SYSTEM DESIGN FOR UNCERTAINTY

Franz S. Hover
Michael S. Triantafyllou

Center for Ocean Engineering
Department of Mechanical Engineering
Massachusetts Institute of Technology
Cambridge, Massachusetts USA

Latest Revision: December 1, 2009

Contents

1 INTRODUCTION 1

2 LINEAR SYSTEMS 2
2.1 Definition of a System . . . 2
2.2 Time-Invariant Systems . . . 3
2.3 Linear Systems . . . 4
2.4 The Impulse Response and Convolution . . . 5
2.5 Causal Systems . . . 7
2.6 An Example of Finding the Impulse Response . . . 7
2.7 Complex Numbers . . . 8
2.8 Fourier Transform . . . 9
2.9 The Angle of a Transfer Function . . . 11
2.10 The Laplace Transform . . . 11
2.10.1 Definition . . . 12
2.10.2 Convergence . . . 12
2.10.3 Convolution Theorem . . . 12
2.10.4 Solution of Differential Equations by Laplace Transform . . . 14

3 PROBABILITY 15
3.1 Events . . . 15
3.2 Conditional Probability . . . 16
3.3 Bayes' Rule . . . 17
3.4 Random Variables . . . 18
3.5 Continuous Random Variables and the Probability Density Function . . . 19
3.6 The Gaussian PDF . . . 20
3.7 The Cumulative Probability Function . . . 20
3.8 Central Limit Theorem . . . 22

4 RANDOM PROCESSES 23
4.1 Time Averages . . . 23
4.2 Ensemble Averages . . . 24
4.3 Stationarity . . . 24
4.4 The Spectrum: Definition . . . 26
4.5 Wiener-Khinchine Relation . . . 26
4.6 Spectrum Interpretation . . . 27

5 SHORT-TERM STATISTICS 29
5.1 Central Role of the Gaussian and Rayleigh Distributions . . . 30
5.2 Frequency of Upcrossings . . . 31
5.3 Maxima At and Above a Given Level . . . 31
5.4 1/N'th Highest Maxima . . . 32
5.5 1/N'th Average Value . . . 32
5.6 The 100-Year Wave: Estimate from Short-Term Statistics . . . 33

6 WATER WAVES 35
6.1 Constitutive and Governing Relations 35
6.2 Rotation and Viscous Effects 36
6.3 Velocity Potential 37
6.4 Linear Waves 38
6.5 Deepwater Waves 39
6.6 Wave Loading of Stationary and Moving Bodies 40
6.7 Limits of the Linear Theory 43
6.8 Characteristics of Real Ocean Waves 43

7 OPTIMIZATION 46
7.1 Single-Dimension Continuous Optimization 46
7.2 Multi-Dimensional Continuous Optimization 48
7.3 Linear Programming 51
7.4 Integer Linear Programming 52
7.5 Min-Max Optimization for Discrete Choices 54
7.6 Dynamic Programming 55
7.7 Solving Dynamic Programming on a Computer 57

8 STOCHASTIC SIMULATION 59
8.1 Monte Carlo Simulation 59
8.2 Making Random Numbers 61
8.3 Grid-Based Techniques 62
8.4 Issues of Cost and Accuracy 65

9 KINEMATICS OF MOVING FRAMES 67
9.1 Rotation of Reference Frames 67
9.2 Differential Rotations 69
9.3 Rate of Change of Euler Angles 71
9.4 A Practical Example: Dead Reckoning 72

10 VEHICLE INERTIAL DYNAMICS 73
10.1 Momentum of a Particle 73
10.2 Linear Momentum in a Moving Frame 74
10.3 Example: Mass on a String 75
10.3.1 Moving Frame Affixed to Mass 75
10.3.2 Rotating Frame Attached to Pivot Point 76
10.3.3 Stationary Frame 76
10.4 Angular Momentum 76
10.5 Example: Spinning Book 78
10.5.1 x-axis 79
10.5.2 y-axis 79
10.5.3 z-axis 79
10.6 Parallel Axis Theorem 80
10.7 Basis for Simulation 80

10.8 What Does an Inertial Measurement Unit Measure? 81

11 CONTROL FUNDAMENTALS 83

11.1 Introduction 83

11.1.1 Plants, Inputs, and Outputs 83

11.1.2 The Need for Modeling 83

11.1.3 Nonlinear Control 83

11.2 Partial Fractions 84

11.2.1 Partial Fractions: Unique Poles 84

11.2.2 Partial Fractions: Complex-Conjugate Poles 85

11.3 Stability in Linear Systems 86

11.4 Stability $\Longleftrightarrow$ Poles in LHP 86

11.5 General Stability 86

11.6 Representing Linear Systems 87

11.6.1 Standard State-Space Form 87

11.6.2 Converting a State-Space Model into a Transfer Function 88

11.6.3 Converting a Transfer Function into a State-Space Model 88

11.7 Block Diagrams and Transfer Functions of Feedback Systems 89

11.7.1 Block Diagrams: Fundamental Form 89

11.7.2 Block Diagrams: General Case 89

11.7.3 Transfer Functions 90

11.8 PID Controllers 91

11.9 Example: PID Control 92

11.9.1 Proportional Only 92

11.9.2 Proportional-Derivative Only 92

11.9.3 Proportional-Integral-Derivative 93

11.10Heuristic Tuning 93

12 CONTROL SYSTEMS – LOOPSHAPING 94

12.1 Introduction 94

12.2 Roots of Stability – Nyquist Criterion 94

12.2.1 Mapping Theorem 94

12.2.2 Nyquist Criterion 95

12.2.3 Robustness on the Nyquist Plot 96

12.3 Design for Nominal Performance 96

12.4 Design for Robustness 97

12.5 Robust Performance 98

12.6 Implications of Bode's Integral 99

12.7 The Recipe for Loopshaping 100

13 MATH FACTS 101

13.1 Vectors 101

13.1.1 Definition 101

13.1.2 Vector Magnitude 101

13.1.3 Vector Dot or Inner Product 102

13.1.4 Vector Cross Product 102
13.2 Matrices 102
13.2.1 Definition 102
13.2.2 Multiplying a Vector by a Matrix 103
13.2.3 Multiplying a Matrix by a Matrix 103
13.2.4 Common Matrices 104
13.2.5 Transpose 104
13.2.6 Determinant 104
13.2.7 Inverse 105
13.2.8 Eigenvalues and Eigenvectors 106
13.2.9 Modal Decomposition 107

1 INTRODUCTION

The material in these lecture notes is used in a junior- and senior-level course at MIT's Department of Mechanical Engineering, entitled *Design of Electromechanical Robotic Systems*. The key ingredient in this course - and what differentiates it from many other related courses - is the uncertainty that is encountered whenever a built system is actually installed in the field. We may find uncertainty in the operating environment, whether it is a windy airspace, a bumpy road, or an obstacle-strewn factory floor. We also find uncertainty in the parameters that define a system, for example, masses and stiffnesses and dampings, torque constants, and physical size. Finally, since complex electromechanical systems involve sensors and actuators, we have to acknowledge uncertainty in measurement and feedback control. Ultimately, such systems are meant to accomplish specific objectives, and the designer's task is to achieve robustness, performance, and cost-effectiveness in the presence of uncertainty.

The notes given here are terse but intended to be self-contained. The goal is to provide fundamental relations useful for modeling and creating systems that have to operate with uncertainty. As a motivation, we focus a lot of attention on ocean waves as a prototypical random environment, and carry out simplified, linear motion and force analysis for marine structures and vehicles. For teaching, I augment these notes with presentations that include some machine elements, fatigue, current navigation technology, and other topics. A separate compilation of solved homework problems is also made available, in the hopes that a) students will find a quick ramp-up to the problems of interest, and b) I can keep coming up with interesting problems! Finally, the class has an intensive, hands-on laboratory project in design of a field-able robotic system.

I will be grateful for any corrections, and for suggestions that will improve these notes.

Franz Hover
Cambridge, MA

2 LINEAR SYSTEMS

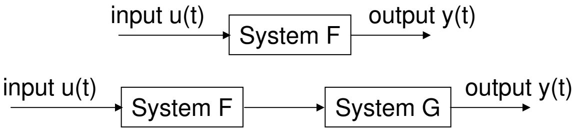
We will discuss what we mean by a linear time-invariant system, and then consider several useful transforms.

2.1 Definition of a System

In short, a system is any process or entity that has one or more well-defined inputs and one or more well-defined outputs. Examples of systems include a simple physical object obeying Newtonian mechanics, and the US economy!

Systems can be physical, or we may talk about a mathematical description of a system. The point of modeling is to capture in a mathematical representation the behavior of a physical system. As we will see, such representation lends itself to analysis and design, and certain restrictions such as linearity and time-invariance open a huge set of available tools.

We often use a block diagram form to describe systems, and in particular their interconnections:



In the second case shown, $y(t) = G[F[u(t)]]$.

Looking at structure now and starting with the most abstracted and general case, we may write a system as a function relating the input to the output; as written these are both functions of time:

$$y(t) = F[u(t)]$$

The system captured in F can be a multiplication by some constant factor - an example of a static system, or a hopelessly complex set of differential equations - an example of a dynamic system. If we are talking about a dynamical system, then by definition the mapping from $u(t)$ to $y(t)$ is such that the current value of the output $y(t)$ depends on the *past history* of $u(t)$. Several examples are:

$$\begin{aligned} y(t) &= \int_{t-3}^{t} u^2(t_1)dt_1, \\ y(t) &= u(t) + \sum_{n=1}^{N} u(t - n\delta t). \end{aligned}$$

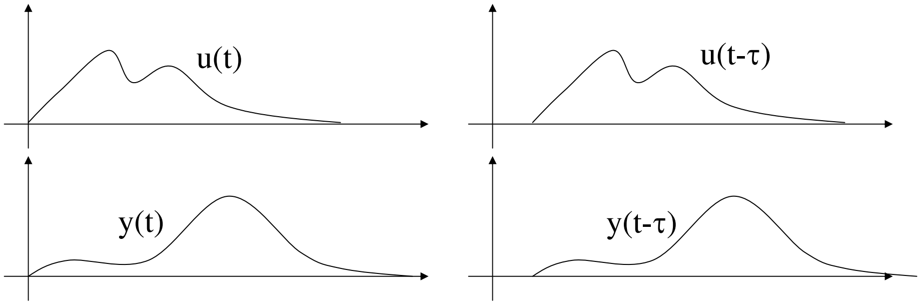
In the second case, δt is a constant time step, and hence $y(t)$ has embedded in it the current input plus a set of N delayed versions of the input.

2.2 Time-Invariant Systems

A dynamic system is time-invariant if shifting the input on the time axis leads to an equivalent shifting of the output along the time axis, with no other changes. In other words, a time-invariant system maps a given input trajectory $u(t)$ no matter when it occurs:

$$y(t-\tau) = F[u(t-\tau)].$$

The formula above says specifically that if an input signal is delayed by some amount τ, so will be the output, and with no other changes.



An example of a physical time-varying system is the pitch response of a rocket, $y(t)$, when the thrusters are being steered by an angle $u(t)$. You can see first that this is an inverted pendulum problem, and unstable without a closed-loop controller. It is time-varying because as the rocket burns fuel its mass is changing, and so the pitch responds differently to various inputs throughout its flight. In this case the "absolute time" coordinate is the time since liftoff.

To assess whether a system is time-varying or not, follow these steps: replace $u(t)$ with $u(t-\tau)$ on one side of the equation, replace $y(t)$ with $y(t-\tau)$ on the other side of the equation, and then check if they are equal. Here are several examples.

$$y(t) = u(t)^{3/2}$$

This system is clearly time-invariant, because it is a static map. Next example:

$$y(t) = \int_0^t \sqrt{u(t_1)}dt_1$$

Replace $u(t_1)$ with $u(t_1-\tau)$ in the right hand side and carry it through:

$$\int_0^t \sqrt{u(t_1-\tau)}dt_1 = \int_{-\tau}^{t-\tau} \sqrt{u(t_2)}dt_2.$$

The left hand side is simply

$$y(t-\tau) = \int_0^{t-\tau} \sqrt{u(t_1)}dt_1$$

Clearly the right and left hand sides are not equal (the limits of integration are different), and hence the system is not time-invariant. As another example, consider

$$y(t) = \int_{t-5}^{t} u^2(t_1) dt_1$$

The right-hand side becomes with the time shift

$$\int_{t-5}^{t} u^2(t_1 - \tau) dt_1 = \int_{t-5-\tau}^{t-\tau} u^2(t_2) dt_2,$$

whereas the left-hand side is

$$y(t - \tau) = \int_{t-5-\tau}^{t-\tau} u^2(t_1) dt_1;$$

the two sides of the defining equation are equal under a time shift τ, and so this system is time-invariant.

A subtlety here is encountered when considering inputs that are zero before time zero - this is the usual assumption in our work, namely $u(t) = 0$ for $t \leq 0$. While linearity is not affected by this condition, time invariance is, because the assumption is inconsistent with *advancing* a signal in time. Clearly part of the input would be truncated! Restricting our discussion to signal *delays* (the insertion of $-\tau$ into the argument, where strictly $\tau > 0$) resolves the issue, and preserves time invariance as needed.

2.3 Linear Systems

Next we consider linearity. Roughly speaking, a system is linear if its behavior is scale-independent; a result of this is the superposition principle. More precisely, suppose that $y_1(t) = F[u_1(t)]$ and $y_2(t) = F[u_2(t)]$. Then linearity means that for any two constants α_1 and α_2,

$$y(t) = \alpha_1 y_1(t) + \alpha_2 y_2(t) = F[\alpha_1 u_1(t) + \alpha_2 u_2(t)].$$

A simple special case is seen by setting $\alpha_2 = 0$:

$$y(t) = \alpha_1 y_1(t) = F[\alpha_1 u_1(t)],$$

making clear the scale-invariance. If the input is scaled by α_1, then so is the output. Here are some examples of linear and nonlinear systems:

$$\begin{aligned} y(t) &= c\frac{du}{dt} \text{ (linear and time-invariant)} \\ y(t) &= \int_0^t u(t_1) dt_1 \text{ (linear but not time-invariant)} \\ y(t) &= 2u^2(t) \text{ (nonlinear but time-invariant)} \\ y(t) &= 6u(t) \text{ (linear and time-invariant).} \end{aligned}$$

Linear, time-invariant (LTI) systems are of special interest because of the powerful tools we can apply to them. Systems described by sets of linear, ordinary or differential differential equations having constant coefficients are LTI. This is a large class! Very useful examples include a mass m on a spring k, being driven by a force $u(t)$:

$$my''(t) + ky(t) = u(t),$$

where the output $y(t)$ is interpreted as a position. A classic case of an LTI partial differential equation is transmission of lateral waves down a half-infinite string. Let m be the mass per unit length, and T be the tension (constant on the length). If the motion of the end is $u(t)$, then the lateral motion satisfies

$$m\frac{\partial^2 y(t,x)}{\partial t^2} = T\frac{\partial^2 y(t,x)}{\partial x^2}$$

with $y(t, x = 0) = u(t)$. Note that the system output y is not only a function of time but also of space in this case.

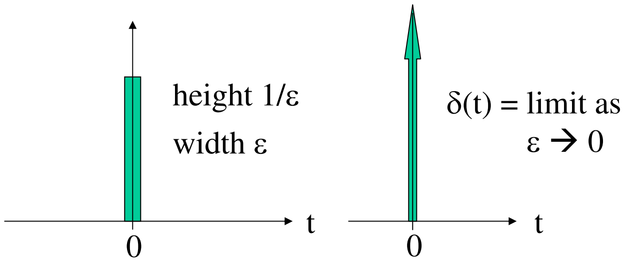
2.4 The Impulse Response and Convolution

A fundamental property of LTI systems is that they obey the convolution operator. This operator is defined by

$$y(t) = \int_{-\infty}^{\infty} u(t_1)h(t-t_1)dt_1 = \int_{-\infty}^{\infty} u(t-t_1)h(t_1)dt_1.$$

The function $h(t)$ above is a particular characterization of the LTI system known as the *impulse response* (see below). The equality between the two integrals should be clear since the limits of integration are infinite. The presence of the t_1 and the $-t_1$ term inside the integrations tells you that we have integrals of products - but that one of the signals is turned around. We will describe the meaning of the convolution more fully below.

To understand the impulse response, first we need the concept of the impulse itself, also known as the delta function $\delta(t)$. Think of a rectangular box centered at time zero, of width (time duration) ϵ, and height (magnitude) $1/\epsilon$; the limit as $\epsilon \longrightarrow 0$ is the δ function. The area is clearly one in any case.



The inner product of the delta function with any function is the value of the function at zero time:

$$\int_{-\infty}^{\infty} f(t)\delta(t)dt = \int_{-\epsilon/2}^{\epsilon/2} f(t)\delta(t)dt = f(t=0)\int_{-\epsilon/2}^{\epsilon/2}\delta(t)dt = f(0).$$

More usefully, the delta function can pick out the function value at a given, nonzero time ξ:

$$\int_{-\infty}^{\infty} f(t)\delta(t-\xi)dt = f(\xi).$$

Returning now to the impulse response function $h(t)$, it is, quite simply, the output of the LTI system, when driven by the delta function as input, that is $u(t) = \delta(t)$, or $h(t) = F[\delta(t)]$. In practical terms, we can liken $h(t)$ to the response of a mechanical system when it is struck very hard by a hammer!

Next we put the delta function and the convolution definition together, to show explicitly that the response of a system to arbitrary input $u(t)$ is the convolution of the input and the impulse response $h(t)$. This is what is stated in the definition given at the beginning of this section. First, we note that

$$\begin{aligned} u(t) &= \int_{-\infty}^{\infty} u(\xi)\delta(\xi - t)d\xi \\ &= \int_{-\infty}^{\infty} u(\xi)\delta(t-\xi)d\xi \text{ (because the impulse is symmetric about zero time).} \end{aligned}$$

Now set the system response $y(t) = F[u(t)]$, where F is an LTI system - we will use its two properties below.

$$\begin{aligned} y(t) &= F\left[\int_{-\infty}^{\infty} u(\xi)\delta(t-\xi)d\xi\right] \\ &= \int_{-\infty}^{\infty} u(\xi)F[\delta(t-\xi)]d\xi \text{ (using linearity)} \\ &= \int_{-\infty}^{\infty} u(\xi)h(t-\xi)d\xi \text{ (using time invariance),} \end{aligned}$$

and this indeed is the definition of convolution, often written as $y(t) = h(t) * u(t)$.

