\dot{p}}wq + Z_{\dot{q}}wp + Z_{\dot{r}}\dot{w} \\ &\quad -K_{\dot{p}}pq + K_{\dot{q}}(p^2 - q^2) + K_{\dot{r}}(\dot{p} - qr) \\ &\quad +M_{\dot{q}}pq + M_{\dot{r}}(\dot{q} + pr) \\ &\quad +N_{\dot{r}}\dot{r} \end{aligned} \tag{273}$$

23.4 Derivation of Kirchhoff's Relations

We can derive Kirchhoff's relation for a lagrangian $L(\vec{v}, \vec{\omega}, t)$, involving the velocity $\vec{v}$ and angular velocity $\vec{\omega}$, whose components will be expressed in a local coordinate system rotating with the angular velocity $\vec{\omega}$, i.e. in a reference system fixed on the body. The principle to satisfy is that of least action, i.e. to minimize the integral I (Crandall *et al.*, 1968):

$$I = \int_{t_1}^{t_2} L(\vec{v}, \vec{\omega}, t) dt \tag{274}$$

At the minimum value of I - the admissible condition - the variation of I, δI, and hence of L with the velocity $\vec{v}$ and angular rate $\vec{\omega}$ vanishes.
Our condition $\delta I = 0$, as written, involves only the lagrangian, which in the more general case is the kinetic co-energy minus the potential energy of the system. Since we are considering the motion of a body in an unbounded, homogeneous fluid, there is no potential energy, so the lagrangian is exactly the kinetic energy:

$$L = E_k. \tag{275}$$

Hamilton's principle in its general form also accounts for applied forces and moments $\vec{\vec{\Xi}}$. They are defined to align with the generalized, infinitesimal linear and angular displacements $\delta\vec{\eta}$ and $\delta\vec{\phi}$, leading to

$$\delta I = \int_{t_1}^{t_2} \left[\delta L(\vec{v}, \vec{\omega}, t) + \left\langle \vec{\vec{\Xi}}_{u,v,w}, \delta\vec{\eta} \right\rangle + \left\langle \vec{\vec{\Xi}}_{p,q,r}, \delta\vec{\phi} \right\rangle \right] dt. \tag{276}$$

Now, the lagrangian is invariant under coordinate transformation, so it is a function of the *free* vectors of velocity and angular velocity. Using the notation detailed in the Nomenclature section below, $\delta\vec{\eta}$ and $\delta\vec{\phi}$ are interpreted as free vectors, while $\delta\underline{\eta}$ and $\delta\underline{\phi}$ are the projections of the free vectors onto a given reference frame. The following relationships link the displacements with the body-referenced rates:

$$\underline{v} = \frac{\partial \underline{\eta}}{\partial t} + \underline{\omega} \otimes \underline{\eta}, \text{ and} \tag{277}$$

$$\underline{\omega} = \frac{\partial \underline{\phi}}{\partial t}. \tag{278}$$

A variation of the lagrangian at a given time t is, to first order,

$$\delta L(\vec{v}, \vec{\omega}, t) = \left\langle \frac{\partial L}{\partial \vec{v}}, \delta\vec{v} \right\rangle + \left\langle \frac{\partial L}{\partial \vec{\omega}}, \delta\vec{\omega} \right\rangle \tag{279}$$

The variations $\delta\vec{v}$ and $\delta\vec{\omega}$, in view of equations 277 and 278, can be written as

$$\begin{aligned} \delta\vec{v} &= \delta\underline{v}^T \hat{x} + \{\underline{v}^T \delta\hat{x}\} \\ &= \left(\frac{\partial \delta\underline{\eta}}{\partial t} + \underline{\omega} \otimes \delta\underline{\eta} \right)^T \hat{x} + \vec{v} \times \delta\vec{\phi}, \end{aligned} \tag{280}$$

$$\begin{aligned} \delta\vec{\omega} &= \delta\underline{\omega}^T\hat{x} + \{\underline{\omega}^T\delta\hat{x}\} \\ &= \left(\frac{\partial\delta\underline{\phi}}{\partial t}\right)^T\hat{x} + \vec{\omega}\times\delta\vec{\phi}. \end{aligned} \tag{281}$$

The terms $\{\underline{v}^T\delta\hat{x}\}$ and $\{\underline{w}^T\delta\hat{x}\}$ above represent the effects of the variation of body orientation, and lead to one of the more subtle points of the derivation. It can be shown that $\delta\hat{x}$, the displacement of the unit triad $\hat{x}$ is $\delta\vec{\phi}\times\hat{x}$, leading for example to $\underline{v}^T(\delta\vec{\phi}\times\hat{x})$. This form, however, fails to capture what is meant by $\{\underline{v}^T\delta\hat{x}\}$ and $\{\underline{w}^T\delta\hat{x}\}$, specifically: how the free vectors $\delta\vec{v}$ and $\delta\vec{w}$ change as the triad rotates, but the projections $\underline{v}$ and $\underline{w}$ *remain constant*. With this in mind, one can easily derive the proper interpretations, as written above.
We now return to the evaluation of the lagrangian L. Combining terms, we see that there will be five inner products to consider. Writing the terms involving the time derivatives as δL_1 and δI_1, we have from an integration by parts

$$\begin{aligned} \delta I_1 &= \int_{t_1}^{t_2}\left[\left\langle\frac{\partial L}{\partial\vec{v}},\left(\frac{\partial\delta\underline{\eta}}{\partial t}\right)^T\hat{x}\right\rangle + \left\langle\frac{\partial L}{\partial\vec{\omega}},\left(\frac{\partial\delta\underline{\phi}}{\partial t}\right)^T\hat{x}\right\rangle\right]dt \\ &= -\int_{t_1}^{t_2}\left[\left\langle\frac{\partial}{\partial t}\frac{\partial L}{\partial\vec{v}},\delta\vec{\eta}\right\rangle + \left\langle\frac{\partial}{\partial t}\frac{\partial L}{\partial\vec{\omega}},\delta\vec{\phi}\right\rangle\right]dt. \end{aligned} \tag{282}$$

There are no terms remaining at the time boundaries because the lagrangian is zero at these points.
In evaluating the remaining terms, in δI_2, one needs to use the following property of triple vector products:

$$<\vec{a},\vec{b}\times\vec{c}>=<\vec{c},\vec{b}\times\vec{a}>=<\vec{b},\vec{c}\times\vec{a}>. \tag{283}$$

Hence, the variation of the action I_2, expressed in terms of vectors, is

$$\begin{aligned} \delta I_2 &= \int_{t_1}^{t_2}\left[\left\langle\frac{\partial L}{\partial\vec{v}},\vec{\omega}\times\delta\vec{\eta}+\vec{v}\times\delta\vec{\phi}\right\rangle + \left\langle\frac{\partial L}{\partial\vec{\omega}},\vec{\omega}\times\delta\vec{\phi}\right\rangle\right]dt \\ &= -\int_{t_1}^{t_2}\left[\left\langle\vec{\omega}\times\frac{\partial L}{\partial\vec{v}},\delta\vec{\eta}\right\rangle + \left\langle\vec{v}\times\frac{\partial L}{\partial\vec{v}}+\vec{\omega}\times\frac{\partial L}{\partial\vec{\omega}},\delta\vec{\phi}\right\rangle\right]dt. \end{aligned} \tag{284}$$

Finally, we write the part of δI due to the generalized forces:

$$\delta I_3 = \int_{t_1}^{t_2}\left[\left\langle\vec{\Xi}_{u,v,w},\delta\vec{\eta}\right\rangle + \left\langle\vec{\Xi}_{p,q,r},\delta\vec{\phi}\right\rangle\right]dt. \tag{285}$$

The Kirchoff relations follow directly from combining the three action variations. The kinetic co-energy we developed is that of the fluid, and so the generalized forces indicate what forces the body would exert on the fluid; with a sign change, these are the forces that the fluid exerts on the body.

23.5 Nomenclature

23.5.1 Free versus Column Vector

We make the distinction between a **free vector** $\vec{f}$, which is an element of a linear vector field, and a **column vector** $\underline{f}$ which denotes the components of the vector $\vec{f}$ in a given coordinate system. The connection between the two concepts is given via the free triad $\hat{x}$, containing as elements the unit vectors of the chosen Cartesian system, $\hat{i}, \hat{j}, \hat{k}$, i.e.:

$$\hat{x} = (\hat{i}, \hat{j}, \hat{k})^T \tag{286}$$

The notation is hybrid, but convenient since it allows us to write:

$$\vec{f} = \underline{f}^T \hat{x} \tag{287}$$

where $\underline{f}^T$ denotes the transpose of $\underline{f}$, i.e. a row vector. The product between the row vector $\underline{f}^T$ and the column vector $\hat{x}$ is in the usual matrix multiplication sense.

23.5.2 Derivative of a Scalar with Respect to a Vector

The derivative of a scalar L with respect to a vector $\vec{x}$ is denoted as:

$$\frac{\partial L}{\partial \vec{x}} \tag{288}$$

and is a vector with the same dimension as $\vec{x}$, whose element i is the derivative of L with respect to the ith element of $\vec{x}$, and in the same direction:

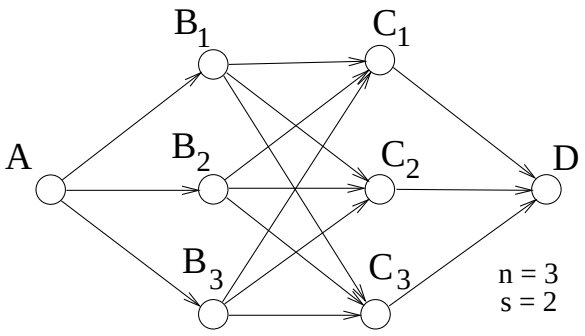
$$\left(\frac{\partial L}{\partial \vec{x}}\right)_i = \frac{\partial L}{\partial \vec{x}_i} \tag{289}$$

23.5.3 Dot and Cross Product

The dot product of two free vectors $\vec{f}$ and $\vec{g}$ is denoted as $< \vec{f}, \vec{g} >$, and we use the following notation for column vectors: $< \underline{f}^T \hat{x}, \underline{g}^T \hat{x} >$. The cross product of two vectors $\vec{f}$ and $\vec{g}$ is $\vec{z} = \vec{f} \times \vec{g}$, denoted in terms of column vectors as $\underline{z} = \underline{f} \otimes \underline{g}$.

24 APPENDIX 3: LQR VIA DYNAMIC PROGRAMMING

There are at least two conventional derivations for the LQR; we present here one based on *dynamic programming*, due to R. Bellman. The key observation is best given through a loose example.



24.1 Example in the Case of Discrete States

Suppose that we are driving from Point A to Point C, and we ask what is the shortest path in miles. If A and C represent Los Angeles and Boston, for example, there are *many* paths to choose from! Assume that one way or another we have found the best path, and that a Point B lies along this path, say Las Vegas. Let X be an arbitrary point east of Las Vegas. If we were to now solve the optimization problem for getting from only Las Vegas to Boston, this same arbitrary point X would be along the new optimal path as well.
The point is a subtle one: the optimization problem from Las Vegas to Boston is easier than that from Los Angeles to Boston, and the idea is to use this property *backwards* through time to evolve the optimal path, beginning in Boston.