An intuitive understanding of convolution can be gained by thinking of the input as an infinite number of scaled delta functions, placed very closely together on the time axis. Explaining the case with the integrand $u(t-\xi)h(\xi)$, we see the convolution integral will call up all these virtual impulses, referenced to time t, and multiply them by the properly shifted impulse responses. Consider one impulse only that occurs at time $t = 2$, and we are interested in the response at $t = 5$. Then $u(t) = \delta(t-2)$ or $u(t-\xi) = \delta(t-2-\xi)$. The integrand will thus be nonzero only when $t-2-\xi$ is zero, or $\xi = t-2$. Now $h(\xi) = h(t-2)$ will be $h(3)$ when $t = 5$, and hence it provides the impulse response three time units after the impulse occurs, which is just what we wanted.

2.5 Causal Systems

All physical systems respond to input only *after* the input is applied. In math terms, this means $h(t) = 0$ for all $t < 0$. For convenience, we also usually consider input signals to be zero before time zero. The convolution is adapted in a very reasonable way:

$$\begin{aligned} y(t) &= \int_{-\infty}^{\infty} u(\xi)h(t-\xi)d\xi \\ &= \int_{0}^{\infty} u(\xi)h(t-\xi)d\xi \\ &= \int_{0}^{t} u(\xi)h(t-\xi)d\xi. \end{aligned}$$

The lower integration limit is set by the assumption that $u(t) = 0$ for $t < 0$, and the upper limit is set by the causality of the impulse response. The complementary form with integrand $u(t-\xi)h(\xi)$ also holds.

2.6 An Example of Finding the Impulse Response

Let's consider the differential equation $mx''(t) + bx'(t) + cx(t) = \delta(t)$, with the initial conditions of $x(0) = x'(0) = 0$. We have

$$\begin{aligned} \int_{-\epsilon/2}^{\epsilon/2} [mx'' + bx' + cx]dt &= \int_{-\epsilon/2}^{\epsilon/2} \delta(t)dt = 1, \text{ so that} \\ m(x'(0^+) - x'(0^-)) &= 1 \end{aligned}$$

The + superscript indicates the instant just after zero time, and the − superscript indicates the instant just before zero time. The given relation follows because at time zero the velocity and position are zero, so it must be the acceleration which is very large. Now since $x'(0-) = 0$, we have $x'(0+) = 1/m$. This is very useful - the initial velocity after the mass is hit with a $\delta(t)$ input. In fact, this *replaces* our previous initial condition $x'(0) = 0$, and we can treat the differential equation as homogeneous from here on. With $x(t) = c_1 e^{s_1 t} + c_2 e^{s_2 t}$, the governing equation becomes $ms_i^2 + bs_i + k = 0$ so that

$$s = -\frac{b}{2m} \pm \frac{\sqrt{b^2 - 4km}}{2m}.$$

Let $\sigma = b/2m$ and

$$\omega_d = \sqrt{\frac{k}{m} - \frac{b^2}{4m^2}},$$

and assuming that $b^2 < 4km$, we find

$$h(t) = \frac{1}{m\omega_d} e^{-\sigma t} sin(\omega_d t), \quad t \geq 0.$$

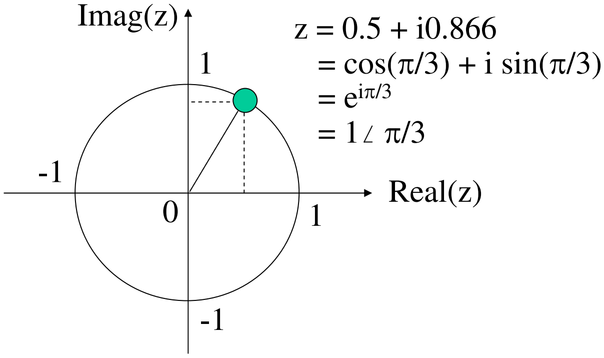
As noted above, once the impulse response is known for an LTI system, responses to all inputs can be found:

$$x(t) = \int_0^t u(\tau)h(t-\tau)d\tau.$$

In the case of LTI systems, the impulse response is a complete definition of the system, in the same way that a differential equation is, with zero initial conditions.

2.7 Complex Numbers

The complex number $z = x + iy$ is interpreted as follows: the real part is x, the imaginary part is y, and $i = \sqrt{-1}$ (imaginary). DeMoivre's theorem connects complex z with the complex exponential. It states that $\cos\theta + i\sin\theta = e^{i\theta}$, and so we can visualize any complex number in the two-plane, where the axes are the real part and the imaginary part. We say that $Re\left(e^{i\theta}\right) = cos\theta$, and $Im\left(e^{i\theta}\right) = sin\theta$, to denote the real and imaginary parts of a complex exponential. More generally, $Re(z) = x$ and $Im(z) = y$.



A complex number has a magnitude and an angle: $|z| = \sqrt{x^2 + y^2}$, and $\arg(z) = atan2(y,x)$. We can refer to the $[x,y]$ description of z as Cartesian coordinates, whereas the [magnitude, angle] description is called polar coordinates. This latter is usually written as $z = |z|\angle\arg(z)$. Arithmetic rules for two complex numbers z_1 and z_2 are as follows:

$$\begin{aligned}
z_1 + z_2 &= (x_1 + x_2) + i(y_1 + y_2) \\
z_1 - z_2 &= (x_1 - x_2) + i(y_1 - y_2) \\
z_1 \cdot z_2 &= |z_1||z_2|\angle\arg(z_1) + \arg(z_2) \\
z_1/z_2 &= \frac{|z_1|}{|z_2|}\angle\arg(z_1) - \arg(z_2)
\end{aligned}$$

Note that, as given, addition and subtraction are most naturally expressed in Cartesian coordinates, and multiplication and division are cleaner in polar coordinates.

2.8 Fourier Transform

The Fourier transform is the underlying principle for frequency-domain description of signals. We begin with the Fourier series.

Consider a signal $f(t)$ continuous on the time interval $[0, T]$, which then repeats with period T off to negative and positive infinity. It can be shown that

$$\begin{aligned} f(t) &= A_o + \sum_{n=1}^{\infty}[A_n \cos(n\omega_o t) + B_n \sin(n\omega_o t)], \text{ where} \\ \omega_o &= 2\pi/T, \\ A_0 &= \frac{1}{T}\int_0^T f(t)dt, \\ A_n &= \frac{2}{T}\int_0^T f(t)\cos(n\omega_o t)dt, \text{ and} \\ B_n &= \frac{2}{T}\int_0^T f(t)\sin(n\omega_o t)dt. \end{aligned}$$

This says that the time-domain signal $f(t)$ has an exact (if you carry all the infinity of terms) representation of a constant plus scaled cosines and sines. As we will see later, the impact of this second, frequency-domain representation is profound, as it allows an entirely new set of tools for manipulation and analysis of signals and systems. A compact form of these expressions for the Fourier series can be written using complex exponentials:

$$\begin{aligned} f(t) &= \sum_{n=-\infty}^{\infty} C_n e^{in\omega_o t}, \text{ where} \\ C_n &= \frac{1}{T}\int_0^T f(t)e^{-in\omega_o t}dt. \end{aligned}$$

Of course, C_n can be a complex number.

In making these inner product calculations, orthogonality of the harmonic functions is useful:

$$\begin{aligned} \int_0^{2\pi} \sin nt \ \sin mt \ dt &= 0, \text{ for } n \geq 1, \ m \geq 1, \ n \neq m \\ \int_0^{2\pi} \cos nt \ \cos mt \ dt &= 0, \text{ for } n \geq 1, \ m \geq 1, \ n \neq m \\ \int_0^{2\pi} \sin nt \ \cos mt \ dt &= 0, \text{ for } n \geq 1, \ m \geq 1. \end{aligned}$$

Now let's go to a different class of signal, one that is not periodic, but has a finite integral of absolute value. Obviously, such a signal has to approach zero at distances far from the origin. We can write a more elegant transformation:

$$\begin{aligned} F(\omega) &= \int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt, \\ f(t) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)e^{i\omega\tau}d\tau. \end{aligned}$$

This is the real Fourier transform: a time-domain signal is transformed into a (complex) frequency-domain version, and it can be transformed back. On working it through, we see that derivatives and integrals look this way through the transform:

$$\begin{aligned} f(t) &\longleftrightarrow F(\omega) \\ \frac{d^n f(t)}{dt^n} &\longleftrightarrow (i\omega)^n F(w) \\ \int_{-\infty}^{t} f(\tau)d\tau &\longleftrightarrow \frac{1}{i\omega}F(w). \end{aligned}$$

Another very important property of the Fourier transform is Parseval's Relation:

$$\int_{-\infty}^{\infty} f^2(t)dt = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(\omega)F^*(\omega)d\omega = \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega,$$

where the $*$-superscript indicates the complex conjugate. We will give more properties for the related Laplace transform in a later section. But as is, the Fourier transform is immediately useful for solving linear differential equations with constant coefficients (LTI systems):

$$\begin{aligned} mx'' + bx' + cx = u(t) \quad \longleftrightarrow \quad & [-m\omega^2 + i\omega b + k]X(\omega) = U(\omega), \text{ so that} \\ & X(\omega) = \frac{1}{-m\omega^2 + i\omega b + k}U(\omega). \end{aligned}$$

Hence, the action of the differential equation to relate $f(t)$ with $x(t)$ is, in the frequency domain, captured by the function

$$H(\omega) = \frac{1}{-m\omega^2 + i\omega b + k}$$

Putting two and two together, we then assert that $X(\omega) = H(\omega)U(\omega)$; in the Fourier space, the system response is the product of impulse response function, and the input! To back this up, we show now that convolution in the time-domain is equivalent to multiplication in the frequency domain:

$$\begin{aligned} X(\omega) &= \mathcal{F}\left[\int_{-\infty}^{\infty} u(\tau)h(t-\tau)d\tau\right] \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(\tau)h(t-\tau)e^{-i\omega t}dtd\tau \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} u(\tau)h(\xi)e^{-i\omega(\xi+\tau)}d\xi d\tau \quad \text{because } e^{-i\omega t} = e^{-i\omega(t-\tau+\tau)} \\ &= \int_{-\infty}^{\infty} e^{-i\omega\xi}h(\xi)d\xi \int_{-\infty}^{\infty} e^{-i\omega\tau}u(\tau)d\tau \\ &= H(\omega)U(\omega). \end{aligned}$$

The central role of the impulse response should be reiterated here. It is a complete definition of the system, and for systems of differential equations, it is a specific function of the parameters and of the frequency ω. The Fourier Transform of the impulse response called the system *transfer function,* and we often refer to the transfer function as "the system," even though it is actually a (transformed) signal.

By way of summary, we can write

$$\begin{aligned} y(t) &= h(t) * u(t), \text{ and} \\ Y(\omega) &= H(\omega)U(\omega). \end{aligned}$$

2.9 The Angle of a Transfer Function

A particularly useful property of the Fourier (and Laplace) transform is that the magnitude of the transfer function scales a sinusoidal input, and the angle of the transfer function adds to the angle of the sinusoidal input. In other words,

$$\begin{aligned} u(t) &= u_o \cos(\omega_o t + \psi) \longrightarrow \\ y(t) &= u_o |H(\omega_o)| cos(\omega_o t + \psi + \arg(H(\omega_o)). \end{aligned}$$

To prove the above relations, we'll use the complex exponential:

$$\begin{aligned} u(t) &= Re\left(u_o e^{i(\omega_o t + \psi)}\right) \\ &= Re\left(\tilde{u}_o e^{i\omega_o t}\right), \text{ making } u_o e^{i\psi} = \tilde{u}_o \text{ complex; then} \\ y(t) &= h(t) * u(t) \\ &= \int_{-\infty}^{\infty} h(\tau) u(t-\tau) d\tau \\ &= \int_{-\infty}^{\infty} h(\tau) Re\left(\tilde{u}_o e^{i\omega_o (t-\tau)}\right) d\tau \\ &= Re\left(\int_{-\infty}^{\infty} h(\tau) e^{-i\omega_o \tau} d\tau \;\; \tilde{u}_o e^{i\omega_o t}\right) \\ &= Re\left(H(\omega_o) u_o e^{i(\omega_o t + \psi)}\right) \\ &= u_o |H(\omega_o)| \cos(\omega_o t + \psi + \arg(H(\omega_o))). \end{aligned}$$

As an example, let $u(t) = 4\cos(3t + \pi/4)$, and $H(\omega) = 2i\omega/5$. Then $H(\omega_o) = H(3) = 6i/5 = 1.2\angle\pi/2$. Thus, $y(t) = 4.8\cos(3t + 3\pi/4)$.

2.10 The Laplace Transform

The causal version of the Fourier transform is the Laplace transform; the integral over time includes only positive values and hence only deals with causal impulse response functions. In our discussion, the Laplace transform is chiefly used in control system analysis and design.

2.10.1 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.

2.10.2 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.

2.10.3 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:

$$
\begin{array}{rrcl}
 & y(t) & \longleftrightarrow & Y(s) \\
\text{(Impulse)} & \delta(t) & \longleftrightarrow & 1 \\
\text{(Unit Step)} & 1(t) & \longleftrightarrow & \dfrac{1}{s} \\
\text{(Unit Ramp)} & t & \longleftrightarrow & \dfrac{1}{s^2} \\
 & e^{-\alpha t} & \longleftrightarrow & \dfrac{1}{s+\alpha} \\
 & \sin \omega t & \longleftrightarrow & \dfrac{\omega}{s^2+\omega^2} \\
 & \cos \omega t & \longleftrightarrow & \dfrac{s}{s^2+\omega^2} \\
 & e^{-\alpha t}\sin \omega t & \longleftrightarrow & \dfrac{\omega}{(s+\alpha)^2+\omega^2} \\
 & e^{-\alpha t}\cos \omega t & \longleftrightarrow & \dfrac{s+\alpha}{(s+\alpha)^2+\omega^2} \\
 & \dfrac{1}{b-a}(e^{-at}-e^{-bt}) & \longleftrightarrow & \dfrac{1}{(s+a)(s+b)} \\
 & \dfrac{1}{ab}\left[1+\dfrac{1}{a-b}(be^{-at}-ae^{-bt})\right] & \longleftrightarrow & \dfrac{1}{s(s+a)(s+b)} \\
 & \dfrac{\omega_n}{\sqrt{1-\zeta^2}}\; e^{-\zeta\omega_n t}\sin \omega_n\sqrt{1-\zeta^2}t & \longleftrightarrow & \dfrac{\omega_n^2}{s^2+2\zeta\omega_n s+\omega_n^2} \\
 & 1-\dfrac{1}{\sqrt{1-\zeta^2}}\; e^{-\zeta\omega_n t}\sin\left(\omega_n\sqrt{1-\zeta^2}t+\phi\right) & \longleftrightarrow & \dfrac{\omega_n^2}{s(s^2+2\zeta\omega_n s+\omega_n^2)} \\
 & \left(\phi=\tan^{-1}\dfrac{\sqrt{1-\zeta^2}}{\zeta}\right) & & \\
\text{(Pure Delay)} & y(t-\tau)1(t-\tau) & \longleftrightarrow & Y(s)e^{-s\tau} \\
\text{(Time Derivative)} & \dfrac{dy(t)}{dt} & \longleftrightarrow & sY(s)-y(0) \\
\text{(Time Integral)} & \displaystyle\int_0^t y(\tau)d\tau & \longleftrightarrow & \dfrac{Y(s)}{s}+\dfrac{\int_{0-}^{0+}y(t)dt}{s}
\end{array}
$$

$$\begin{aligned}
Y(s) &= \int_0^\infty y(t)e^{-st}dt \\
&= \int_0^\infty \left[\int_0^t g(t-\tau)\, x(\tau)\, d\tau\right] e^{-st}\, dt \\
&= \int_0^\infty \left[\int_0^\infty g(t-\tau)\, h(t-\tau)\, x(\tau)\, d\tau\right] e^{-st}dt \\
&= \int_0^\infty x(\tau) \left[\int_0^\infty g(t-\tau)\, h(t-\tau)\, e^{-st}\, dt\right] d\tau \\
&= \int_0^\infty x(\tau)\, G(s)e^{-s\tau}\, d\tau \\
&= G(s)X(s),
\end{aligned}$$

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 $x(t)$.

2.10.4 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.

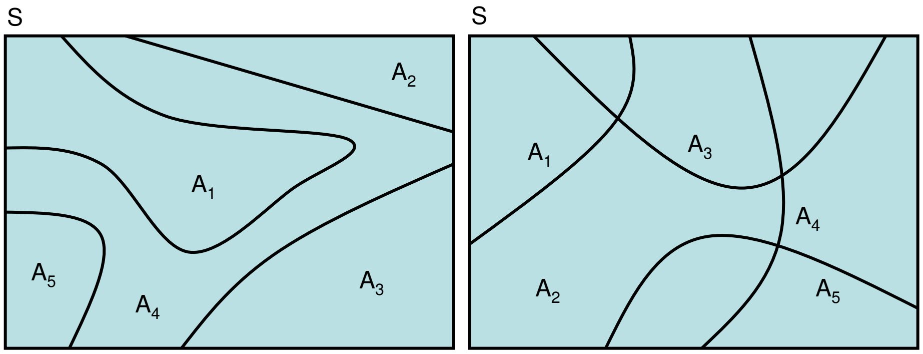
Specific examples of this procedure are given in a later section on transfer functions.

3 PROBABILITY

In this section, we discuss elements of probability, as a prerequisite for studying random processes.

3.1 Events

Define an event space S that has in it a number of events A_i. If the set of possible events A_i covers the space completely, then we will always get one of the events when we take a sample. On the other hand, if some of the space S is not covered with an A_i then it is possible that a sample is not classified as any of the events A_i. Events A_i may be overlapping in the event space, in which case they are *composite* events; a sample may invoke multiple events. But the A_i may not overlap, in which case they are *simple* events, and a sample brings only one event A_i, or none if the space S is not covered. In the drawing below, simple events cover the space on the left, and composite events cover the space on the right.