Example: Nodal Travel. We now add some structure to the above experiment. Consider now traveling from point A (Los Angeles) to Point D (Boston). Suppose there are only three places to cross the Rocky Mountains, B_1, B_2, B_3, and three places to cross the Mississippi River, C_1, C_2, C_3.[3] By way of notation, we say that the path from A to B_1 is AB_1. Suppose that all of the paths (and distances) from A to the B-nodes are known, as are those from the B-nodes to the C-nodes, and the C-nodes to the terminal point D. There are nine unique paths from A to D.
A brute-force approach sums up the total distance for all the possible paths, and picks the shortest one. In terms of computations, we could summarize that this method requires nine additions of three numbers, equivalent to eighteen additions of two numbers. The *comparison* of numbers is relatively cheap.
The dynamic programming approach has two steps. First, from each B-node, pick the best path to D. There are three possible paths from B_1 to D, for example, and nine paths total from the B-level to D. Store the best paths as $B_1D|_{opt}, B_2D|_{opt}, B_3D|_{opt}$. This operation involves nine additions of two numbers.
Second, compute the distance for each of the possible paths from A to D, *constrained to the optimal paths from the B-nodes onward*: $AB_1 + B_1D|_{opt}$, $AB_2 + B_2D|_{opt}$, or $AB_3 + B_3D|_{opt}$. The combined path with the shortest distance is the total solution; this second step involves three sums of two numbers, and total optimization is done in twelve additions of two numbers. Needless to say, this example gives only a mild advantage to the dynamic programming

[3]Apologies to readers not familiar with American geography.

approach over brute force. The gap widens vastly, however, as one increases the dimensions of the solution space. In general, if there are s layers of nodes (e.g., rivers or mountain ranges), and each has width n (e.g., n river crossing points), the brute force approach will take (sn^s) additions, while the dynamic programming procedure involves only $(n^2(s-1)+n)$ additions. In the case of $n = 5$, $s = 5$, brute force requires 6250 additions; dynamic programming needs only 105.

24.2 Dynamic Programming and Full-State Feedback

We consider here the regulation problem, that is, of keeping $x_{desired} = 0$. The closed-loop system thus is intended to reject disturbances and recover from initial conditions, but not necessarily follow y-trajectories. There are several necessary definitions. First we define an instantaneous *penalty* function $l(x(t), u(t))$, which is to be greater than zero for all nonzero x and u. The *cost* associated with this penalty, along an optimal trajectory, is

$$J = \int_0^\infty l(x(t), u(t))dt, \tag{290}$$

i.e., the integral over time of the instantaneous penalty. Finally, the *optimal return* is the cost of the optimal trajectory remaining after time t:

$$V(x(t), u(t)) = \int_t^\infty l(x(\tau), u(\tau))d\tau. \tag{291}$$

.

We have directly from the dynamic programming principle

$$V(x(t), u(t)) = \min_u \left\{ l(x(t), u(t))\delta t + V(x(t+\delta t), u(t+\delta t)) \right\}. \tag{292}$$

The minimization of $V(x(t), u(t))$ is made by considering all the possible control inputs u in the time interval $(t, t+\delta t)$. As suggested by dynamic programming, the return at time t is constructed from the return at $t+\delta t$, and the differential component due to $l(x, u)$. If V is smooth and has no explicit dependence on t, as written, then

$$\begin{aligned} V(x(t+\delta t), u(t+\delta t)) &= V(x(t), u(t)) + \frac{\partial V}{\partial x}\frac{dx}{dt}\,\delta t + h.o.t. \longrightarrow \\ &= V(x(t), u(t)) + \frac{\partial V}{\partial x}(Ax(t) + Bu(t))\delta t. \end{aligned} \tag{293}$$

Now control input u in the interval $(t, t+\delta t)$ cannot affect $V(x(t), u(t))$, so inserting the above and making a cancellation gives

$$0 = \min_u \left\{ l(x(t), u(t)) + \frac{\partial V}{\partial x}(Ax(t) + Bu(t)) \right\}. \tag{294}$$

We next make the assumption that $V(x, u)$ has the following form:

$$V(x, u) = \frac{1}{2}x^T P x, \tag{295}$$

where P is a symmetric matrix, and positive definite.[4][5] It follows that

$$\begin{aligned} \frac{\partial V}{\partial x} &= x^T P \longrightarrow \\ 0 &= \min_u \left\{ l(x,u) + x^T P(Ax + Bu) \right\}. \end{aligned} \tag{296}$$

We finally specify the instantaneous penalty function. The LQR employs the special quadratic form

$$l(x,u) = \frac{1}{2} x^T Q x + \frac{1}{2} u^T R u, \tag{297}$$

where Q and R are both symmetric and positive definite. The matrices Q and R are to be set by the user, and represent the main "tuning knobs" for the LQR. Substitution of this form into the above equation, and setting the derivative with respect to u to zero gives

$$\begin{aligned} 0 &= u^T R + x^T P B \\ u^T &= -x^T P B R^{-1} \\ u &= -R^{-1} B^T P x. \end{aligned} \tag{298}$$

The **gain matrix** for the feedback control is thus $K = R^{-1} B^T P$. Inserting this solution back into equation 297, and eliminating u in favor of x, we have

$$0 = \frac{1}{2} x^T Q x - \frac{1}{2} x^T P B R^{-1} B^T P + x^T P A x.$$

All the matrices here are symmetric except for PA; since $x^T P A x = x^T A^T P x$, we can make its effect symmetric by letting

$$x^T P A x = \frac{1}{2} x^T P A x + \frac{1}{2} x^T A^T P x,$$

leading to the final matrix-only result

$$0 = Q + PA + A^T P - P B R^{-1} B^T P. \tag{299}$$

25 Further Robustness of the LQR

The most common robustness measures attributed to the LQR are a one-half gain reduction in any input channel, an infinite gain amplification in any input channel, or a phase error of plus or minus sixty degrees in any input channel. While these are general properties

[4] Positive definiteness means that $x^T P x > 0$ for all nonzero x, and $x^T P x = 0$ if $x = 0$.

[5] This suggested form for the optimal return can be confirmed after the optimal controller is derived.

that have a clear graphical implication on the Bode or Nyquist plot, other useful robustness conditions can be developed. These include robustness to uncertainty in the real *coefficients* of the model (e.g., coefficients in the A matrix), and certain nonlinearities, including control switching and saturation. We will use the Lyapunov stability and the LMI formulation of matrix problems in this section to expand these ideas.
Saturation nonlinearities in particular are ubiquitous in control system application; we find them in "railed" conditions of amplifiers, rate and position limits in control surface actuators, and in well-meaning but ad hoc software limits. As shown below, moderate robustness in saturation follows from the basic analysis, but much stronger results can be obtained with new tools.
When the LQR is used to define the loop shape in the loop transfer recovery method (as opposed to the Kalman filter in the more common case), these robustness measures hold.

25.1 Tools

25.1.1 Lyapunov's Second Method

The idea of Lyapunov's Second Method is the following: if a positive definite function of the state $\vec{x}$ can be found, with $V(\vec{x}) = 0$ only when $\vec{x} = \vec{0}$, and if $dV(\vec{x})/dt < 0$ for all time, then the system is stable. A useful analogy for the Lyapunov function $V(\vec{x})$ is energy, dissipated in a passive mechanical system by damping, or in a passive electrical system through resistance.

25.1.2 Matrix Inequality Definition

Inequalities in the case of matrices are in the sense of positive and negative (semi) definiteness. Positive definite A means $\vec{x}^{\,T} A \vec{x} > 0$ for all $\vec{x}$; positive semidefinite A means $\vec{x}^{\,T} A \vec{x} \geq 0$ for all $\vec{x}$. With A and B square and of the same dimension,

$$A < B \text{ means } \vec{x}^{\,T} A \vec{x} < \vec{x}^{\,T} B \vec{x}, \text{ for all } \vec{x}. \tag{300}$$

Also, we say for the case of a scalar γ,

$$A < \gamma \text{ means } A - \gamma I < 0. \tag{301}$$

25.1.3 Franklin Inequality

A theorem we can use to assist in the Lyapunov analysis is the following, attributed to Franklin (1969).

$$A^T B + B^T A \leq \gamma A^T A + \frac{1}{\gamma} B^T B, \text{ for all real } \gamma > 0. \tag{302}$$

The scalar γ is a free parameter we can specify. It is assumed that the matrices A and B are of consistent dimensions.

25.1.4 Schur Complement

Consider the symmetric block matrix defined as follows:

$$M = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix}. \tag{303}$$

The *Schur complements of blocks* A and D are defined, respectively, as

$$\begin{aligned} \Gamma_A &= D - B^T A^{-1} B \\ \Gamma_D &= A - B D^{-1} B^T, \end{aligned} \tag{304}$$

and an important property is that

$$M > 0 \longleftrightarrow \Gamma_A > 0 \text{ and } \Gamma_D > 0 \text{ and } A > 0 \text{ and } D > 0. \tag{305}$$

The Schur complements thus capture the sign of M, but in a smaller dimensioned matrix equation. The fact that both A and D have to be positive definite, independent of the Schur complements, is obvious by considering $\vec{x}$ that involve only A or D, e.g., $\vec{x} = [\vec{1}_{1\times n_A}\ \vec{0}_{1\times n_D}]^T$, where n_A and n_D are the dimensions of A and D.

25.1.5 Proof of Schur Complement Sign

It is easy to verify that

$$M = \begin{bmatrix} A & B \\ B^T & D \end{bmatrix} = \begin{bmatrix} I & 0 \\ B^T A^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & D - B^T A^{-1} B \end{bmatrix} \begin{bmatrix} I & A^{-1} B \\ 0 & I \end{bmatrix} \tag{306}$$

The outer matrices are nonsingular so there is a one-to-one transformation

$$\vec{z} = \begin{bmatrix} I & A^{-1} B \\ 0 & I \end{bmatrix} \vec{x}, \tag{307}$$

and thus $M > 0$ is equivalent to $A > 0$ and $\Gamma_A > 0$. The proof for Γ_D is completely analogous.

25.1.6 Schur Complement of a Nine-Block Matrix

For the purposes of derivation below, consider now the symmetric matrix

$$M = \begin{bmatrix} A & B & C \\ B^T & D & 0 \\ C^T & 0 & E \end{bmatrix}. \tag{308}$$

Using the Schur complement of the diagonal block including D and E, we have

$$M > 0 \longleftrightarrow A - B D^{-1} B^T - C E^{-1} C^T > 0 \text{ and } D > 0 \text{ and } E > 0. \tag{309}$$

Again, positive definiteness of a large matrix is reflected in a smaller matrix inequality.

25.1.7 Quadratic Optimization with a Linear Constraint

Consider the quadratic function $V = \vec{x}^{\,T} P \vec{x}$, P symmetric and positive definite, and the minimization of V subject to a constraint $\vec{c}^{\,T}\vec{x} = a$, where $\vec{c}^{\,T}$ is a row vector, and a is a scalar. The solution is:

$$\min V = \frac{a^2}{\vec{c}^{\,T} P^{-1} \vec{c}}. \tag{310}$$

To show that this is true, augment the objective with a LaGrange multiplier to become

$$\tilde{V} = V + \lambda(\vec{c}^{\,T}\vec{x} - a). \tag{311}$$

Setting $\partial\tilde{V}/\partial\vec{x} = 0$ gives $2\vec{x}^{\,T}P + \lambda\vec{c}^{\,T} = 0$. Post-multiply this equation by $P^{-1}\vec{c}$, and solve to give $\lambda = -2a/\vec{c}^{\,T}P^{-1}\vec{c}$. Then, using $2\vec{x}^{\,T}P + \lambda\vec{c}^{\,T} = 0$ again, post-multiply by $\vec{x}/2$, and insert λ, concluding the proof.