Intuitively, the probability of an event is the fraction of the number of positive outcomes to the total number of outcomes. Assign to each event a probability, so that we have

$$\begin{aligned} p_i &= p(A_i) \geq 0 \\ p(S) &= 1. \end{aligned}$$

That is, each defined event A_i has a probability of occurring that is greater than zero, and the probability of getting a sample from the entire event space is one. Hence, the probability has the interpretation of the area of the event A_i. It follows that the probability of A_i is exactly one minus the probability of A_i not occuring:

$$p(A_i) = 1 - p(\bar{A}_i).$$

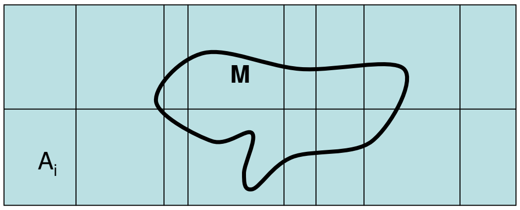
Furthermore, we say that if A_i and A_j are non-overlapping, then the probability of either A_i or A_j occuring is the same as the sum of the separate probabilities:

$$p(A_i \cup A_j) = p(A_i) + p(A_j).$$

Similarly if the A_i and A_j do overlap, then the probability of either or both occurring is the sum of the separate probabilities minus the sum of both occurring:

$$p(A_i \cup A_j) \quad = \quad p(A_i) + p(A_j) - p(A_i \cap A_j).$$

As a tangible example, consider a six-sided die. Here there are six events $A_1, A_2, A_3, A_4, A_5, A_6$, corresponding with the six possible values that occur in a sample, and $p(A_i) = 1/6$ for all i. The event that the sample is an even number is $M = A_2 \cup A_4 \cup A6$, and this is a composite event.



3.2 Conditional Probability

If a composite event M is known to have occurred, a question arises as to the probability that one of the constituent simple events A_i occurred. This is written as $P(A_j|M)$, read as "the probability of A_j, given M," and this is a conditional probability. The key concept here is that M replaces S as the event space, so that $p(M) = 1$. This will have the natural effect of inflating the probabilities of events that are part of event M, and in fact

$$p(A_j|M) = \frac{p(A_j \cap M)}{p(M)}.$$

Referring to our die example above, if M is the event of an even result, then we have

$$\begin{aligned} M &= A_2 \cup A_4 \cup A_6 \\ p(M \cap A_2) &= p(A_2) = 1/6 \\ p(M) &= 1/2 \longrightarrow \\ p(A_2|M) &= \frac{1/6}{1/2} = 1/3. \end{aligned}$$

Given that an event result was observed (composite event M), the probability that a two was rolled is 1/3. Now if all the A_j are independent (simple) events and M is a composite event, then we can write an opposing rule:

$$p(M) = p(M|A_1)p(A_1) + \cdots + p(M|A_n)p(A_n).$$

This relation collects conditional probabilities of M given each separate event A_i. Its logic is easily seen in a graph. Here is an example of how to use it in a practical problem. Box

A has 2000 items in it of which 5% are defective; box B has 500 items with 40% defective; boxes C and D each contain 1000 items with 10% defective. If a box is picked at random, and one item is taken from that box, what is the probability that it is defective? M is the composite event of a defective item, so we are after $p(M)$. We apply the formula above to find

$$p(M) = 0.05 \times 0.25 + 0.40 \times 0.25 + 0.10 \times 0.25 + 0.10 \times 0.25 = 0.1625.$$

3.3 Bayes' Rule

Consider a composite event M and a simple event A_i. We have from conditional probability above

$$p(A_i|M) = \frac{p(A_i \cap M)}{p(M)}$$
$$p(M|A_i) = \frac{p(A_i \cap M)}{p(A_i)},$$

and if we eliminate the denominator on the right-hand side, we find that

$$\begin{aligned} p(M|A_i) &= \frac{p(A_i|M)p(M)}{p(A_i)} \\ p(A_i|M) &= \frac{p(M|A_i)p(A_i)}{p(M)}. \end{aligned}$$

The second of these is most interesting - it gives the probability of a simple event, conditioned on the composite event, in terms of the composite event conditioned on the simple one! Recalling our above formula for $p(M)$, we thus derive Bayes' rule:

$$p(A_i|M) = \frac{p(M|A_i)p(A_i)}{p(M|A_1)p(A_1) + \cdots + p(M|A_n)p(A_n)}.$$

Here is an example of its use. Consider a medical test that is 99% accurate - it gives a negative result for people who do not have the disease 99% of the time, and it gives a positive result for people who do have the disease 99% of the time. Only one percent of the population has this disease. Joe just got a positive test result: What is the probability that he has the disease? The composite event M is that he has the disease, and the simple events are that he tested positive (+) or he tested negative (−). We apply

$$\begin{aligned} p(M|+) &= \frac{p(+|M)p(M)}{p(+)} \\ &= \frac{p(+|M)p(M)}{p(+|M)p(M) + p(+|\bar{M})p(\bar{M})} \\ &= \frac{0.99 \times 0.01}{0.99 \times 0.01 + 0.01 \times 0.99} \\ &= 1/2. \end{aligned}$$

This example is not well appreciated by many healthcare consumers!

Here is another example, without so many symmetries. Box A has nine red pillows in it and one white. Box B has six red pillows in it and nine white. Selecting a box at random and pulling out a pillow at random gives the result of a red pillow. What is the probability that it came from Box A? M is the composite event that it came from Box A; the simple event is that a red pillow was collected (R). We have

$$\begin{aligned} p(M|R) &= \frac{p(R|M)p(M)}{p(R)} \\ &= \frac{p(R|M)p(M)}{p(R|M)p(M) + p(R|\bar{M})p(\bar{M})} \\ &= \frac{0.9 \times 0.5}{0.9 \times 0.5 + 0.4 \times 0.5} \\ &= 0.692. \end{aligned}$$

3.4 Random Variables

Now we assign to each event A_i in the sample space a given value: each A_i corresponds with an x_i. For instance, a coin toss resulting in heads could be equated with a \$1 reward, and each tails could trigger a \$1 loss. Dollar figures could be assigned to each of the faces of a die. Hence we see that if each event A_i has a probability, then so will the numerical values x_i.

The average value of x_i can be approximated of course by sampling the space N times, summing all the x's, and dividing by N. As N becomes bigger, this computation will give an increasingly accurate result. In terms of probabilities the formula for the *expected value* is

$$\bar{x} = E(x) = \sum_{i=1}^{n} p(A_i)x_i.$$

The equivalence of this expected value with the numerical average is seen as follows: if the space is sampled N times, and the number of results $[A_i, x_i]$ is k_i, then $p(A_i) \simeq k_i/N$.

Superposition is an important property of the expectation operator:

$$E(x + y) = E(x) + E(y).$$

The mean of a function of x is defined using probabilities of the random variable x:

$$E\left[f(x(\xi))\right] = \sum_{i=1}^{n} f(x_i)p_i.$$

Another important property of a random variable is the variance - a measure of how much the x varies from its own mean:

$$\sigma^2 = E\left[(x - \bar{x})^2\right]$$

$$= \quad E(x^2) - \bar{x}^2.$$

The second line is apparent because $E(-2x\bar{x}) = -2\bar{x}^2$. Note we use the symbol σ^2 for variance; the standard deviation σ is just the square root, and has the same units as does the random variable x.

3.5 Continuous Random Variables and the Probability Density Function

Let us suppose now the random event has infinitely many outcomes: for example, the random variable x occurs *anywhere* in the range of $[0, 1]$. Clearly the probability of hitting any specific point is zero (although not impossible). We proceed this way:

$$p(x \text{ is in the range}[x_o, x_o + dx]) = \underline{p}(x_o)dx,$$

where $\underline{p}(x_o)$ is called the *probability density function.* Because all the probabilities that comprise it have to add up to one, we have

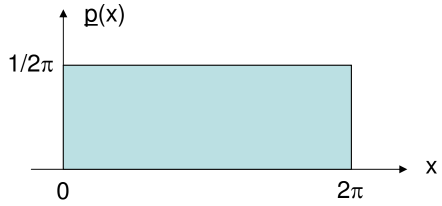
$$\int_{-\infty}^{\infty} \underline{p}(x)dx = 1.$$

With this definition, we can calculate the mean of the variable x and of a function of the variable $f(x)$:

$$\begin{aligned} E\left[x\right] &= \int_{-\infty}^{\infty} x\underline{p}(x)dx, \\ E\left[f(x)\right] &= \int_{-\infty}^{\infty} f(x)\underline{p}(x)dx. \end{aligned}$$

Here are a few examples. Consider a random variable that is equally likely to occur at any value between zero and 2π. Considering the area under $\underline{p}$ has to be one, we know then that $\underline{p}(x) = 1/2\pi$ for $x = [0, 2\pi]$ and it is zero everywhere else.

$$\begin{aligned} E(x) &= \pi \\ \sigma^2(x) &= \pi^2/3 \\ \sigma(x) &= \pi/\sqrt{3} \\ E(\cos x) &= \int_0^{2\pi} \frac{1}{2\pi} \cos x \; dx = 0 \\ E(\cos^2 x) &= \frac{1}{2}. \end{aligned}$$



The earlier concept of conditional probability carries over to random variables. For instance, considering this same example we can write

$$\begin{aligned} E\left[x|x>\pi\right] &= \int_0^{2\pi} x\underline{p}(x|x>\pi)\,dx \\ &= \int_\pi^{2\pi} x\frac{\underline{p}(x)}{p(x>\pi)}dx = \frac{3\pi}{2}. \end{aligned}$$

The denominator in the integral inflates the original pdf by a factor of two, and the limits of integration cause only values of x in the range of interest to be used.

3.6 The Gaussian PDF

The normal or Gaussian pdf is one of the most popular distributions for describing random variables, partly because many physical systems do exhibit Gaussian variability, and partly because the Gaussian pdf is amenable to some very powerful tools in design and analysis. It is

$$\underline{p}(x) = \frac{1}{\sigma\sqrt{2\pi}}e^{(x-\bar{x})^2/2\sigma^2},$$

where σ and σ^2 are the standard deviation and variance, respectively, and $\bar{x}$ is the mean value. By design, this pdf always has area one. The cumulative probability function is

$$\begin{aligned} P(x) &= \frac{1}{2} + \text{erf}\left(\frac{x-\bar{x}}{\sigma}\right), \text{ where} \\ \text{erf}(\xi) &= \frac{1}{\sqrt{2\pi}}\int_0^\xi e^{-\xi^2/2}d\xi. \end{aligned}$$

Don't try to compute the *error function* erf(); look it up in a table or call a subroutine! The Guassian distribution has a shorthand: $N(\bar{x}, \sigma^2)$. The arguments are the mean and variance.

3.7 The Cumulative Probability Function

The *cumulative probability function* is closely related to the pdf $\underline{p}(x)$:

$$\begin{aligned} P(x_o) &= p(x \leq x_o) = \int_{-\infty}^{x_o} \underline{p}(x)dx, \text{ so that} \\ \underline{p}(x_o) &= \frac{dP(x_o)}{dx}. \end{aligned}$$

The probability density function is the derivative of the cumulative probability function. P is important because it lets us now transform the complete pdf of a random variable into the pdf of a function of the random variable. Let us say $y = f(x)$; the key idea is that for a monotonic function $f(x)$ (monotonic means the function is either strictly increasing or strictly decreasing with x),

$$p(x \leq x_o) = p(y \leq y_o = f(x_o));$$

these probabilities are the same, although we will see some subtleties to do with multiple values if the function is not monotonic. Here is a first example: let $y = ax + b$. In the case that $a > 0$, then

$$ax + b \leq y_o \text{ when } x \leq \frac{y_o - b}{a} \longrightarrow$$
$$p(y \leq y_o) = \int_{-\infty}^{\frac{y_o - b}{a}} \underline{p}(x)\, dx.$$

The case when $a < 0$ has simply

$$p(y \leq y_o) = 1 - \int_{-\infty}^{\frac{y_o - b}{a}} \underline{p}(x)\, dx.$$

All that we have done here is modify the upper limit of integration, to take the function into account. Now suppose that $y < y_o$ or $y > y_o$ over several disjoint regions of x. This will be the case if $f(x)$ is not monotonic. An example is $y = x^2$, which for a given value of y_o clearly has two corresponding x_o's. We have

$$\begin{array}{rcl} p(y \geq y_o) & = & p(x \leq -\sqrt{y_o}) + p(x \geq \sqrt{y_o}), \text{ or equivalently} \\ p(y \leq y_o) & = & 1 - p(x \leq -\sqrt{y_o}) - p(x \geq \sqrt{y_o}) \end{array}$$

and there is of course no solution if $y_o < 0$. The use of pdf's for making these calculations, first in the case of monotonic $f(x)$, goes like this:

$$\begin{array}{rcl} \underline{p}(y)|dy| & = & \underline{p}(x)|dx|, \text{ so that} \\ \underline{p}(y) & = & \underline{p}(x) / \left| \dfrac{dy}{dx} \right| . \end{array}$$

In the case of non-monotonic $f(x)$, a given value of y corresponds with $x_1, \cdots, x_n$. The correct extension of the above is

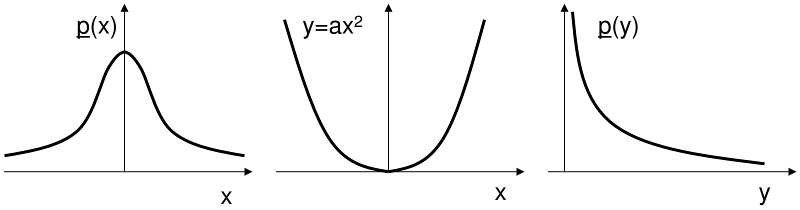
$$\underline{p}(y) = \underline{p}(x_1) / \left| \frac{dy(x_1)}{dx} \right| + \cdots + \underline{p}(x_n) / \left| \frac{dy(x_n)}{dx} \right| .$$

Here is a more detailed example. Consider the Gaussian or normal distribution $N(0, \sigma^2)$:

$$\underline{p}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2},$$

and let $y = ax^2$. For a given (positive) y, there are two solutions for x:

$$x_1 = -\sqrt{\frac{y}{a}}, \quad x_2 = \sqrt{\frac{y}{a}}.$$



Now $dy/dx = 2ax$ so that

$$\begin{aligned}
\left|\frac{dy(x_1)}{dx}\right| &= \left|\frac{dy(x_2)}{dx}\right| = 2a|x| = 2a\sqrt{\frac{y}{a}} = 2\sqrt{ay} \longrightarrow \\
\underline{p}(y) &= \underline{p}(x_1)/\left|\frac{dy(x_1)}{dx_1}\right| + \underline{p}(x_2)/\left|\frac{dy(x_2)}{dx_2}\right| \\
&= \frac{1}{\sigma\sqrt{2\pi}}\left\{\frac{1}{2\sqrt{ay}}e^{-y/2a\sigma^2} + \text{ same }\right\}, \text{ giving finally} \\
&= \frac{1}{\sigma\sqrt{2\pi ay}}e^{-y/2\sigma^2 a}.
\end{aligned}$$

3.8 Central Limit Theorem

A rather amazing property of random variables is captured in the central limit theorem; that a sum of random variables taken from distributions - even many different distributions - approaches a single Gaussian distribution as the number of samples gets large. To make this clear, let x_1 come from a distribution with mean $\bar{x}_1$ and variance σ_1^2, and so on up to x_n, where n is the number of samples. Let $y = \sum_{i=1}^{n} x_i$. As $n \to \infty$,

$$\begin{aligned}
\underline{p}(y) &= N(\bar{y}, \sigma_y^2), \text{ with} \\
\bar{y} &= \sum_{i=1}^{n} \bar{x}_i, \\
\sigma_y^2 &= \sum_{i=1}^{n} \sigma_i^2.
\end{aligned}$$

This is easy to verify numerically, and is at the heart of Monte Carlo simulation techniques. As a practical matter in using the theorem, it is important to remember that as the number of trials goes to infinity so will the variance, even if the mean does not (for example, if the underlying means are all zero). Taking more samples does not mean that the variance of the sum decreases, or even approaches any particular value.

4 RANDOM PROCESSES

From the essential aspects of probability we now move into the time domain, considering random signals. For this, assign to each random event A_i a complete signal, instead of a single scalar: $A_i \longrightarrow x_i(t)$. The set of all the functions that are available (or the menu) is call the *ensemble* of the random process. An example case is to roll a die, generating $i = [1, 2, 3, 4, 5, 6]$ and suppose $x_i(t) = t^i$.

In the general case, there could be infinitely many members in the ensemble, and of course these functions could involve some other variables, for example $x_i(t, y, z)$, where y and z are variables not related to the random event A_i. Any particular $x_i(t)$ can be considered a regular, deterministic function, if the event is known. $x(t_o)$, taken at a specific time but without specification of which event has occurred, is a random variable.

4.1 Time Averages

The theory of random processes is built on two kinds of probability calculations: those taken across time and those taken across the ensemble. For time averages to be taken, we have to consider a specific function, indexed by i:

$$\begin{aligned}
m(x_i(t)) &= \lim_{T\to\infty} \frac{1}{T}\int_0^T x_i(t)dt \text{ (mean)} \\
V^t(x_i(t)) &= \lim_{T\to\infty} \frac{1}{T}\int_0^T [x_i(t) - m(x_i(t))]^2 dt \text{ (variance on time)} \\
R_i^t(\tau) &= \lim_{T\to\infty} \frac{1}{T}\int_0^T [x_i(t) - m(x_i(t))][x_i(t+\tau) - m(x_i(t))]dt \text{ (autocorrelation).}
\end{aligned}$$

The mean and variance have new symbols, but are calculated in a way that is consistent with our prior definitions. The autocorrelation is new and plays a central role in the definition of a spectrum. Notice that is an inner product of the function's deviation from its mean, with a delayed version of the same, such that $R(0) = V^t$.

Consider the roll of a die, and the generation of functions $x_i(t) = a\cos(i\omega_o t)$. We have

$$\begin{aligned}
m(x_i(t)) &= \lim_{T\to\infty} \int_0^T a\cos(i\omega_o t)dt = 0 \\
V^t(x_i(t)) &= \lim_{T\to\infty} \frac{1}{T}\int_0^T a^2\cos^2(i\omega_o t)dt = \frac{a^2}{2} \\
R_i^t(\tau) &= \lim_{T\to\infty} \frac{1}{T}\int_0^T a^2\cos(i\omega_o t)\cos(i\omega_o(t+\tau))dt = \frac{a^2}{2}\cos(i\omega_o\tau).
\end{aligned}$$

In this case, the autocorrelation depends explicitly on the event index i, and has a peak of $a^2/2$ at $i\omega_o\tau = 2\pi k$, where k is an integer. These values for τ are precisely separated by the period of the i'th harmonic in the ensemble. When the functions line up, we get a positive R^t; when they are out of phase, we get a negative R^t.