25.2 Comments on Linear Matrix Inequalities (LMI's)

The reason that solving $M > 0$ is simpler than the equivalent Schur complement inequality is that, as long as $M(\vec{y})$ is *affine* in the unknowns $\vec{y}$, the solution set is *convex*. Affine means that

$$M(\vec{y}) = M_o + K[\vec{y}\;\vec{y}\;...\;\vec{y}]N, \tag{312}$$

i.e., that M is a constant matrix plus terms linear in each element of $\vec{y}$. Note the matrices K and N here, as well as the columnwise expansion of $\vec{y}$, are needed to put elements of $\vec{y}$ in arbitrary locations of M.

Convex means that if $\vec{y}_1$ and $\vec{y}_2$ are feasible solutions to the inequality, then so is $\vec{y}_3 = \alpha\vec{y}_1 + (1-\alpha)\vec{y}_2$, for any α between zero and one. A convex solution set has no hidden corners or shadows, because every point on a straight line between two solutions is also a solution. Computer programs solving convex optimization problems can always get the global optimum efficiently, or determine that no solution exists.

From the Schur complements above, clearly we can transform a (suitable) matrix equation that is quadratic or higher in the unknowns into an affine $M(\vec{y})$, and therefore make an efficient solution. The term *linear matrix inequality* refers of course to the fact that M is affine and hence linear in the unknown variables, but also to the methods of analysis and numerical solution, wherein the idea is to recognize

$$M(\vec{y}) = M_o + M_1 y_1 + M_2 y_2 + ...M_n y_n, \tag{313}$$

where n is the dimension of the unknown vector $\vec{y}$.

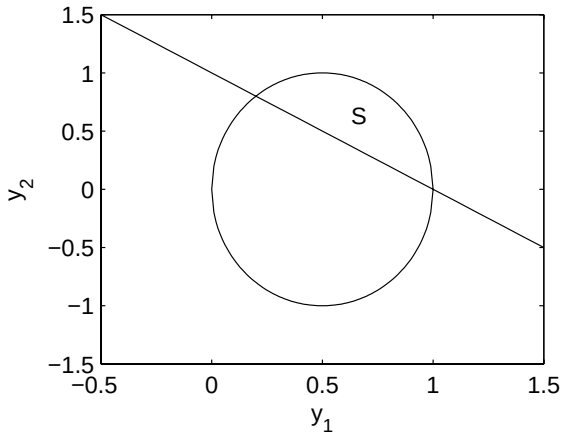
Example. Consider the LMI

$$M(\vec{y}) = \begin{bmatrix} y_1 - 1 & 2y_2 - 1 \\ y_2 + 1 & y_2 - y_1 \end{bmatrix} > 0. \tag{314}$$

We use the theorem that $M > 0$ is achieved if each of the submatrices cornered at the (1,1) element has positive determinant. Then $M > 0$ is equivalent to

$$\begin{aligned} y_1 &> 1 \\ (y_1 - 1)(y_2 - y_1) - (2y_2 - 1)(y_2 + 1) &> 0. \end{aligned} \tag{315}$$

The solution set is shown as S in the figure below; M is affine in $\vec{y}$ and the solution set is convex.



25.3 Parametric Uncertainty in A and B Matrices

25.3.1 General Case

Let S be the solution to the matrix Riccati equation

$$A^T S + SA + Q - PBR^{-1}B^T P = 0. \tag{316}$$

We consider here perturbations to the design plant matrices A and B of the form $\dot{\vec{x}} = (A + \Delta A)\vec{x} + (B + \Delta B)\vec{u}$. The control in the LQR is $\vec{u} = -R^{-1}B^T P\vec{x}$. Select a Lyapunov function of the form $V(\vec{x}) = \vec{x}^{\,T} S\vec{x}$. Then

$$\begin{aligned} \dot{V} &= \dot{\vec{x}}^{\,T} S\vec{x} + \vec{x}^{\,T} S\dot{\vec{x}} \\ &= \vec{x}^{\,T}\Big[A^T S + SA + \Delta A^T S + S\Delta A - 2SBR^{-1}B^T S - \\ &\qquad SBR^{-1}\Delta B^T S - S\Delta BR^{-1}BS\Big]\vec{x}. \end{aligned} \tag{317}$$

We give the following special structures to the perturbations:

$$\begin{aligned} \Delta A &= D_A \Delta_A E_A \\ \Delta B &= D_B \Delta_B E_B. \end{aligned} \tag{318}$$

The idea is to capture all of the uncertainty in Δ_A and Δ_B and capture the structure of the perturbations using the other four matrices. We have, using the Franklin inequality,

$$\begin{aligned} \dot{V} &= \vec{x}^{\,T} \Big[A^T S + SA - 2SBR^{-1}B^T S + E_A^T \Delta_A^T D_A^T S + S D_A \Delta_A E_A \\ &\qquad - SBR^{-1} E_B^T \Delta_B^T D_B^T S - S D_B \Delta_B E_B R^{-1} B S \Big] \vec{x} \qquad (319) \\ &\leq \vec{x}^{\,T} \Bigg[A^T S + SA - 2SBR^{-1}B^T S + \gamma_A E_A^T \Delta_A^T \Delta_A E_A + \frac{1}{\gamma_A} S D_A D_A^T S \\ &\qquad + \gamma_B S B R^{-1} E_B^T \Delta_B^T \Delta_B E_B R^{-1} B^T S + \frac{1}{\gamma_B} S D_B D_B^T S \Bigg] \vec{x} \\ &\leq \vec{x}^{\,T} \Bigg[A^T S + SA - 2SBR^{-1}B^T S + \gamma_A E_A^T E_A + \frac{1}{\gamma_A} S D_A D_A^T S \\ &\qquad + \gamma_B S B R^{-1} E_B^T E_B R^{-1} B^T S + \frac{1}{\gamma_B} S D_B D_B^T S \Bigg] \vec{x}, \qquad (320) \end{aligned}$$

where the last inequality holds if

$$\begin{aligned} \Delta_A^T \Delta_A &\leq I, \\ \Delta_B^T \Delta_B &\leq I. \qquad (321) \\ & \qquad (322) \end{aligned}$$