4.2 Ensemble Averages

The other set of statistics we can compute are across the ensemble, but at a particular time. Set $y_i = x_i(t_o)$ where t_o is a specific time. Then, considering again the six harmonics from above, we have

$$E(y) = \sum p_i y_i = \sum_{i=1}^{6} \frac{1}{6} a \cos(i\omega_o t_o)$$
$$E(y^2) = \sum p_i y_i^2 = \sum_{i=1}^{6} \frac{1}{6} a^2 \cos^2(i\omega_o t_o).$$

We can see from this simple example that in general, time averages (which are independent of time, but dependent on event) and ensemble statistics (which are independent of event, but dependent on time) are not the same. Certainly one could compute ensemble statistics on time averages, and vice versa, but we will not consider such a procedure specifically here.

The ensemble autocorrelation function is now a function of the time and of the delay:

$$\begin{aligned} R(t,\tau) &= E(x(t)x(t+\tau)) \text{ or} \\ R(t,\tau) &= E\left[\{x(t) - E(x(t))\}\{x(t+\tau) - E(x(t+\tau))\}\right]. \end{aligned}$$

The second form here explicitly takes the mean values into account, and can be used when the process has nonzero mean. The two versions are not necessarily equal as written.

4.3 Stationarity

A stationary random process is one whose ensemble statistics do not depend on time. Intuitively, this means that if we were to sample a sequence of processes, at the same time within each process, and compute statistics of this data set, we would find no dependence of the statistics on the time of the samples. Aircraft engine noise is a stationary process in level flight, whereas the sound of live human voices is not. For a stationary process, $m(t) = m$, i.e., the ensemble mean has no dependence on time. The same is true for the other statistics: $V(t) = R(t,0) = V$, and $R(t,\tau) = R(\tau)$. Formally, a stationary process has *all* ensemble statistics independent of time, whereas our case that the mean, variance, and autocorrelation functions are independent of time defines a (weaker) *second-order stationary* process.

Here is an example: $y_i(t) = a\cos(\omega_o t + \theta_i)$, where θ_i is a random variable, distributed uniformly in the range $[0, 2\pi]$. Is this process stationary? We have to show that all three of the ensemble statistics are independent of time:

$$\begin{aligned} E(y(t)) &= \frac{1}{2\pi}\int_o^{2\pi} a\cos(\omega_o t + \theta)d\theta = 0 \\ R(t,\tau) &= E(y(t)y(t+\tau)) \\ &= \frac{1}{2\pi}\int_0^{2\pi} a^2 \cos(\omega_o t + \theta)\cos(\omega_o(t+\tau) + \theta)d\theta \end{aligned}$$

$$\begin{aligned} &= \frac{1}{2}a^2\cos(\omega_o\tau) \\ V(t) &= R(t,0). \end{aligned}$$

Thus the process is second-order stationary.

As noted above, the statistics of a stationary process are not necessarily the same as the time averages. A very simple example of this is a coin toss, in which heads triggers $x_1(t) = 1$ and $x_2(t) = 2$. Clearly the mean on time of $x_1(t)$ is one, but the ensemble mean at any time is $E(x(t_o)) = 1.5$. This difference occurs here even though the process is obviously stationary.

When the ensemble statistics and the time averages are the same, we say that the process is *ergodic*. Continuing our example above, let us calculate now the time averages:

$$\begin{aligned} m(y_i(t)) &= \lim_{T\to\infty}\frac{1}{T}\int_0^T a\cos(\omega_o t+\theta_i)dt \\ &= \lim_{T\to\infty}\frac{1}{T}a\frac{1}{\omega_o}\sin(\omega_o t+\theta_i)|_0^T \\ &= 0; \\ R^t(\tau) &= \lim_{T\to\infty}\frac{1}{T}\int_0^T a^2\cos(\omega_o t+\theta_i)\cos(\omega_o(t+\tau)+\theta_i)dt \\ &= \frac{1}{2}a^2\cos(\omega_o\tau); \\ V^t &= R^t(0) = \frac{a^2}{2}. \end{aligned}$$

So a sinusoid at random phase is an ergodic process. Indeed, this form is a foundation for modeling natural random processes such as ocean waves, atmospheric conditions, and various types of noise. In particular, it can be verified that the construction

$$y(t) = \sum_{n=1}^{N} a_n\cos(\omega_n t+\theta_n),$$

where the θ_n are independently and uniformly distributed in $[0, 2\pi]$, is stationary and ergodic. It has mean zero, and autocorrelation

$$R(\tau) = \sum_{n=1}\frac{a_n^2}{2}\cos(\omega_n\tau).$$

We now make two side notes. Under stationary and ergodic conditions, the autocorrelation function is symmetric on positive and negative τ because we can always write

$$R(\tau) = E(x(t)x(t+\tau)) = E(x(t'-\tau)x(t')), \text{ where } t' = t+\tau.$$

Furthermore, we have the inequality that $R(0) \geq |R(\tau)|$ for any τ. To see this,

$$0 \leq E[(x(t)+x(t+\tau))^2] \quad = \quad E[x(t)^2] + 2E[x(t)x(t+\tau)] + E[x(t+\tau)^2]$$

$$\begin{aligned} &= 2R(0) + 2R(\tau);\ \text{similarly,} \\ 0 \leq E[(x(t) - x(t+\tau))^2] &= E[x(t)^2] - 2E[x(t)x(t+\tau)] + E[x(t+\tau)^2] \\ &= 2R(0) - 2R(\tau). \end{aligned}$$

The only way both of these can be true is if $R(0) \geq |R(\tau)|$.

4.4 The Spectrum: Definition

Given an ergodic process $y(t)$, with mean zero and autocorrelation $R(\tau)$, the power spectral density of $y(t)$, or the spectrum, is the *Fourier transform of the autocorrelation*:

$$\begin{aligned} S(\omega) &= \int_{-\infty}^{\infty} R(\tau) e^{-i\omega\tau} d\tau \\ R(\tau) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) e^{i\omega\tau} d\omega. \end{aligned}$$

The spectrum is a real and even function of frequency ω, because the autocorrelation is real and even. Expanding the above definition,

$$S(\omega) = \int_{-\infty}^{\infty} R(\tau)(\cos\omega\tau - i\sin\omega\tau) d\tau,$$

and clearly only the cosine will create an inner product with $R(\tau)$.

4.5 Wiener-Khinchine Relation

Recall from our discussion of the Fourier transform that convolution in the time domain of the impulse response $h(t)$ and an arbitrary system input $u(t)$, is equivalent to multiplication in the frequency domain of the Fourier transforms. This is a property in particular of linear, time-invariant systems. Now we can make some additional strong statements in the case of random processes.

If $u(t)$ is stationary and ergodic, and the system is LTI, then the output $y(t)$ is also stationary and ergodic. The statistics are related using the spectrum:

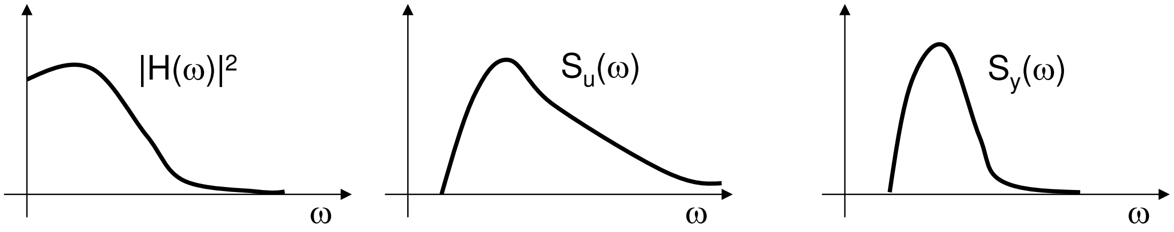
$$S_y(\omega) = |H(\omega)|^2 S_u(\omega).$$

This can be seen as a variant on the transfer function from the Fourier transform. Here, the quantity $|H(\omega)|^2$ transforms the spectrum of the input to the spectrum of the output. It can be used to map the statistical properties of the input (such as an ocean wave field) to statistical properties of the output. In ocean engineering, this is termed the response amplitude operator, or RAO.

To prove this, we will use the convolution property of LTI systems.

$$\begin{aligned}
y(t) &= \int_{-\infty}^{\infty} h(\tau)u(t-\tau)d\tau, \text{ so that} \\
R_y(t,\tau) &= E[y(t)y(t+\tau)], \\
&= E\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(\tau_1)u(t-\tau_1)h(\tau_2)u(t+\tau-\tau_2)d\tau_1 d\tau_2\right\} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\tau_1 d\tau_2 h(\tau_1)h(\tau_2)E[u(t-\tau_1)u(t+\tau-\tau_2)] \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\tau_1 d\tau_2 h(\tau_1)h(\tau_2)R_u(\tau-\tau_2+\tau_1) \\
&\quad \text{(because the input is stationary and ergodic, } R_u \text{ does not depend on time)} \\
S_y(\omega) &= \int_{-\infty}^{\infty} R_y(\tau)e^{-i\omega\tau}d\tau \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\tau d\tau_1 d\tau_2 e^{-i\omega\tau} h(\tau_1)h(\tau_2)R_u(\tau-\tau_2+\tau_1); \text{ now let } \xi = \tau-\tau_2+\tau_1 \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} d\xi d\tau_1 d\tau_2 e^{-i\omega(\xi+\tau_2-\tau_1)} h(\tau_1)h(\tau_2)R_u(\xi) \\
&= \int_{-\infty}^{\infty} d\xi e^{-i\omega\xi}R_u(\xi)\int_{-\infty}^{\infty} e^{i\omega\tau_1}h(\tau_1)\int_{-\infty}^{\infty} d\tau_2 e^{-i\omega\tau_2}h(\tau_2) \\
&= S_u(\omega)H^*(\omega)H(\omega).
\end{aligned}$$

Here we used the $*$-superscript to denote the complex conjugate, and finally we note that $H^*H = |H|^2$.



4.6 Spectrum Interpretation

Let us consider now the stationary and ergodic random process with description:

$$y(t) = \sum_{n=1}^{N} a_n \cos(\omega_n t + \psi_n),$$

where ψ_n is a random variable with uniform distribution in the range $[0, 2\pi]$. As mentioned previously, this process has autocorrelation

$$R(\tau) = \frac{1}{2}\sum_{n=1}^{N} a_n^2 \cos\omega_n\tau;$$

and then

$$S(\omega) = \frac{1}{2}\sum_{n=1}^{N} a_n^2 \pi [\delta(\omega - \omega_n) + \delta(\omega + \omega_n)].$$

As with the Fourier transform, each harmonic in the time domain maps to a pair of delta functions in the frequency domain. However, unlike the Fourier transform, there is no phase angle associated with the spectrum - the two delta functions are both positive and both real.

Further, a real process has infinitely many frequency components, so that the spectrum really become a continuous curve. For example the Bretschneider wave spectrum in ocean engineering is given by

$$S^+(\omega) = \frac{5}{16}\frac{\omega_m^4}{\omega^5} H_{1/3}^2 e^{-5\omega_m^4/4\omega^4}$$

where ω is frequency in radians per second, ω_m is the modal (most likely) frequency of any given wave, and $H_{1/3}$ is the significant wave height. The + superscript on $S(\omega)$ indicates a "one-sided spectrum," wherein all the energy at positive and negative frequencies has been collected into the positive frequencies. We also take into account a factor of $1/2\pi$ (for reasons given below), to make the formal definition

$$\begin{array}{rll} S^+(\omega) = & \frac{1}{\pi} S(\omega), & \text{for } \omega \geq 0, \text{ and} \\ & 0, & \text{for } \omega < 0. \end{array}$$

What is the justification for the factor of $1/2\pi$? Consider that

$$\begin{aligned} R(\tau) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega) e^{i\omega\tau} d\omega \longrightarrow \\ R(0) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} S(\omega) d\omega \\ &= \frac{2}{2\pi}\int_{0}^{\infty} S(\omega) d\omega, \end{aligned}$$

and therefore that

$$\sigma^2 = R(0) = \int_0^{\infty} S^+(\omega) d\omega.$$

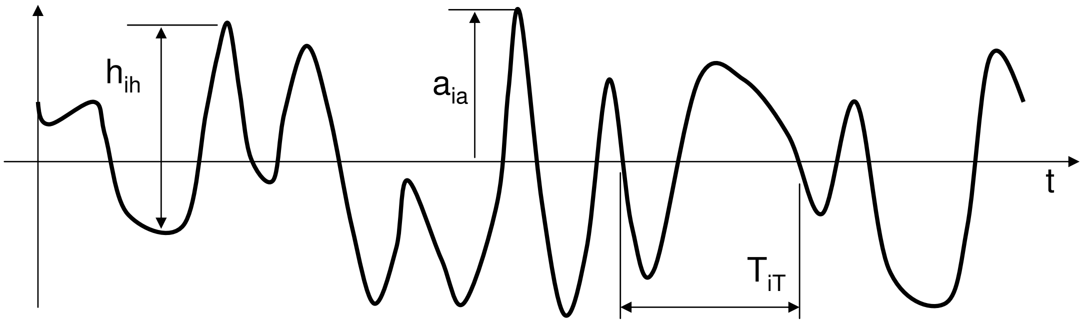
In words, the area under the one-sided spectrum is exactly equal to the variance, or the square of the standard deviation of the process.

5 SHORT-TERM STATISTICS

The spectrum contains information about the magnitude of each frequency component in a stationary and ergodic random process. A summation of harmonic functions with random phase satisfies ergodicity and stationarity, and this will be a dominant model of a random process in our discussion. Also, the central limit theorem provides that a random process of sufficient length and ensemble size has a Gaussian distribution.

The primary calculation is the frequency-domain multiplication of an input spectrum by a transfer function magnitude squared, so as to obtain the output spectrum. Further, a Gaussian input driving an LTI system will create a Gaussian output. Most importantly, the input and output spectra have statistical information that we will be able to integrate into the system design process. In this section, we focus on short-term statistics, namely those which will apply to a random process that is truly stationary. An easy example is a field of ocean waves: over the course of minutes or hours, the process is stationary, but over days the effects of distant storms will change the statistics.

Considering specific "events" within a random process, several of the most important are the amplitude a_{i_a}, the height h_{i_h}, and the period T_{i_T}. The index here is counting through the record the number of amplitude measurements, height measurements, and period measurements. In the figure below, the period is measured specifically between zero downcrossings, and the amplitude is the maximum value reached after an upcrossing and before the next downcrossing. The height goes from the minimum after a zero downcrossing to the maximum after the following zero upcrossing. These definitions have to be applied consistently, because sometimes (as shown) there are fluctuations that do not cross over the zero line.



We will focus on statistics of the amplitude a; the spectrum used below is that of a. Let us define three even moments:

$$\begin{aligned}
M_0 &= \int_0^\infty S^+(\omega)d\omega \\
M_2 &= \int_0^\infty \omega^2 S^+(\omega)d\omega \\
M_4 &= \int_0^\infty \omega^4 S^+(\omega)d\omega.
\end{aligned}$$

We know already that M_0 is related to the variance of the process. Without proof, these are combined into a "bandwidth" parameter that we will use soon:

$$\epsilon^2 = 1 - \frac{M_2^2}{M_o M_4}.$$

Physically, ϵ is near one if there are many local minima and maxima between zero crossings (*broadband*), whereas it is near zero if there is only one maxima after a zero upcrossing before returning to zero (*narrow-band*).

5.1 Central Role of the Gaussian and Rayleigh Distributions

The Central Limit Theorem - which states that the sum of a large number of random variables approaches a Gaussian - ensures that stationary and ergodic processes create a data trace that has its samples normally distributed. For example, if a histogram of the samples from an ocean wave measurement system is plotted, it will indicate a normal distribution. Roughly speaking, in any given cycle, the trace will clearly spend more time near the extremes and less time crossing zero. But for the random process, these peaks are rarely repeated, while the zero is crossed nearly every time. It is recommended to try a numerical experiment to confirm the result of a normal distribution:

$$p(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \mathrm{e}^{-y^2/2\sigma_y^2},$$

where the standard deviation is σ_y and the mean is zero. As indicated above, the standard deviation is precisely the square root of the area under the one-sided spectrum.

In contrast with the continuous trace above, heights are computed only once for each cycle. Heights are defined to be positive only, so there is a lower limit of zero, but there is no upper limit. Just as the signal y itself can theoretically reach arbitrarily high values according to the normal distribution, so can heights. It can be shown that the distribution of heights from a Gaussian process is Rayleigh:

$$p(h) = \frac{h}{4\sigma_y^2} \mathrm{e}^{-h^2/8\sigma_y^2},$$

where σ here is the standard deviation of the *underlying* normal process. The mean and standard deviation of the height itself are different:

$$\begin{aligned} \bar{h} &= \sqrt{2\pi}\sigma_y \simeq 2.5\sigma_y \\ \sigma_h &= \sqrt{8-2\pi}\sigma_y \simeq 1.3\sigma_y. \end{aligned}$$

Notice that the Rayleigh pdf has an exponential of the argument squared, but that this exponential is also multiplied by the argument; this drives the pdf to zero at the origin. The cumulative distribution is the simpler Rayleigh cpf:

$$p(h < h_o) = 1 - \mathrm{e}^{-h_o^2/8\sigma_y^2};$$

$P(h)$ looks like half of a Gaussian pdf, turned upside down! A very useful formula that derives from this simple form is that

$$p(h > h_o) = 1 - p(h < h_o) = e^{-2h_o^2/(\bar{h}^{1/3})^2}.$$

This follows immediately from the cumulative probability function, since $\bar{h}^{1/3} = 4\sigma_y$. It is confirmed that $p(h > \bar{h}^{1/3}) = e^{-2} \simeq 0.13$.

5.2 Frequency of Upcrossings

The first statistic we discuss is the frequency with which the process exceeds a given level; we consider upcrossings only of the positive value A. Now let $\bar{f}(A)$ be the average frequency of upcrossings past A, in upcrossings per second. Then $\bar{f}(0)$ is the average frequency of zero upcrossing, or $1/\bar{T}$, the inverse of the average period, $E(T)$. The formulas are

$$\begin{aligned} \bar{f}(0) &= \frac{1}{2\pi}\sqrt{\frac{M_2}{M_0}} \\ \bar{f}(A) &= \frac{1}{2\pi}\sqrt{\frac{M_2}{M_0}}e^{-A^2/2M_0}. \end{aligned}$$

With M_0 equal to the variance, the exponential here clearly is of the Gaussian form. Here is an example of the use of these equations in design. An fixed ocean platform is exposed to storm waves of standard deviation two meters and average period eight seconds. How high must the deck be to only be flooded every ten minutes, on average?

This problem does not involve any transfer function since the platform is fixed. If it were floating, there would be some motion and we would have to transform the wave spectrum into the motion spectrum. All we have to do here is invert the equation to solve for A, given that $\bar{f}(A) = 1/(60 \times 10)$, $M_0 = 4$ and $\bar{T} = 8$ or $\bar{f}(0) = 1/8$:

$$A = \sqrt{-2M_0 \ln\left(\bar{T}\bar{f}(A)\right)} = 5.87m.$$

This result gives a flavor of how valuable these statistics will be - even though the standard deviation of wave elevation is only two meters, every ten minutes we should expect a six-meter amplitude!

5.3 Maxima At and Above a Given Level

Now we look at the probability of any maximum amplitude a_{ia} reaching or exceeding a given level. We normalize the amplitude with the random process variance, i.e., $\eta = a/\sqrt{M_0}$ and

$\bar{\eta} = A/\sqrt{M_0}$. The results are very useful for calculating extreme loads. First,

$$\begin{aligned} p(\eta = \bar{\eta}) &= \frac{\epsilon}{\sqrt{2\pi}} e^{-\bar{\eta}^2/2\epsilon^2} + \phi(\bar{\eta}q/\epsilon) \frac{\bar{\eta}q}{\sqrt{2\pi}} e^{-\bar{\eta}^2/2} \text{ where} \\ q &= \sqrt{1-\epsilon^2}, \\ \phi(\xi) &= \int_{-\infty}^{\xi} e^{-u^2/2} du \text{ (related to the error function erf).} \end{aligned}$$

With large amplitudes being considered and small ϵ (a narrow-banded process), we can make some approximations to find:

$$\begin{aligned} p(\eta = \bar{\eta}) &\approx \frac{2q}{1+q} \bar{\eta} e^{-\bar{\eta}^2/2} \longrightarrow \\ p(\eta > \bar{\eta}) &\approx \frac{2q}{1+q} e^{-\bar{\eta}^2/2}. \end{aligned}$$

The second relation here is the more useful, as it gives the probability that the (nondimensional) amplitude will *exceed* a given value. It follows directly from the former equation, since (roughly) the cumulative distribution is the derivative of the probability density.

5.4 1/N'th Highest Maxima

We next define a statistic that is the average lowest value of the 1/N'th highest peaks. The construction goes this way: From the record, list all of the highest peaks. Rank and then collect the highest 1/N fraction of these numbers. The average lowest value of this set is the 1/N'th highest maxima. For instance, let the peaks be [6 7 11 5 3 4 8 5 9 4 2 5]. There are twelve numbers here, and the one-third highest maxima $a^{1/3}$ is around 6.5, because [7 11 8 9] exceed it. We use the superscript $1/N$ to denote the 1/N'th highest maxima of a quantity.

Building on the previous section, we have that

$$\begin{aligned} p(a > a^{1/N}) = \frac{1}{N} &\approx \frac{2q}{1+q} \exp\left(-(a^{1/N})^2/2M_0\right), \text{ so that} \\ a^{1/N} &\approx \sqrt{2M_0 \ln\left(\frac{2q}{1+q}N\right)}. \end{aligned}$$

5.5 1/N'th Average Value

We can similarly define a statistic that is the average of the 1/N'th highest peaks. In this case, we are after the average of this collection of 1/N peaks:

$$\begin{aligned} \bar{a}^{1/N} &= E(a|a > a^{1/N}) \\ &= \int_{a^{1/N}}^{\infty} a \, p(a = a_m | a_m > a^{1/N}) da. \end{aligned}$$

Note that we use the dummy variable a_m. We have then the conditional probability

$$p(a = a_m | a_m > a^{1/N}) = \frac{p[(a = a_m) \cap (a_m > a^{1/N})]}{p(a_m > a^{1/N})}.$$

Working in nondimensional form, we have

$$\begin{aligned} \bar{\eta}^{1/N} &= \int_{\eta^{1/N}}^{\infty} \frac{1}{1/N} \eta p(\eta = \eta_m) d\eta \\ &= \frac{2qN}{1+q} \int_{\eta^{1/N}}^{\infty} \eta^2 e^{\eta^2/2} d\eta. \end{aligned}$$

Here are a few explicit results for amplitude and height:

$$\begin{aligned} \bar{a}^{1/3} &= 1.1\sqrt{M_0} \text{ to } 2\sqrt{M_0} \\ \bar{a}^{1/10} &= 1.8\sqrt{M_0} \text{ to } 2.5\sqrt{M_0}. \end{aligned}$$

The amplitudes here vary depending on the parameter ϵ - this point is discussed in the Principles of Naval Architecture, page 20. Here are some 1/N'th average heights:

$$\begin{aligned} \bar{h}^{1/3} &= 4.0\sqrt{M_0} \\ \bar{h}^{1/10} &= 5.1\sqrt{M_0}. \end{aligned}$$

The value $\bar{h}^{1/3}$ is the significant wave height, the most common description of the size of waves. It turns out to be very close to the wave size reported by experienced mariners.

Finally, here are expected highest heights in N observations - which is not quite either of the 1/N'th maximum or the 1/N'th average statistics given above:

$$\begin{aligned} \bar{h}100) &= 6.5\sqrt{M_0} \\ \bar{h}(1000) &= 7.7\sqrt{M_0} \\ \bar{h}10000) &= 8.9\sqrt{M_0}. \end{aligned}$$

5.6 The 100-Year Wave: Estimate from Short-Term Statistics

For long-term installations, it is important to characterize the largest wave to be expected in an extremely large number of cycles. We will make such a calculation here, although as indicated in our discussion of Water Waves, the Rayleigh distribution does not adequately capture extreme events over such time scales. Spectra and the consequent Rayleigh height distribution are *short-term properties only.*

The idea here is to equate $p(h > h_o)$ from the distribution with the definition that in fact $h > h_o$ once in 100 years. Namely, we have

$$p(h > h_{100yr}) = \frac{1}{100\text{years}/\bar{T}} = \mathrm{e}^{-2h_{100yr}^2/\bar{h}^{1/3}},$$

where $\bar{T}$ is the average period. As we will see, uncertainty about what is the proper $\bar{T}$ has little effect in the answer. Looking at the first equality, and setting $\bar{T} = 8$ seconds and $\bar{h}^{1/3} = 2$ meters as example values, leads to

$$\begin{aligned} 2.5 \times 10^{-9} &= \mathrm{e}^{-2h_{100yr}^2/\bar{h}^{1/3}}; \\ \log(2.5 \times 10^{-9}) &= -2h_{100yr}^2/4; \\ h_{100yr} &= 6.3 \text{ meters, or } 3.1\bar{h}^{1/3}. \end{aligned}$$

According to this calculation, the 100-year wave height is approximately three times the significant wave height. Because $\bar{T}$ appears inside the logarithm, an error of twofold in $\bar{T}$ changes the extreme height estimate only by a few percent.

6 WATER WAVES

Surface waves in water are a superb example of a stationary and ergodic random process. The model of waves as a nearly linear superposition of harmonic components, at random phase, is confirmed by measurements at sea, as well as by the linear theory of waves, the subject of this section.

We will skip some elements of fluid mechanics where appropriate, and move quickly to the cases of two-dimensional, inviscid and irrotational flow. These are the major assumptions that enable the linear wave model.

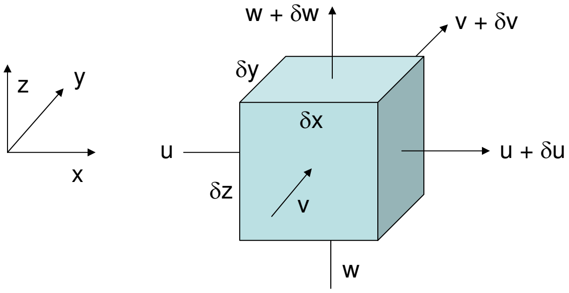
6.1 Constitutive and Governing Relations

First, we know that near the sea surface, water can be considered as incompressible, and that the density ρ is nearly uniform. In this case, a simple form of conservation of mass will hold:

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0,$$

where the Cartesian space is $[x, y, z]$, with respective particle velocity vectors $[u, v, w]$. In words, the above equation says that net flow into a differential volume has to equal net flow out of it. Considering a box of dimensions $[\delta x, \delta y, \delta z]$, we see that any δu across the x-dimension, has to be accounted for by δv and δw:

$$\delta u \delta y \delta z + \delta v \delta x \delta z + \delta w \delta x \delta y = 0.$$



Next, we invoke Newton's law, in the three directions:

$$\begin{aligned}
\rho \left[\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} \right] &= -\frac{\partial p}{\partial x} + \mu \left[\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right]; \\
\rho \left[\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} + w\frac{\partial v}{\partial y} + u\frac{\partial v}{\partial z} \right] &= -\frac{\partial p}{\partial y} + \mu \left[\frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} + \frac{\partial^2 v}{\partial z^2} \right]; \\
\rho \left[\frac{\partial w}{\partial t} + w\frac{\partial w}{\partial x} + u\frac{\partial w}{\partial y} + v\frac{\partial w}{\partial z} \right] &= -\frac{\partial p}{\partial z} + \mu \left[\frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} + \frac{\partial^2 w}{\partial z^2} \right] - \rho g.
\end{aligned}$$

Here the left-hand side of each equation is the acceleration of the fluid particle, as it moves through the differential volume. The terms such as $u\frac{\partial u}{\partial x}$ capture the fact that the force balance is for a moving particle; the chain rule expansion goes like this in the x-direction:

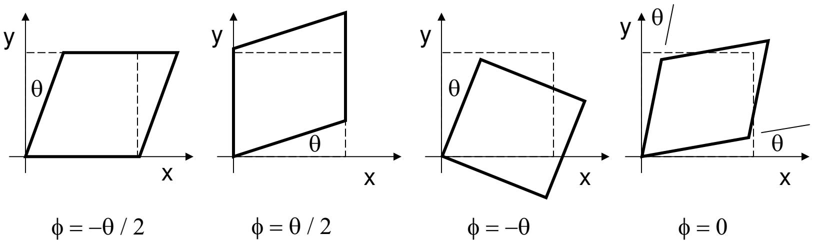
$$\frac{du}{dt} = \frac{\partial u}{\partial t} + \frac{\partial u}{\partial x}\frac{\partial x}{\partial t} + \frac{\partial u}{\partial y}\frac{\partial y}{\partial t} + \frac{\partial u}{\partial z}\frac{\partial z}{\partial t},$$

where $u = \partial x / \partial t$ and so on.

On the right side of the three force balance equations above, the differential pressure clearly acts to slow the particle (hence the negative sign), and viscous friction is applied through absolute viscosity μ. The third equation also has a term for gravity, leading in the case of zero velocities to the familiar relation $p(z) = -\rho g z$, where z is taken positive upward from the mean free surface.

6.2 Rotation and Viscous Effects

In a fluid, unlike for rigid bodies, rotation angle is taken as the average of the angular deflections of the faces. Hence, a net rotation only occurs if the deflection of each face is additive. If they are opposing, then we have only a shearing of the element, with no rotation. Several cases are illustrated below for the two-dimensional case.



Now the rotation rate in the z direction is $\frac{\partial v}{\partial x}$ (the counterclockwise deflection rate of the horizontal face in the plots above), minus $\frac{\partial u}{\partial y}$ (clockwise deflection rate of the vertical face in the plots above). Giving the three dimensional rotation rate vector symbol $\vec{\omega}$, we have

$$\vec{\omega} = \left[\frac{\partial w}{\partial y} - \frac{\partial v}{\partial z} \quad \frac{\partial u}{\partial z} - \frac{\partial w}{\partial x} \quad \frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} \right]^T .$$

Despite this attention, we will now argue that rotational effects are negligible in large water waves. The center of the argument is the fact that a spherical particle will have no rotation except through shear forces. At the same time, however, the Reynolds number in ocean-scale waves has to be taken into account; it is the ratio of inertial to viscous forces

$$Re = \frac{Ud}{\nu},$$

where characteristic speed and length scales are U and d respectively, with ν the kinematic viscosity $(\mu\rho)$. The kinematic viscosity of water at typical ocean temperatures is $1e-6m^2/s$. In contrast, velocities encountered in ocean waves are on the order of $10m/s$, with flow structures on the scale of meters or more. Hence the Reynolds number is very large, and the viscous forces may be neglected. This means in particular that $\vec{\omega}$ is zero and that we will neglect all terms with μ in the force balance.

Note that the inviscid and irrotational assumption is not necessarily valid near solid boundaries, where very small flow structures associated with turbulence result from the no-slip boundary condition.

6.3 Velocity Potential

We introduce the vector field $\phi(\vec{x}, t)$ to satisfy the following relation:

$$\vec{V} = \left\{ \begin{array}{c} u \\ v \\ w \end{array} \right\} = \left[\frac{\partial \phi}{\partial x} \quad \frac{\partial \phi}{\partial y} \quad \frac{\partial \phi}{\partial z} \right]^T = \nabla \phi.$$

The conservation of mass is transformed to

$$\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial^2 \phi}{\partial x^y} + \frac{\partial^2 \phi}{\partial z^2} = \nabla^2 \cdot \phi = 0.$$

Considering Newton's law, the first force balance (x-direction) that we gave above is

$$\rho \left[\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} + w\frac{\partial u}{\partial z} \right] \quad = \quad -\frac{\partial p}{\partial x};$$

this becomes, substituting the velocity potential ϕ,

$$\rho \left[\frac{\partial^2 \phi}{\partial t \partial x} + \frac{\partial \phi}{\partial x}\frac{\partial^2 \phi}{\partial x^2} + \frac{\partial \phi}{\partial y}\frac{\partial^2 \phi}{\partial y \partial x} + \frac{\partial \phi}{\partial z}\frac{\partial^2 \phi}{\partial z \partial x} \right] \quad = \quad -\frac{\partial p}{\partial x}.$$

Integrating on x we find

$$p + \rho\frac{\partial \phi}{\partial t} + \frac{1}{2}\rho\left(u^2 + v^2 + w^2\right) = C,$$

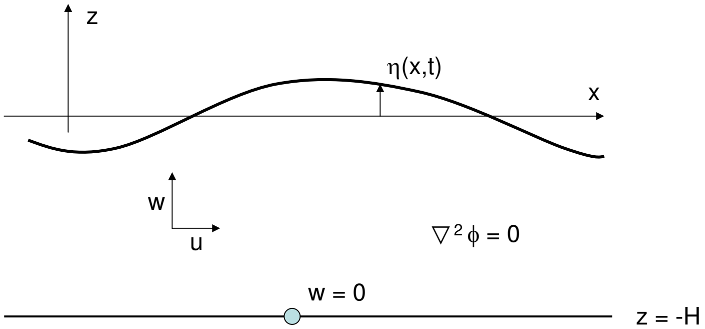
where C is a constant. The other two force balance equations are precisely the same but with the addition of gravity effects in the z-direction. Hence a single equation for the whole field is

$$p + \rho\frac{\partial \phi}{\partial t} + \frac{1}{2}\rho\left(u^2 + v^2 + w^2\right) + \rho g z = C.$$

This is the Bernoulli equation.

6.4 Linear Waves

We consider small amplitude waves in two dimensions x and z, and call the surface deflection $\eta(x,t)$, positive in the positive z direction. Within the fluid, conservation of mass holds, and on the surface we will assume that $p = p_a$, the atmospheric pressure, which we can take to be zero since it is merely an offset to all the pressures under the surface. At the seafloor of course $w = 0$ because flow cannot move into and out of the boundary; at the same time, because of the irrotational flow assumption, the velocity u can be nonzero right to the boundary.



Taking a look at the relative size of terms, we see that $(u^2 + v^2 + w^2)/2$ is much smaller than gz - consider that waves have frequency of one radian per second or so, with characteristic size scale one meter (vertical), whereas g is of course order ten. Hence the Bernoulli equation near the surface simplifies to:

$$\rho \frac{\partial \phi}{\partial t} + \rho g \eta \approx 0 \text{ at } z = 0.$$

Next, we note the simple fact from our definitions that

$$\frac{\partial \eta}{\partial t} \approx \frac{\partial \phi}{\partial z} \text{ at } z = 0.$$

In words, the time rate of change of the surface elevation is the same as the z-derivative of the potential, namely w. Combining these two equations we obtain

$$\frac{\partial^2 \phi}{\partial t^2} + g \frac{\partial \phi}{\partial z} = 0 \text{ at } z = 0.$$

The solution for the surface is a traveling wave

$$\eta(x,t) = a \cos(\omega t - kx + \psi),$$

where a is amplitude, ω is the frequency, k is the wavenumber (see below), and ψ is a random phase angle. The traveling wave has speed ω/k. The corresponding candidate potential is

$$\begin{aligned} \phi(x,z,t) &= -\frac{a\omega}{k} \frac{\cosh(k(z+H))}{\sinh(kH)} \sin(\omega t - kx + \psi), \text{ where} \\ \omega &= 2\pi/T = \sqrt{kg \tanh(kH)} \text{ (dispersion)}, \\ k &= 2\pi/\lambda. \end{aligned}$$

Here λ is the wavelength, the horizontal extent between crests. Let us confirm that this potential satisfies the requirements. First, does it solve Bernoulli's equation at $z = 0$?

$$\begin{aligned}
\frac{\partial^2\phi}{\partial t^2} &= \frac{a\omega^3}{k}\frac{1}{\tanh kH}\sin(\omega t - kx + \psi) \\
&= a\omega g\sin(\omega t - kx + \psi) \text{ and} \\
\frac{\partial\phi}{\partial z} &= -a\omega\sin(\omega t - kx + \psi).
\end{aligned}$$

Clearly Bernoulli's equation at the surface is satisfied. Working with the various definitions, we have further

$$\begin{aligned}
u(x, z, t) &= \frac{\partial\phi}{\partial x} = a\omega\frac{\cosh(k(z + H))}{\sinh(kH)}\cos(\omega t - kx + \psi), \\
w(x, z, t) &= \frac{\partial\phi}{\partial z} = -a\omega\frac{\sinh(k(z + H))}{\sinh(kH)}\sin(\omega t - kx + \psi) \\
p(x, z, t) &\approx -\rho\frac{\partial\phi}{\partial t} - \rho g z \\
&= \rho\frac{a\omega^2}{k}\frac{\cosh(k(z + H))}{\sinh(kH)}\cos(\omega t - kx + \psi) - \rho g z.
\end{aligned}$$

At the surface, $z = 0$, it is clear that the hyperbolic sines in $w(x, z, t)$ cancel. Then taking an integral on time easily recovers the expression given above for surface deflection $\eta(x, t)$. The pressure here is $\rho g\eta$, as would be expected. At depth $z = -H$, $w = 0$ because $\sinh(0) = 0$, thus meeting the bottom boundary condition. The particle trajectories in the x-direction and the z-direction are respectively

$$\begin{aligned}
\xi_p(x, z, t) &= a\frac{\cosh(k(z + H))}{\sinh(kH)}\sin(\omega t - kx + \psi) \\
\eta_p(x, z, t) &= \frac{a}{k}\frac{\cosh(k(z + H))}{\sinh(kH)}\cos(\omega t - kx + \psi).
\end{aligned}$$

Hence the particles' motions take the form of ellipses, clockwise when the wave is moving in the positive x direction.