For a given perturbation model with matrices D_A, E_A, D_B, and E_B, we thus obtain a matrix inequality which, if met, guarantees stability of the system for all $\vec{x}$:

$$\begin{aligned} A^T S + SA - 2SBR^{-1}B^T S + \gamma_A E_A^T E_A + \frac{1}{\gamma_A} S D_A D_A^T S \\ + \gamma_B S B R^{-1} E_B^T E_B R^{-1} B^T S + \frac{1}{\gamma_B} S D_B D_B^T S \;\; < 0. \end{aligned} \tag{323}$$

Since scalars γ_A and γ_B can take on any positive value, they remain to be picked, and in fact judicious choices are needed to maximize the robustness guarantees – i.e., to minimize the conservativeness.

25.3.2 Uncertainty in B

Putting aside ΔA for the moment, and substituting in the Ricatti equation into the above leads to

$$-Q - SBR^{-1}B^TS + \gamma_B SBR^{-1}E_B^TE_BR^{-1}B^TS + \frac{1}{\gamma_B}SD_BD_B^TS < 0. \tag{324}$$

Following the presentation of gain and phase margins of the LQR in a previous section, we here consider gain margins through the diagonal matrix N, such that

$$\Delta_B = BN = B\sqrt{R^{-1}|N|} \times I \times \sqrt{R|N|} = D_B\Delta_B E_B. \tag{325}$$

$|N|$ denotes the matrix made up of the absolute values of the elements of N. This Δ_B is a very specific structure, whose rationale becomes evident below. The stability inequality becomes

$$-Q - SBR^{-1}B^TS + \gamma_B SB|N|R^{-1}B^TS + \frac{1}{\gamma_B}SBR^{-1}|N|B^TS < 0, \tag{326}$$

or, equivalently,

$$I - \gamma_B|N| - \frac{1}{\gamma_B}|N| > 0, \tag{327}$$

because Q can be arbitrarily small in the LQR; certainly a large value of Q increases the robustness (see the case of ΔA below, for example). Since N is diagonal, it is sufficient to keep all

$$1 - |N_{i,i}|\left(\gamma_B + \frac{1}{\gamma_B}\right) > 0. \tag{328}$$

Solving for γ_B gives

$$\gamma_B = \frac{1 \pm \sqrt{1 - 4N_{i,i}^2}}{2|N_{i,i}|}, \tag{329}$$

showing that real, positive γ_B are attained for $|N_{i,i}| < 1/2$. Hence, this analysis confirms the one-half gain reduction property derived earlier. On the other hand, it confers only a one-half gain amplification robustness, which is overly conservative. Usage of the Franklin inequality to symmetrize the components involving ΔB is the cause of this, since it can be verified that using Equation 319 directly gives

$$\begin{aligned} -Q - SBR^{-1}(I + N + N)B^TS &< 0, \text{ or} \\ 1 + 2N_{i,i} &> 0, \end{aligned} \tag{330}$$

showing both the one-half reduction and infinite upward gain margins.
In summary, in the special case of diagonal N in the model of ΔB, we crafted D_B and E_B so as to align terms of the form $SBR^{-1}B^TS$, simplifying the analysis. This leads to a conservative result using the Franklin inequality, but the full, standard result on a second look. Better results could be obtained for specific cases, but probably with greater effort.

25.3.3 Uncertainty in A

Uncertainty in the system matrix A has a different role in robustness than does uncertainty in B, which is primarily an issue of gain and does not affect the open-loop stability. Perturbations to the design matrix A cause the open- and closed-loop poles to move, potentially destabilizing either or both of the open- and closed-loop systems. Consider the main inequality for stability again, setting aside ΔB now:

$$-Q - SBR^{-1}B^TS + \gamma_A E_A^T E_A + \frac{1}{\gamma_A} S D_A D_A^T S < 0. \tag{331}$$

Aligning the uncertain terms with $SBR^{-1}B^TS$ (as with ΔB above) proves to be impractical here because it imposes a specific structure on ΔA, and requires that elements of ΔA would have to change together. A more useful result involves a diagonal Δ_A, and for the purposes of illustration, we consider a second order system as follows:

$$\Delta_A = D_A \Delta_A E_A = \begin{bmatrix} d_1 & d_2 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \delta_1 & 0 \\ 0 & \delta_2 \end{bmatrix} I \tag{332}$$

$$= \begin{bmatrix} d_1\delta_1 & d_2\delta_2 \\ 0 & 0 \end{bmatrix}. \tag{333}$$

In addition, we set $B = [b \;\; 0]^T$, so that R is a scalar r, and we set Q diagonal. The resultant inequality is

$$-Q + \gamma_A I - S\left(\begin{bmatrix} b^2/r & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} (d_1^2 + d_2^2)/\gamma_A & 0 \\ 0 & 0 \end{bmatrix} \right) S < 0. \tag{334}$$

The condition for stability will be met if $Q - \gamma_A I > 0$ *and* $b^2/r - (d_1^2 + d_2^2)/\gamma_A > 0$, or equivalently, setting $\gamma_A = \min Q_{i,i}$, if

$$\frac{b^2}{r} \min Q_{i,i} > d_1^2 + d_2^2. \tag{335}$$

Now let, for example,

$$\begin{aligned} A &= \begin{bmatrix} -0.5 & -0.5 \\ 1 & 0 \end{bmatrix}, \\ B &= [1 \;\; 0]^T \\ Q &= diag(1, 1) \\ R &= 1. \end{aligned}$$

This informs us that the robustness condition is $d_1^2 + d_2^2 < 2$, and very substantial perturbations in A are allowed, for example

$$A + \Delta A = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix}.$$

Computed results that confirm the bound are shown in Figure 9. An interesting note is that the gain reduction margin in the presence of this ΔA suffers dramatically, with failure of the robustness if $b < 0.93$. Intuitively, the uncertainty in A has used up almost all of the available safety margin.