Note that there are no nonlinear terms in $[x, y, z, u, v, w, p, \phi]$ in any of these equations, and hence this model for waves is linear. In particular, this means that waves of different frequencies and phases can be superimposed, without changing the behavior of the independent waves.

6.5 Deepwater Waves

In the limit that $H \longrightarrow \infty$, the above equations simplify because

$$\phi(x, z, t) \longrightarrow -\frac{a\omega}{k}e^{kz}\sin(\omega t - kx + \psi).$$

We find that

$$\begin{aligned}
\omega^2 &= kg \text{ (dispersion)} \\
p &= \rho g a e^{kz} \cos(\omega t - kx + \psi) - \rho g z; \\
u &= a\omega e^{kz} \cos(\omega t - kx + \psi); \\
w &= -a\omega e^{kz} \sin(\omega t - kx + \psi); \\
\xi_p &= a e^{kz} \sin(\omega t - kx + \psi); \\
\eta_p &= a e^{kz} \cos(\omega t - kx + \psi).
\end{aligned}$$

The dynamic part of the pressure undergoes an exponential decay in amplitude with depth. This is governed by the wave number k, so that the dynamic pressure is quite low below even one-half wavelength in depth: the factor is $e^{-\pi} \approx 0.05$. Particle motions become circular for the deepwater case. The radii of the circles also decay exponentially with depth.

6.6 Wave Loading of Stationary and Moving Bodies

The elegance of the linear wave theory permits explicit estimation of wave loads on structures, usually providing reasonable first approximations. We break the forces on the body into three classes:

1. The dynamic pressure load integrated over the body surface, with the assumption that the presence of the body does not affect the flow - it is a "ghost" body. We call this the incident wave force.
2. The flow is deflected from its course because of the presence of the body; assuming here that the body is stationary. This is the diffraction wave force.
3. Forces are created on the body by its moving relative still water. This is wavemaking due to the body pushing fluid out of the way. We call this the radiation wave force.

This separation of effects clearly depends on linearizing assumptions. Namely, the moving flow interacts with a stationary body in the incident wave and diffraction forces, whereas the stationary flow interacts with a moving body in the radiation force. Further, among the first two forces, we decompose into a part that is unaffected by the "ghost" body and a part that exists only because of the body's presence.

Without proof, we will state simple formulas for the diffraction and radiation loads, and then go into more detail on the incident wave (pressure) force.

As a prerequisite, we need the concept of added mass: it can be thought of as the fluid mass that goes along with a body when it is accelerated or decelerated. Forces due to added mass will be seen most clearly in experiments under conditions when the body has a low instantaneous speed, and separation drag forces are minimal. The added mass of various two-dimensional sections and three-dimensional shapes can be looked up in tables. As one simple example, the added mass of a long cylinder exposed to crossflow is precisely the mass of the displaced water: $A_m = \pi r^2 \rho$ (per unit length).

A very interesting and important aspect of added mass is its connection with the Archimedes force. We observe that the added mass force on a body accelerating in a still fluid is only one-half that which is seen on a stationary body in an accelerating flow. Why is this? In the case of the accelerating fluid, and regardless of the body shape or size, there must be a pressure gradient in the direction of the acceleration - otherwise the fluid would not accelerate. This non-uniform pressure field integrated over the body will lead to a force. This is entirely equivalent to the Archimedes explanation of why, in a gravitational field, objects float in a fluid. This effect is not at all present if the body is accelerating in a fluid having no pressure gradient. The "mass" that explains this Archimedes force as an inertial effect is in fact the same as the added mass, and hence the factor of two.

For the development of a simple model, we will focus on a body moving in the vertical direction; we term the vertical motion $\xi(t)$, and it is centered at $x = 0$. The vertical wave elevation is $\eta(t, x)$, and the vertical wave velocity is $w(t, x, z)$. The body has beam $2b$ and draft T; its added mass in the vertical direction is taken as A_m. The objective is to write an equation of the form

$$m\xi_t t + C\xi = F_I + F_D + F_R,$$

where m is the material (sectional) mass of the vessel, and C is the hydrostatic stiffness, the product of ρg and the waterplane area: $C = 2b\rho g$.

The diffraction force is

$$F_D(t) = A_m w_t(t, x = 0, z = -T/2).$$

In words, this force pushes the body upward when the wave is accelerating upward. Note the wave velocity is referenced at the center of the body. This is the effect of the accelerating flow encountering a fixed body - but does *not* include the Archimedes force. The Archimedes force is derived from the dynamic pressure in the fluid independent of the body, and captured in the incident wave force below. The radiation force is

$$F_R(t) = -A_m \xi_{tt}.$$

This force pulls the body downward when it is accelerating upward; it is the effect of the body accelerating through still fluid. Clearly there is no net force when the acceleration of the wave is matched by the acceleration of the body: $F_D + F_R = 0$.

Now we describe the incident wave force using the available descriptions from the linear wave theory:

$$\begin{aligned} \eta(t, x) &= a\cos(\omega t - kx + \psi) \text{ and} \\ p(t, x, z) &= \rho g a e^{kz} \cos(\omega t - kx + \psi) - \rho g z. \end{aligned}$$

We will neglect the random angle ψ and the hydrostatic pressure $-\rho g z$ in our discussion. The task is to integrate the pressure force on the bottom of the structure:

$$\begin{aligned} F_I &= \int_{-b}^{b} p(t, x, z = -T) dx \\ &= \rho a g \mathrm{e}^{-kT} \int_{-b}^{b} \cos(\omega t - kx) dx \\ &= \frac{2\rho a g}{k} \mathrm{e}^{-kT} \cos(\omega t) \sin(kb). \end{aligned}$$

As expected, the force varies as $\cos(\omega t)$. The effect of spatial variation in the x-direction is captured in the $\sin(kb)$ term.

If $kb < 0.6$ or so, then $\sin(kb) \approx kb$. This is the case that b is about one-tenth of the wavelength or less, and quite common for smaller vessels in beam seas. Further, $\mathrm{e}^{-kT} \approx 1 - kT$ if $kT < 0.3$ or so. This is true if the draft is less than about one twentieth of the wavelength, also quite common. Under these conditions, we can rewrite F_I:

$$\begin{aligned} F_I &\approx 2\rho g a(1 - kT) b \cos\omega t \\ &= 2b\rho g a \cos\omega t - 2bT\rho\omega^2 a \cos\omega t \\ &= C\eta(t, x = 0) + \nabla\rho w_t(t, x = 0, z = 0). \end{aligned}$$

Here ∇ is the (sectional) volume of the vessel. Note that to obtain the second line we used the deepwater dispersion relation $\omega^2 = kg$.

We can now assemble the complete equation of motion:

$$\begin{aligned} m\xi_{tt} + C\xi &= F_I + F_D + F_R \\ &= C\eta(t, x = 0) + \nabla\rho w_t(t, x = 0, z = 0) + A_m w_t(t, x = 0, z = -T/2) - \\ &\quad A_m \xi_{tt}, \text{ so that} \\ (m + A_m)\xi_{tt} + C\xi &\approx C\eta(t, x = 0) + (\nabla\rho + A_m) w_t(t, x = 0, z = -T/2). \end{aligned}$$

Note that in the last line we have equated the z-locations at which the fluid acceleration w_t is taken, to $z = -T/2$. It may seem arbitrary at first, but if we chose the alternative of $w_t(t, x = 0, z = 0)$, we would obtain

$$\begin{aligned} (m + A_m)\xi_{tt} + C\xi &= C\eta(t, x = 0) + (\nabla\rho + A_m) w_t(t, x = 0, z = 0) \\ (-(m + A_m)\omega^2 + C)\xi &= (C - (\nabla\rho + A_m)\omega^2)\eta(t, x = 0) \longrightarrow \\ \frac{\xi(j\omega)}{\eta(j\omega)} &= 1, \end{aligned}$$

since m is equal to $\nabla\rho$ for a neutrally buoyant body. Clearly the transfer function relating vehicle heave motion to wave elevation cannot be unity - the vessel does not follow all waves equally! If we say that $w_t(t, x = 0, z = -T/2) = \gamma w_t(t, x = 0, z = 0)$, where $\gamma < 1$ is a function of the wavelength and T, the above becomes more suitable:

$$\begin{aligned}(-(m + A_m)\omega^2 + C)\xi &= (C - (\gamma \nabla \rho + A_m)\omega^2)\eta(t, x = 0) \longrightarrow \\ \frac{\xi(j\omega)}{\eta(j\omega)} &= \frac{C - (\gamma \nabla \rho + A_m)\omega^2}{C - (m + A_m)\omega^2}\end{aligned}$$

This transfer function has unity gain at low frequencies and gain $(\gamma \nabla \rho + A_m)/(m + A_m)$ at high frequencies. It has zero magnitude at $\omega = \sqrt{C/(\gamma \nabla \rho + A_m)}$, but very high magnitude (resonance) at $\omega = \sqrt{C/(m + A_m}$. The zero occurs at a higher frequency than the resonance because $\gamma < 1$.

In practice, the approximation that w_t should be taken at $z = -T/2$ is reasonable. However, one significant factor missing from our analysis is damping, which depends strongly on the specific shape of the hull. Bilge keels and sharp corners cause damping, as does the creation of radiated waves.

6.7 Limits of the Linear Theory

The approximations made inevitably affect the accuracy of the linear wave model. Here are some considerations. The ratio of wave height to wavelength is typically in the range of 0.02-0.05; a one-meter wave in sea state 3 has wavelength on the order of fifty meters. When this ratio approaches 1/7, the wave is likely to break. Needless to say, at this point the wave is becoming nonlinear! Ultimately, however, even smaller ocean waves interact with each other in a nonlinear way. There are effects of bottom topography, wind, and currents. Nonlinear interaction causes grouping (e.g., rogue waves), and affects the propagation and directionality of waves. It is impossible to make a forward prediction in time - even with perfect and dense measurements - of a wave field, unless these effects are included. In fact, there is some recent work by Yue's group at MIT on the large-scale prediction problem.

6.8 Characteristics of Real Ocean Waves

The origin of almost all ocean waves is wind. Tides and tsunamis also count as waves, but of course at different frequencies. Sustained wind builds waves bigger in amplitude and longer in wavelength - hence their frequency decreases. Waves need sufficient physical space, called the fetch, to fully develop. When the wind stops (or the wave moves out of a windy area), the amplitude slowly decays, with characteristic time $\tau = g^2/2\nu\omega^4$. This rule says that low-frequency waves last for a very long time!

The spectra of ocean waves are reasonably modeled by the standard forms, including JONSWAP, Pierson-Moskowitz, Ochi, and Bretschneider; these have different assumptions and different applications. The conditions of building seas and decaying seas (swell) are different; in the former case, the spectrum is quite wide whereas it may be narrow for the latter.

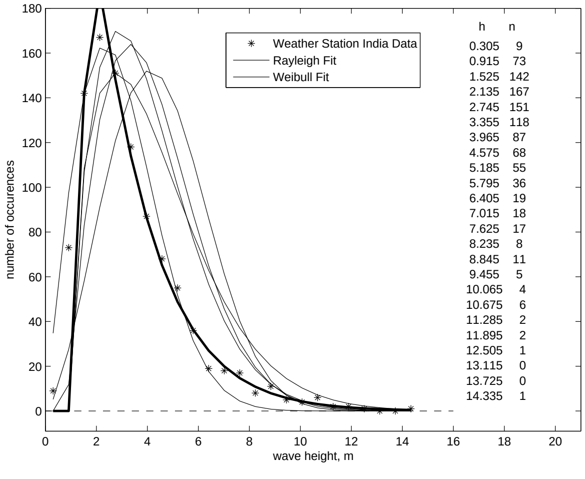
Further details can be found in subject texts, including the Principles of Naval Architecture (E.V., Lewis, ed. SNAME, 1989).

Most important from a design point of view, it has been observed that extreme events do NOT follow the Rayleigh distribution - they are more common. Such dangers are well documented in data on a broad variety of processes including weather, ocean waves, and some social systems. In the case of ocean waves, nonlinear effects play a prominent role, but a second factor which has to be considered for long-term calculations is storms. In periods of many years, intense storms are increasingly likely to occur, and these create short-term extreme seas that may not be well characterized at all in the sense of a spectrum. For the purpose of describing such processes, the Weibull distribution affords some freedom in shaping the ”tail.” The Weibull cpf and pdf are respectively:

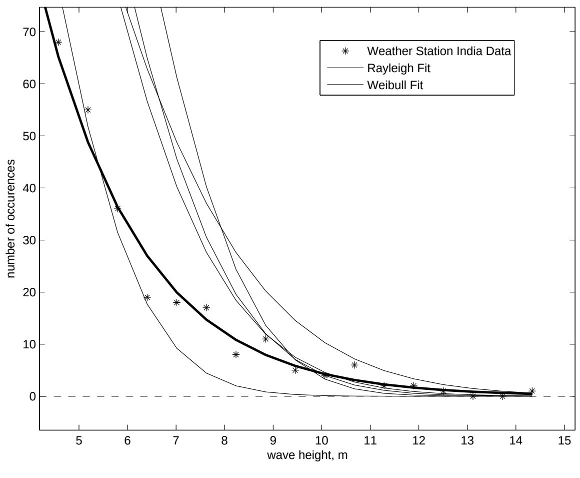
$$\begin{aligned} P(h < h_o) &= 1 - \mathrm{e}^{-(x-\mu)^c/b^c}; \\ p(h) &= \frac{c(x-\mu)^{c-1}}{b^c}\mathrm{e}^{-(x-\mu)^c/b^c}. \end{aligned}$$

It is the choice of c to be a (real) number other than two which makes the Weibull a more general case of the Rayleigh distribution. b is a measure related to the standard deviation, and μ is an offset applied to the argument, giving further flexibility in shaping. Clearly $x > \mu$ is required if c is non-integer, and so μ takes the role of a lower limit to the argument. No observations of h below μ are accounted for in this description.

Here is a brief example to illustrate. Data from Weather Station India was published in 1964 and 1967 (see *Principles of Naval Architecture*), giving a list of observed wave heights taken over a long period. The significant wave height in the long-term record is about five meters, and the average period is about ten seconds. But the distribution is decidedly non-Rayleigh, as shown in the right figure below. Several trial Weibull pdf's are shown, along with an optimal (weighted least-squares) fit in the bold line. The right figure is a zoom of the left, in the tail region.



h	n
0.305	9
0.915	73
1.525	142
2.135	167
2.745	151
3.355	118
3.965	87
4.575	68
5.185	55
5.795	36
6.405	19
7.015	18
7.625	17
8.235	8
8.845	11
9.455	5
10.065	4
10.675	6
11.285	2
11.895	2
12.505	1
13.115	0
13.725	0
14.335	1



Armed with this distribution, we can make the calculation from the cpf that the 100-year wave is approximately 37 meters, or $7.5\bar{h}^{1/3}$. This is a very significant amplification, compared to the factor of three predicted using short-term statistics in Section 5.6, and reinforces the importance of observing and modeling accurately real extreme events.

7 OPTIMIZATION

The engineer is continually faced with non-trivial decisions, and discerning the best among alternatives is one of the most useful and general tasks that one can master. Optimization exists because in nearly every endeavor, one is faced with tradeoffs. Here are some examples:

- Contributing to savings versus achieving enjoyment from purchases made now;
- Buying an expensive bicycle from one of many manufacturers - you are faced with choices on accessories, weight, style, warranty, performance, reputation, and so on;
- Writing a very simple piece of code that can solves a particular problem versus developing a more professional and general-use product;
- Size of the column to support a roof load;
- How fast to drive on the highway;
- Design of strength bulkheads inside an airplane wing assembly

The field of optimization is very broad and rich, with literally hundreds of different classes of problems, and many more methods of solution. Central to the subject is the concept of the *parameter space* denoted as X, which describes the region where specific decisions x may lie. For instance, acceptable models of a product off the shelf might be simply indexed as x_i. x can also be a vector of specific or continuous variables, or a mixture of the two. Also critical is the concept of a *cost* $f(x)$ that is associated with a particular parameter set x. We can say that f will be minimized at the optimal set of parameters x^*:

$$f(x^*) = \min_{x \epsilon X} f(x).$$

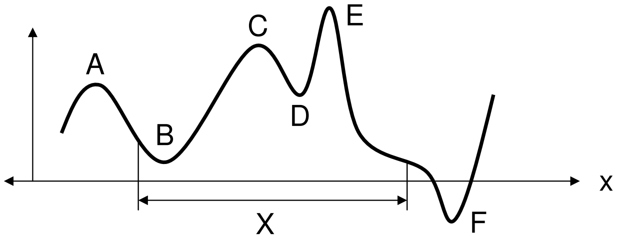
We will develop in this section some methods for continuous parameters and others for discrete parameters. We will consider some concepts also from planning and multi-objective optimization, e.g., the case where there is more than one cost function.

7.1 Single-Dimension Continuous Optimization

Consider the case of only one parameter, and one cost function. When the function is known and is continuous - as in many engineering applications - a very reasonable first method to try is to zero the derivative. In particular,

$$\left[\frac{df(x)}{dx} \right]_{x=x*} = 0.$$

The user has to be aware even in this first problem that there can exist multiple points with zero derivative. These are any locations in X where $f(x)$ is flat, and indeed these could be at local minima and at local and global maxima. Another important point to note is that if X is a finite domain, then there may be *no* location where the function is flat. In this case, the solution could lie along the boundary of X, or take the minimum within X. In the figure below, points A and C are local maxima, E is the global maxima, B and D are local minima, and F is the global minimum shown. However, the solution domain X does not admit F, so the best solution would be B. In all the cases shown, however, we have at the maxima and minima $f'(x) = 0$. Furthermore, at maxima $f''(x) < 0$, and at minima $f''(x) > 0$.



We say that a function $f(x)$ is *convex* if and only if it has everywhere a nonnegative second derivative, such as $f(x) = x^2$. For a convex function, it should be evident that any minimum is in fact a global minimum. (A function is concave if $-f(x)$ is convex.) Another important fact of convex functions is that the graph always lies above any points on a tangent line, defined by the slope at point of interest:

$$f(x + \delta x) \geq f(x) + f'(x)\delta$$

When δ is near zero, the two sides are approximately equal.

Suppose that the derivative-zeroing value x^* cannot be deduced directly from the derivative. We need another way to move toward the minimum. Here is one approach: move in the downhill direction by a small amount that scales with the inverse of the derivative. Letting $\delta = -\gamma/f'(x)$, makes the above equation

$$f(x + \delta) \approx f(x) - \gamma$$

The idea is to take small steps downhill, e.g., $x_{k+1} = x_k + \delta$ where k indicates the k'th guess, and this algorithm is usually just called a gradient method, or something similarly nondescript! While it is robust, one difficulty with this algorithm is how to stop, because it tries to move a constant decrement in f at each step. It will be unable to do so near a flat minimum, although one can of course to modify γ on the fly to improve convergence.

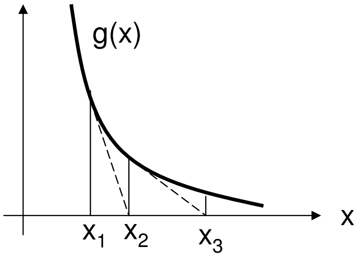
As another method, pose a new problem in which $g(x) = f'(x)$, and we now have to solve $g(x^*) = 0$. Finding the zero of a function is an important problem on its own. Now the convexity inequality above resembles the Taylor series expansion, which is rewritten here in full form:

$$g(x + \delta) = g(x) + g'(x)\delta + \frac{1}{2!}g''(x)\delta^2 + \frac{1}{2!}g'''(x)\delta^3 + \cdots$$

The expansion is theoretically true for any x and any δ (if the function is continuous), and so clearly $g(x+\delta)$ can be at least approximated by the first two terms on the right-hand side. If it is desired to set $g(x+\delta)=0$, then we will have the estimate

$$\delta = -g(x)/g'(x).$$

This is Newton's first-order method for finding the zero of a function. The idea is to shoot down the tangent line to its own zero crossing, and take the new point as the next guess. As shown in the figure, the guesses could go wildly astray if the function is too flat near the zero crossing.



Let us view this another way. Going back to the function f and the Taylor series approximation

$$f(x+\delta) \approx f(x) + f'(x)\delta + \frac{1}{2!}f''(x)\delta^2,$$

we can set to zero the left-hand side derivative with respect to δ, to obtain

$$\begin{aligned} 0 &\approx 0 + f'(x) + f''(x)\delta \longrightarrow \\ \delta &= -f'(x)/f''(x). \end{aligned}$$

This is the same as Newton's method above since $f'(x) = g(x)$. It clearly employs both first and second derivative information, and will hit the minimum in one shot(!) if the function truly is quadratic and the derivatives $f'(x)$ and $f''(x)$ are accurate. In the next section dealing with multiple dimensions, we will develop an analogous and powerful form of this method. Also in the next section, we refer to a line search, which is merely a one-dimensional minimization *along a particular direction*, using a (unspecified) one-dimensional method - such as Newton's method applied to the derivative.

7.2 Multi-Dimensional Continuous Optimization

Now we consider that the cost function f is a function of more than one variable. X is a multi-dimensional space, and x is a vector. We consider again continuous functions. At the minima, as for the single-dimension case, we have certain conditions for the first and second derivatives:

$$\begin{aligned}
\nabla f(x*) &= [f_{x_1}, f_{x_2}, \cdots]_{x=x^*} = [0, 0, \cdots] \\
\nabla^2 f(x) &= \left[\begin{array}{ccc} f_{x_1x_1} & f_{x_1x_2} & \\ f_{x_2x_1} & f_{x_2x_2} & \\ & & \cdots \end{array} \right]_{x=x^*} > 0,
\end{aligned}$$

where the notation that a matrix is greater than zero denotes that it is positive definite. We will present three practical methods for finding the minimum. First is the method of steepest descent. The idea here is to find the downhill direction and take a step δ that way:

$$\begin{aligned}
e &= -\nabla f(x)/||\nabla f(x)|| \\
\delta &= \gamma e.
\end{aligned}$$

Note that, as written, this is a different algorithm than the first method given in one dimension, because here the direction vector e is normalized, and hence the magnitude of the step in x is the same no matter what the steepness of the function. We note also that there exists a value α such that $x + \alpha e$ is the minimizing argument of f, along the e direction and passing through the point x. This is the result of the so-called line search, and a reasonable steepest descent procedure is to perform iteratively a two-step procedure of a gradient calculation followed by a line search in the downhill direction.

The performance of successive line searches can be quite good, or very poor. Poor performance results in some cases because the successive downhill directions are constrained to be orthogonal to each other. This has to be the case because at the minimum on the line, the gradient in the direction of the line is zero by definition. In fact none of these downhill directions may actually point to the minimum, and so many pointless steps might be taken. A solution to this is the conjugate gradient method, wherein we make a useful modification to the downhill directions used for each of the line searches.

We will call the downhill direction vector corresponding with the k'th guess d_k. Letting $g(x) = \nabla f(x)$ and $d_0 = -g(x_0)$, we will let $d_{k+1} = -g(x_{k+1}) + \beta d_k$; this says that we will *deflect* the next search direction from the downhill direction by the term βd_k; the scalar factor β is given by

$$\beta = \frac{g(x_{k+1})^T g(x_{k+1})}{g(x_k)^T g(x_k)}$$

Note here that d is not normalized. The algebra needed to derive this rule is not difficult, and a simple example will illustrate that it is a very powerful one.

Finally, we mention Newton's second-order method in multiple dimensions, using the second derivative. It comes from the multivariable version of the Taylor series expansion:

$$f(x + \delta) \approx f(x) + \nabla f(x)\delta + \frac{1}{2}\delta^T \nabla^2 f(x)\delta.$$

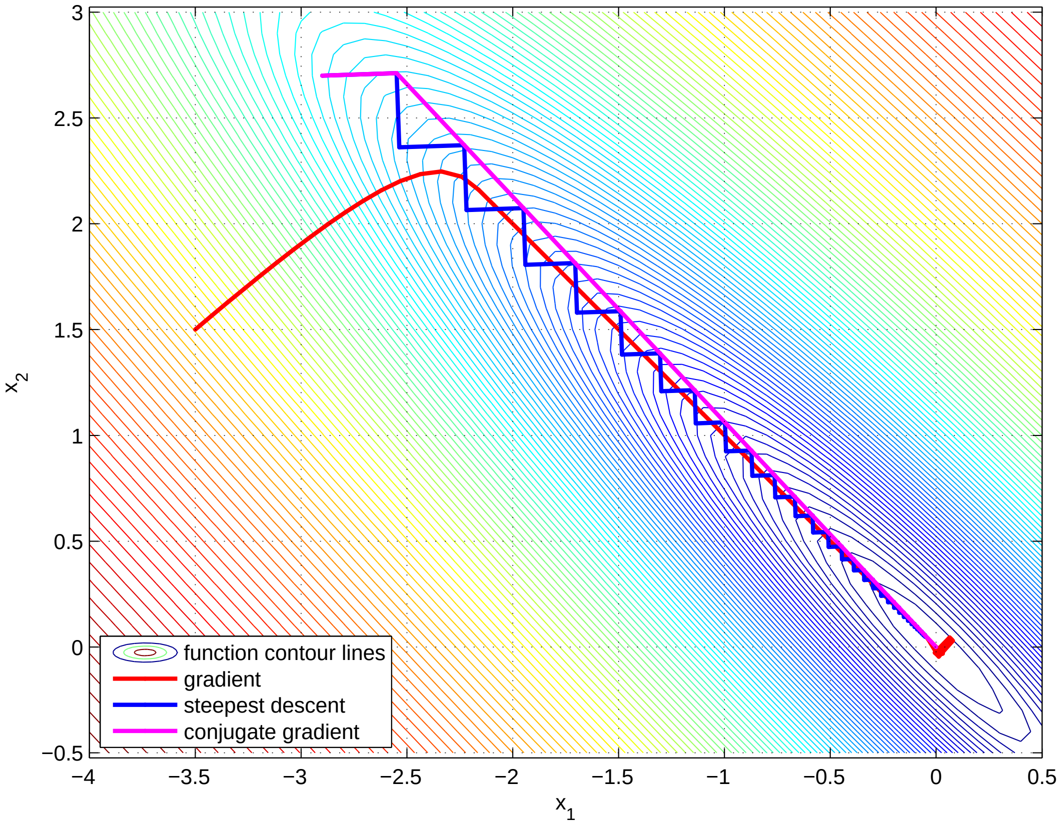
Following from the one-dimensional case, we try to select δ so as to cause $\partial f(x+\delta)/\partial\delta = 0$. This gives

$$\begin{aligned} -\nabla f(x) &= \delta^T \nabla^2 f(x) \longrightarrow \\ \delta &= -[\nabla^2 f(x)]^{-1} \nabla f(x). \end{aligned}$$

In words, Newton's method takes the first and second-derivative information and tries to come up with a specific solution in one step. The algorithm has extremely good convergence properties and is recommended when second derivatives are available.

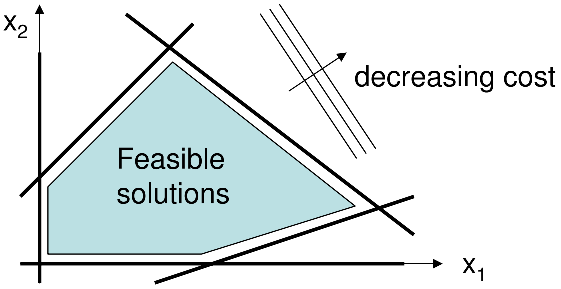
It is important to understand that both the conjugate gradient method and Newton's second-order method get the exact answer in two (conjugate gradient) or one (Newton) tries, when in fact the function is quadratic. Thinking of computational cost, the conjugate gradient algorithm has to have the derivative vector at two points, whereas Newton has to have the gradient plus the Hessian matrix, at the starting point. If these derivatives have to be created numerically and with accuracy, it is clear that the Hessian could be quite expensive. For this reason, the conjugate gradient method may be preferred.

For non-quadratic forms, these algorithms will both do the best they can to approximate and it will take some additional trials. An elegant combination of gradient and Newton methods is found in the Levenberg-Marquardt algorithm.



7.3 Linear Programming

We now consider the case that the cost is a linear function of n parameters. There is clearly no solution unless the parameter space is constrained, and indeed the solution is guaranteed to be on the boundary. The situation is well illustrated in two dimensions ($n = 2$), an example of which is shown below. Here, five linear inequality boundaries are shown; no x are allowed outside of the feasible region. In the general case, both equality and inequality constraints may be present.



The nature of the problem - all linear, and comprising inequality and possibly equality constraints - admits special and powerful algorithms that are effective even in very high dimensions, e.g., thousands. In lower dimensions, we can appeal to intuition gained from the figure to construct a simpler method for small systems, say up to ten unknowns.

Foremost, it is clear from the figure that the solution has to lie on one of the boundaries. The solution in fact lies at a *vertex* of n hypersurfaces of dimension $n-1$. Such a hypersurface is a line in two dimensional space, a plane in three-space, and so on. We will say that a line in three-space is an intersection of two planes, and hence is equivalent to two hypersurfaces of dimension two. It can be verified that a hypersurface of dimension $n-1$ is defined with one equation.

If there are no equality constraints, then these n hypersurfaces forming the solution vertex are a subset of the I inequality constraints. We will generally have $I > n$. If there are also $E < n$ equality constraints, then the solution lies at the intersection of these E equality hypersurfaces and $n-E$ other hypersurfaces taken from the I inequality constraints. Of course, if $E = n$, then we have only a linear system to solve. Thus we have a combinatorics problem; consider the case of inequalities only, and then the mixed case.

- I **inequalities, no equalities.** n of the inequalities will define the solution vertex. The number of combinations of n constraint equations among I choices is $I!/(I-n)!n!$. Algorithm: For each combination (indexed k, say) in turn, solve the linear system of equations to find a solution x_k. Check that the solution does not violate any of the other $I-n$ inequality constraints. Of all the solutions that are feasible (that is, they do not violate any constraints), pick the best one - it is optimal.

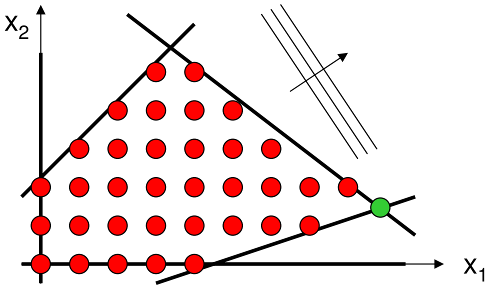
- I **inequalities,** E **equalities.** The solution involves all the equalities, and $n-E$ inequality constraints. The number of combinations of $n-E$ constraint equations among I choices is $I!/(I-n+E)!(n-E)!$. Algorithm: For each combination (indexed with k) in turn, solve the linear set of equations, to give a candidate solution x_k. Check that the solution does not violate any of the remaining $I-n+E$ inequality constraints. Of all the feasible solutions, pick the best one - it is optimal.

The above rough recipe assumes that none of the hypersurfaces are parallel; parallel constraints will not intersect and no solution exists to the linear set of equations. Luckily such cases can be detected easily (e.g., by checking for singularity of the matrix), and classified as infeasible. In the above figure, $I = 5$, and $n = 2$. Hence there are $5!/(5-2)!2! = 10$ combinations to try: AB, AC, AD, AE, BC, BD, BE, CD, CE, DE. Only five are evident, because some of the intersections are outside of the area indicated (Can you find them all?).

The linear programming approach is extremely valuable in many areas of policy and finance, where costs scale linearly with quantity, and inequalities are commonplace. There are also a great many engineering applications, because of the prevalence of linear analysis.

7.4 Integer Linear Programming

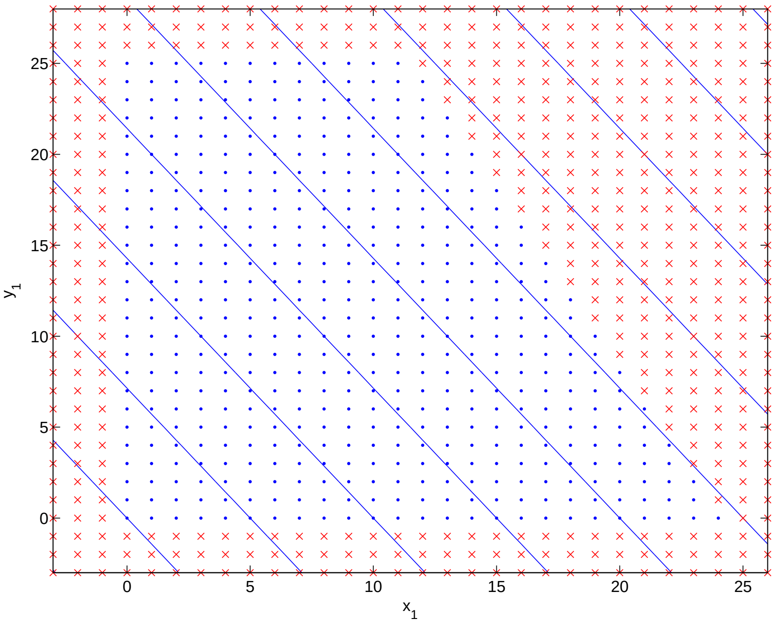
Sometimes the constraint and cost functions are continuous, but only integer solutions are allowed. Such is the case in commodities markets, where it is expected that one will deal in tons of butter, whole automobiles, and so on. The image below shows integer solutions within a feasible domain defined by continuous function inequalities. Note that the requirement of an integer solution makes it far less obvious how to select the optimum point; it is no longer a vertex.



The branch-and-bound method comes to the rescue. In words, what we will do is successively solve continuous linear programming problems, but while imposing new inequality constraints that force the elements into taking integer values.

The method uses two major concepts. The first has to do with bounds and is quite intuitive. Suppose that in a solution domain X_1, the cost has a known *upper bound* $\bar{f}_1$, and that in a different domain X_2, the cost has a known *lower bound* $\underline{f}_2$. Suppose further that $\bar{f}_1 < \underline{f}_2$. If it is desired to minimize the function, then such a comparison clearly suggests we need spend no more time working in X_2. The second concept is that of a branching tree, and an example is the best way to proceed here. We try to maximize[1]

$$\begin{aligned} J &= 1000x_1 + 700x_2, \text{ subject to} \\ 100x_1 + 50x_2 &\leq 2425 \text{ and} \\ x_2 &\leq 25.5, \\ &\quad \text{with both } x_1, x_2 \text{ positive integers.} \end{aligned}$$



Hence we have four inequality constraints; the problem is not dissimilar to what is shown in the above figure. First, we solve the continuous linear programming problem, finding

$$A : J = 29350 : x_1 = 11.5, x_2 = 25.5.$$

Clearly, because neither of the solution elements is an integer, this solution is not valid. But it does give us a starting point in branch and bound: Branch this solution into two, where we consider the integers x_2 closest to 25.5:

$$\begin{aligned} B : J(x_2 \leq 25) &= 29250 : x_1 = 11.75, x_2 = 25 \text{ and} \\ C : J(x_2 \geq 26) &= X, \end{aligned}$$

where the X indicates we have violated one of our original constraints. So there is nothing more to consider along the lines of C. But we pursue B because it still has non-integer solutions, branching x_1 into

$$\begin{aligned} D : J(x_1 \leq 11, x_2 \leq 25) &= 28500 : x_1 = 11, x_2 = 25 \text{ and} \\ E : J(x_1 \geq 12, x_2 \leq 25) &= 29150 : x_1 = 12, x_2 = 24.5. \end{aligned}$$

[1] This problem is from G. Sierksma, Linear and integer programming, Marcel Dekker, New York, 1996.