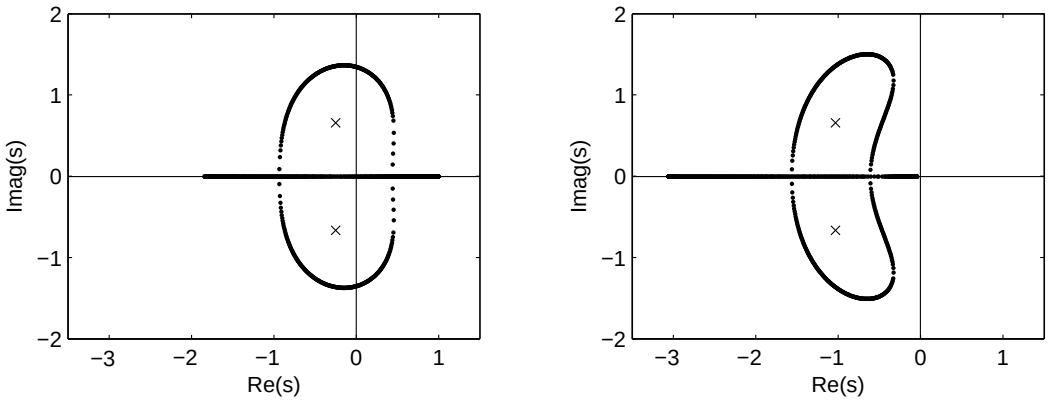


Figure 9: (left) open-loop pole locations along the boundary of the stability region in Equation 335; (right) closed-loop pole locations with LQR.

25.3.4 A and B Perturbations as an LMI

Using the triple LMI of Equation 309, we have, for the combined A and B structured uncertainty models,

$$\begin{aligned} -Q - SBR^{-1}B^TS + \gamma_A E_A^T E_A + \frac{1}{\gamma_A} SD_A D_A^T S \\ +\gamma_B SBR^{-1}E_B^T E_B R^{-1}B^T S + \frac{1}{\gamma_B} SD_B D_B^T S \;\; < \;\; 0 \end{aligned} \tag{336}$$

is equivalent to

$$\begin{bmatrix} -Q - SBR^{-1}B^TS + \gamma_A E_A^T E_A + \gamma_B SBR^{-1}E_B^T E_B R^{-1}B^T S & SD_A & SD_B \\ D_A^T S & -\gamma_A I & 0 \\ D_B^T S & 0 & -\gamma_B I \end{bmatrix} < 0. \tag{337}$$

The solution set $[\gamma_A \quad \gamma_B]$, if it exists, is convex. If a solution does not exist, one could lower the robustness by altering D_A, E_A, D_B, or E_B. Alternatively, increasing Q, the state penalty, where possible may increase robustness.

25.4 Input Nonlinearities

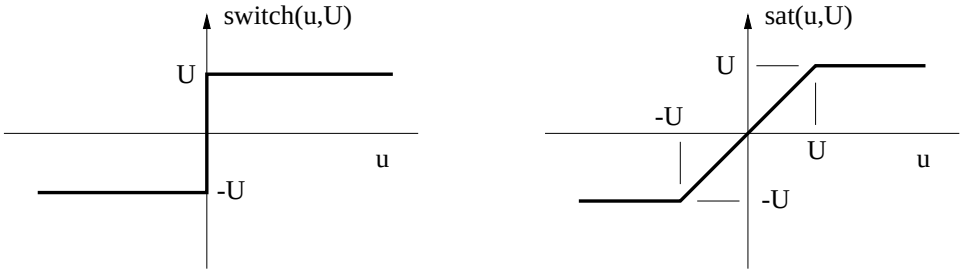
Control systems in practice often make use of switched or saturated action in the control channels. Switched and saturated control channels are defined respectively as

$$\text{switch}(u, U) = U\text{sign}(u), \tag{338}$$

and

$$\text{sat}(u, U) = U\text{sat}(u/U, 1) = \begin{cases} u, & |u| < U \\ U\text{sign}(u), & |u| > U. \end{cases}, \tag{339}$$

where the scalar U is positive. If the second argument of $\text{sat}(\cdot)$ is omitted, it is taken to be one.



Our stability analysis so far is one wherein we seek to guarantee a negative rate in $V(\vec{x})$ for all time. Referencing Equation 330, it is clear that *any* N can be tolerated within the range $(-.5, \infty)$, *even if it is time varying*. It follows directly then that switching and saturated control channels cannot destabilize the LQR system until the gain reduction is more than one-half, that is, $|u| > 2U$. In practice, however, systems routinely operate beyond this point, and this is the aspect of saturated control in particular that we wish to pursue now. This section details the approach of Huang & Lam (2002), which uses LMI's and the structured uncertainty models above.

We say that $f(\vec{u})$ is the diagonal function taking control commands from the LQR ($\vec{u}$) to the input to the plant; elements of f correspond to the saturated operator above. Through scaling the B matrix, we can set $U_i = 1$, and hence f's extreme outputs are ± 1.

First, we assert that under the proper choice of positive definite matrix W, and for certain vectors $\vec{z}$, the following inequality holds:

$$2\vec{z}^T f(R^{-1}\vec{z}) \geq \vec{z}^T W R^{-1}\vec{z}. \tag{340}$$

If the control penalty matrix R is diagonal, and we constrain W also to be diagonal, this condition is met if, for each control channel,

$$2z_i \text{sat}\left(\frac{z_i}{R_{i,i}}\right) \geq z_i^2 \frac{W_{i,i}}{R_{i,i}}. \tag{341}$$

If the channel is not saturated, then we require $W_{i,i} \leq 2$. If the channel is saturated, then it is required instead that

$$2\left|\frac{R_{i,i}}{z_i}\right| \geq W_{i,i}. \tag{342}$$

We will use the inequality of Equation 340 in our Lyapunov analysis, with special attention to these conditions. Because for the LQR we have $z_i = e_i^T B^T S\vec{x}$, where e_i is the unit vector in the i'th direction, Equation 342 effectively places a bound on the state under saturated control, namely

$$|e_i^T B^T S\vec{x}| \leq 2R_{i,i}/W_{i,i}. \tag{343}$$

Viewed this way, clearly we would like to make each $W_{i,i}$ as small as possible. As we show below, however, the matrix W appears in the only negative term of $\dot{V}(\vec{x})$; the optimization then is to maximize the state space satisfying Equation 342, while keeping W big enough to guarantee stability.
With $V(\vec{x}) = \vec{x}^{\,T} S\vec{x}$, and considering uncertainty ΔB only, we have

$$\dot{V}(\vec{x}) \;=\; \vec{x}^{\,T}(SA + A^T S)\vec{x} - 2\vec{x}^{\,T} SBf(R^{-1}B^T S\vec{x}) - 2\vec{x}^{\,T} S\Delta Bf(R^{-1}B^T S\vec{x}). \tag{344}$$

Note that ΔB never appears inside f because f operates on the control u, calculated with B alone. Applying the Franklin inequality to the last term yields

$$\begin{aligned}\dot{V}(\vec{x}) \;\leq\; & \vec{x}^{\,T}(SA + A^T S)\vec{x} - 2\vec{x}^{\,T} SBf(R^{-1}B^T S\vec{x}) \\ & +\gamma_B \vec{x}^{\,T} SD_B D_B^T S\vec{x} + \frac{1}{\gamma_B} f^T(R^{-1}B^T S\vec{x})E_B^T E_B f(R^{-1}B^T S\vec{x})).\end{aligned} \tag{345}$$

Next, consider the remaining asymmetric term. Applying Equation 340, under the proper constraints on W, we obtain

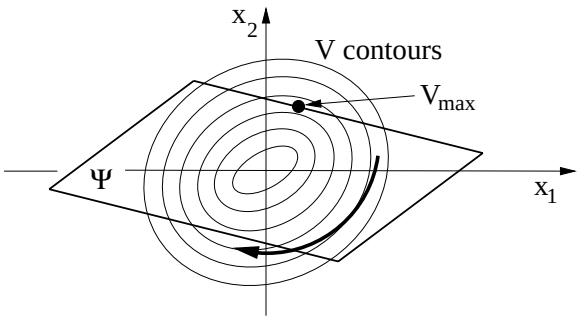
$$\begin{aligned}\dot{V}(\vec{x}) \;\leq\; & \vec{x}^{\,T}(SA + A^T S)\vec{x} - \vec{x}^{\,T} SBWR^{-1}B^T S\vec{x} \\ & +\gamma_B \vec{x}^{\,T} SD_B D_B^T S\vec{x} + \frac{1}{\gamma_B} f^T(R^{-1}B^T S\vec{x})E_B^T E_B f(R^{-1}B^T S\vec{x})).\end{aligned} \tag{346}$$

Finally, the last term is bounded, giving

$$SA + A^T S - SBWR^{-1}B^T S + \gamma_B SD_B D_B^T S + \frac{1}{\gamma_B} SBR^{-1}E_B^T E_B R^{-1}B^T S < 0 \tag{347}$$

as the sufficient condition for stability. This last step is extremely conservative, since in fact $|f_i| \leq 1$; Huang and Lam address this point.

The supporting conditions on elements of W are linear: $W_{i,i}$ must be as in Equation 342, and less than two. As noted above, these translate to linear constraints on $\vec{x}$. Call the set of states that satisfy these conditions Ψ. The problem now is that although the Lyapuonov function derivative is clearly less than zero for states within Ψ, the state trajectory could leave Ψ. This is entirely likely since V is a quadratic function, leading to elliptical contours as in the figure below.



As a result, the range of states leading to guaranteed stability is limited to those that are known not to leave Ψ. In other words, we have to settle for the largest ellipse completely contained in Ψ. We now invoke Equation 310, referencing Equation 343 as the set of linear constraints, to yield

$$V_{max} = \min_i \frac{4R_{i,i}^2}{W_{i,i}^2 e_i^T B^T S B e_i} \tag{348}$$

We would like to maximize V_{max}, which is equivalent to maximizing the minimum value of all the possible right hand sides of the above equation. This itself can be posed as an optimization in a new scalar k:

$$\text{minimize } k \text{ such that } kI - \frac{1}{2}R^{-1}W\Gamma > 0, \tag{349}$$

where Γ is diagonal, with $\Gamma_{i,i} = \sqrt{e_i^T B^T S B e_i}$. It follows that any $\vec{x}$ is stable when $V(\vec{x}) = \vec{x}^{\,T} S \vec{x} < 1/k^2$.

In summary, we have three constraints and one optimization: Equation 347 for Lyapunov stability, $0 \leq W_{i,i} \leq 2$ for cases where channels are not saturated, and Equation 349 for maximizing the level of V - an elliptical region in the state space - for which stability is guaranteed.

As in the case of no input saturation, this can be posed and solved as set of LMI's (which itself can be posed as a single LMI):

$$\begin{bmatrix} SA + A^T S - SBWR^{-1}B^T S + \gamma_B S D_B D_B^T S & SBR^{-1}E_B^T \\ E_B R^{-1} B^T S & \gamma_B I \end{bmatrix} < 0, \tag{350}$$

$$\frac{1}{2}R^{-1}W\Gamma < k, \tag{351}$$

$$W < 2, \tag{352}$$

$$0 < W. \tag{353}$$

To solve, pick a value for k and find a feasible solution. Decrease k as far as possible; the feasible solution achieves the maximum V from which the range of state space can be found from $V = 1/k^2$.