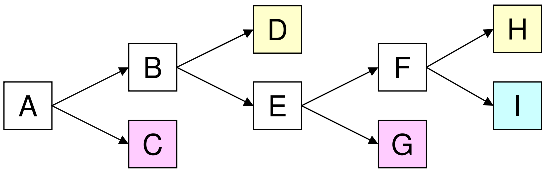
D does not need to be pursued any further, since it is has integer solutions; we store D as a possible optimum. Expanding E in x_2, we get

$$\begin{aligned} F : J(x_1 \geq 12, x_2 \leq 24) &= 29050 : x_1 = 12.25, x_2 = 24, \text{ and} \\ G : J(x_1 \geq 12, x_2 \geq 25) &= X. \end{aligned}$$

G is infeasible because it violates one of our original inequality constraints, so this branch dies. F has non-integer solutions so we branch in x_1:

$$\begin{aligned} H : J(x_1 \leq 12, x_2 \leq 24) &= 28800 : x_1 = 12, x_2 = 24 \text{ and} \\ I : J(x_1 \geq 13, x_2 \leq 24) &= 28750 : x_1 = 13, x_2 = 22.5. \end{aligned}$$

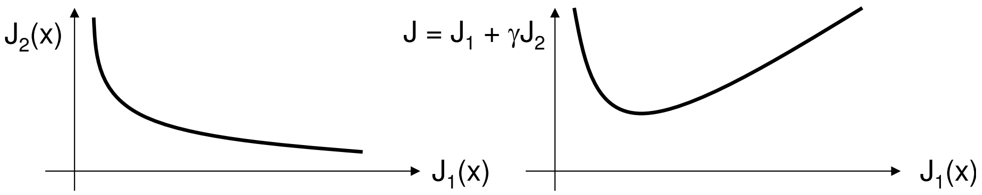
Now I is a non-integer solution, but even so it is not as good as H, which does have integer solution; so there is nothing more to pursue from I. H is better than D, the other available integer solution - so it is the optimal.



There exist many commercial programs for branch-and-bound solutions of integer linear programming problems. We implicitly used the upper-vs.-lower bound concept in terminating at I: if a non-integer solution is dominated by any integer solution, no branches from it can do better either.

7.5 Min-Max Optimization for Discrete Choices

A common dilemma in optimization is the existence of multiple objective functions. For example, in buying a motor, we have power rating, weight, durability and so on to consider. Even if it were a custom job - in which case the variables can be continuously varied - the fact of many objectives makes it messy. Clean optimization problems minimize one function, but if we make a meta-function out of many simpler functions (e.g., a weighted sum), we invariably find that the solution is quite sensitive to how we constructed the meta-function. This is not as it should be! The figure below shows on the left a typical tradeoff of objectives - one is improved at the expense of another. Both are functions of the underlying parameters. Constructing a meta-objective on the right, there is a indeed a minimum (plotted against $J_1(x)$), but its location depends directly on γ.



A very nice resolution to this problem is the min-max method. What we look for is the candidate solution with the smallest normalized deviation from the peak performance across objectives. Here is an example that explains. Four candidates are to be assessed according to three performance metrics; higher is better. They have raw scores as follows:

	Metric I	Metric II	Metric III
Candidate A	3.5	9	80
Candidate B	2.6	10	90
Candidate C	4.0	8.5	65
Candidate D	3.2	7.5	86

Note that each metric is evidently given on different scales. Metric I is perhaps taken out of five, Metric II is out of ten perhaps, and Metric III could be out of one hundred. We make four basic calculations:

- Calculate the range (max minus the min) for each metric: we get [1.4, 2.5, 35].
- Pick out the maximum metric in each metric: we have [4.0, 10, 90].
- Finally, replace the entries in the original table with the normalized deviation from the best:

	Metric I	Metric II	Metric III
Candidate A	(4.0-3.5)/1.4 = 0.36	(10-9)/2.5 = 0.4	(90-80)/35 = 0.29
Candidate B	(4.0-2.6)/1.4 = 1	(10-10)/2.5 = 0	(90-90)/35 = 0
Candidate C	(4.0-4.0)/1.4 = 0	(10-8.5)/2.5 = 0.6	(90-65)/35 = 1
Candidate D	(4.0-3.2)/1.4 = 0.57	(10-7.5)/2.5 = 1	(90-86)/35 = 0.11

- For each candidate, select the worst (highest) deviation: we get [0.4, 1, 1, 1].

The candidate with the lowest worst (min of the max!) deviation is our choice: Candidate A.

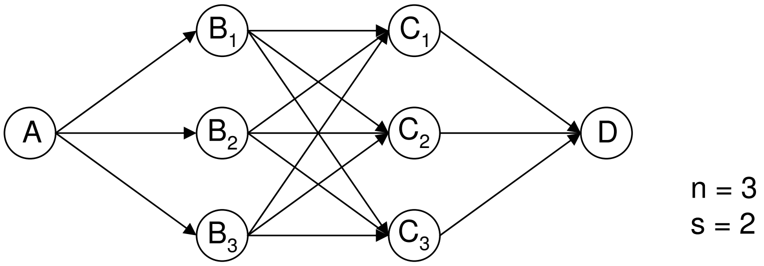
The min-max criterion can break a log-jam in the case of multiple objectives, but of course it is not without pitfalls. For one thing, are the metrics all equally important? If not, would a weighted sum of the deviations add any insight? We also notice that the min-max will throw out a candidate who scores at the bottom of the pack in any metric; this may or may not be perceived as fair. In broader terms, such decision-making can have fascinating social aspects.

7.6 Dynamic Programming

We introduce a very powerful approach to solving a wide array of complicated optimization problems, especially those where the space of unknowns is very high, e.g., it is a trajectory itself, or a complex sequence of actions, that is to be optimized. Only an introductory description here is given, focussing on *shortest-path* problems. A great many procedure and planning applications can be cast as shortest-path problems.

We begin with the essential concept. 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 a new optimization problem for getting only from 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. 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 the total optimization is done in twelve additions of two numbers.

Needless to say, this example gives only a mild advantage to the dynamic programming 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 15625 additions; dynamic programming needs only 105!

7.7 Solving Dynamic Programming on a Computer

Certainly the above algorithm can be implemented as written - moving backward from the end to the beginning, keeping track at each stage only of the optimal trajectories from that stage forward. This decision will involve some careful recording and indexing. A very simple algorithm called value iteration may be more accessible on your first try. As we will show in an example below, value iteration also allows us to consider problems where distinct stages are not clear.

It goes this way:

1. Index all of the possible configurations, or nodes, of the system (cities).

2. With each configuration, create a list of where we can go to from that node - probably this is a list of indices (cities that are plausibly part of an optimal path). The starting node (Los Angeles) is pointed to by no other nodes, whereas the end node (Boston) points to none.

3. For each of these simple paths defined from node to node, assign a cost of transition (simple driving miles between the cities).

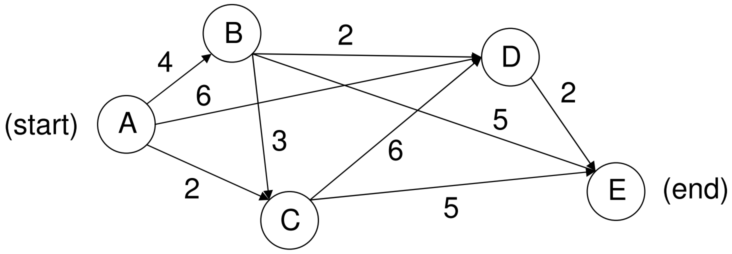
4. Now assign to each of these configurations an *initial guess* for what is the cost from this node to the end state (optimum total miles from each city to Boston). Clearly the costs-to-go for nodes that point to the terminal node are well-known, but none of the others are.

5. Sweep through all the configurations (except the terminal one), picking the best path out, based on the local path and the estimated cost at the next node. At each node, we have only to keep track of the best next node index, and the new estimated cost-to-go.

6. Repeat to convergence!

This algorithm can be shown to converge always, and has a number of variants and enhancements. An example makes things clearer:

Node	Points to Nodes	With Costs	Initial Estimate of Cost to Go
A (initial)	B,C,D	4,2,6	10
B	C,D,E	3,2,5	10
C	D,E	6,5	10
D	E	2	2 (known)
E	(terminal)	NA	NA



And here is the evolution of the value iteration:

	A	B	C	D
iteration	cost-to-go	cost-to-go	cost-to-go	cost-to-go
0	NA	10	10	2(E)
1	min(14,12,8) = 8(D,E)	min(13,4,5) = 4(D,E)	min(8,5) = 5(E)	2(E)
2	min(8,7,8) = 7(C,E)	4(D,E)	5(E)	2(E)

We can end safely after the second iteration because the path from A involves C, which cannot change from its value after the first iteration, because it connects all the way through to E.

8 STOCHASTIC SIMULATION

Whereas in optimization we seek a set of parameters $\vec{x}$ to minimize a cost, or to maximize a reward function $J(\vec{x})$, here we pose a related but different question. Given a system S, it is desired to understand how *variations in the defining parameters $\vec{x}$ lead to variations in the system output*. We will focus on the case where $\vec{x}$ is a set of random variables, that can be considered unchanging - they are static. In the context of robotic systems, these unknown parameters could be masses, stiffness, or geometric attributes. How does the system behavior depend on variations in these physical parameters? Such a calculation is immensely useful because real systems have to be robust against modeling errors.

At the core of this question, the random parameters x_i in our discussion are described by distributions; for example each could have a pdf $p(x_i)$. If the variable is known to be normal or uniformly distributed, then of course it suffices to specify the mean and variance, but in the general case, more information may be needed.

8.1 Monte Carlo Simulation

Suppose that we make N simulations, each time drawing the needed random parameters x_i from a random number "black box" (about which we will give more details in the next section). We define the high-level output of our system S to be $g(\vec{x})$. For simplicity, we will say that $g(\vec{x})$ is a scalar. $g(\vec{x})$ can be virtually any output of interest, for example: the value of one state at a given time after an impulsive input, or the integral over time of the trajectory of one of the outputs, with a given input. In what follows, will drop the vector notation on x for clarity.

Let the *estimator* G of $g(x)$ be defined as

$$G = \frac{1}{N} \sum_{j=1}^{N} g(x_j).$$

You recognize this as a straight average. Indeed, taking the expectation on both sides,

$$E(G) = \frac{1}{N} \sum_{j=1}^{N} E(g(x_j)),$$

it is clear that $E(G) = E(g)$. At the same time, however, we do not know $E(g)$; we calculate G understanding that with a very large number of trials, G should approach $E(g)$. Now let's look at the variance of the estimator. This conceptually results from an infinite number of estimator trials, each one of which involves N evaluations of g according to the above definition. It is important to keep in mind that such a variance involves samples of the estimator (each involving N evaluations) - not the underlying function $g(x)$. We have

$$\begin{aligned}
\sigma^2(G) &= \sigma^2\left[\frac{1}{N}\sum_{j=1}^{N} g(x_j)\right] \\
&= \frac{1}{N^2}\sigma^2\left[\sum_{j=1}^{N} g(x_j)\right] \\
&= \frac{1}{N^2}\sum_{j=1}^{N}\sigma^2(g) \\
&= \frac{1}{N}\sigma^2(g).
\end{aligned}$$

This relation is key. The first equality follows from the fact that $\sigma^2(nx) = n^2\sigma^2(x)$, if n is a constant. The second equality is true because $\sigma^2(x+y) = \sigma^2(x) + \sigma^2(y)$, where x and y are random variables. The major result is that $\sigma^2(G) = \sigma^2(g)$ if only one-sample trials are considered, but that $\sigma^2(G) \to 0$ as $N \to \infty$. Hence with a large enough N, we can indeed expect that our G will be very close to $E(g)$.

Let us take this a bit further, to get an explicit estimate for the error in G as we go to large N. Define a nondimensional estimator error

$$\begin{aligned}
q &= \frac{G - E(g)}{\sigma(G)} \\
&= \frac{(G - E(g))\sqrt{N}}{\sigma(g)},
\end{aligned}$$

where the second equality comes from the result above. We call the factor $\sigma(g)/\sqrt{N}$ the standard error. Invoking the central limit theorem, which guarantees that the distribution of G becomes Gaussian for large enough N, we have

$$\begin{aligned}
\lim_{N\to\infty} \text{prob}(a < q < b) &= \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt \\
&= F(a) - F(b),
\end{aligned}$$

where $F(x)$ is the cumulative probability function of the standard Gaussian variable:

$$F(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt$$

Looking up some values for $F(x)$, we see that the nondimensional error is less than one in 68.3% of trials; it is less than two in 95.4% of trials, and less than three in 99.7% of trials. The 99.7% confidence interval corresponds with

$$\begin{aligned}
-3 \leq {} & (G - E(g))\sqrt{N}/\sigma(g) & \leq 3 \rightarrow \\
-3\sigma(g)/\sqrt{N} \leq {} & G - E(g) & \leq 3\sigma(g)/\sqrt{N}.
\end{aligned}$$

In general, quadrupling the number of trials improves the error by a factor of two.

So far we have been describing a single estimator G, which recovers the mean. The mean, however, is in fact an integral over the random domain:

$$E(g) = \int_{x \epsilon X} p(x) g(x) dx,$$

where $p(x)$ is the pdf of random variable x. So the Monte Carlo estimator G is in fact an integrator:

$$G \simeq \int_{x \epsilon X} p(x) g(x) dx.$$

We can just as easily define estimators of statistical moments:

$$G_n = \frac{1}{N} \sum_{j=1}^{N} x_j^n g(x_j) \simeq \int_{x \exists X} x^n p(x) g(x) dx,$$

which will follow the same basic convergence trends of the mean estimator G. These moments can be calculated all using the same N evaluations of $g(x)$.

The above equation gives another point of view to understand how the Monte Carlo approach works: the effect of the probability density function in the integral is replaced by the fact that random variables in MC are drawn from the same distribution. In other words, a high $p(x)$ in a given area of the domain X amplifies $g(x)$ there. MC does the same thing, because there are in fact more x drawn from this area, in making the N evaluations.

8.2 Making Random Numbers

The Monte Carlo method requires that we fire into our evaluation $g(x)$ a group of N random numbers (or sets of random numbers), drawn from a distribution (or a set of distributions for more than one element in x). Here we describe how to generate such data from simple distributions.

Note that both the normal and the uniform distributions are captured in standard MATLAB commands.

We describe in particular how to generate samples of a given distribution, from random numbers taken from an underlying *uniform* distribution. First, we say that the cumulative probability function of the uniform distribution is

$$P(w) = \begin{cases} 0, & w \leq 0 \\ w, & 0 < w < 1 \\ 1, & w \geq 1 \end{cases}$$

If $x = r(w)$ where r is the transformation we seek, recall that the cumulative probabilities are

$$P(x) = P(r(w)) = P(w) = w,$$

and the result we need is that

$$w = P(x) = \int_{-\infty}^{x} p(x)dx.$$

Our task is to come up with an x that goes with the uniformly distributed w - it is not as hard as it would seem. As an example, suppose we want to generate a normal variable x (zero mean, unity variance). We have

$$\begin{aligned} P(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt &= w \\ F(x) &= w, \text{ or} \\ x &= F^{-1}(w), \end{aligned}$$

where $F(x)$ is the cumulative probability function of a standard Gaussian variable (zero mean, unity variance), and can be looked up or calculated with standard routines. Note $F(x)$ is related within a scale factor to the error function (erf).

As another example, consider the exponential distribution

$$p(x) = \lambda e^{-\lambda x};$$

this distribution is often used to describe the time of failure in complex systems. We have

$$\begin{aligned} P(x) = \int_{0}^{x} \lambda e^{-\lambda t} dt &= w \\ 1 - e^{-\lambda x} &= w \text{ or} \\ x &= -\frac{\log(1-w)}{\lambda}. \end{aligned}$$

Similarly, this procedure applied to the Rayleigh distribution

$$p(x) = xe^{-x^2/2}$$

gives $x = \sqrt{-2\log(1-w)}$. In these formulas, we can replace $(w-1)$ with w throughout; w is uniformly distributed on the interval [0,1], so they are equivalent.

8.3 Grid-Based Techniques

As noted above, moment calculations on the output variable are essentially an integral over the domain of the random variables. Given this fact, an obvious approach is simply to focus on a high-quality integration routine that uses some *fixed* points x - in the Monte Carlo method, these were chosen at random. In one dimension, we have the standard trapezoid rule:

$$\begin{aligned}
\int_a^b g(x)dx &\approx \sum_{i=1}^{n} w_i g(x_i), \text{ with} \\
w_1 &= (b-a)/2(n-1) \\
w_n &= (b-a)/2(n-1) \\
w_2, \cdots, w_{n-1} &= (b-a)/(n-1) \\
x_i &= a + (i-1)(b-a)/(n-1).
\end{aligned}$$

Here the w_i are simply weights, and g is to be evaluated at the different abscissas x_i. This rule has error of order $1/n^2$, meaning that a doubling in the number of points n gives a fourfold improvement in the error. To make an integration in two dimensions we take the tensor product of two single-dimension calculations:

$$\int_a^b \int_c^d g(x)dx \approx \sum_{i=0}^{n_1} \sum_{j=0}^{n_2} w_i w_j g(x_{ij}).$$

Here the abscissas have two elements, taken according to the grids in the two directions. In many applications, the trapezoid rule in multi-dimensions works quite well and can give a good understanding of the various moments of the given system. Extensions such as Romberg integration, with the tensor product, can be applied to improve results.

A particularly powerful technique involving orthogonal polynomials is also available and it gives truly remarkable accuracy for smooth functions. For our development here, we will focus on normal distributions of random variables. These pertain to a particular set of orthogonal polynomials known as the Hermite (pronounced "hermit") polynomials. These are:

$$\begin{aligned}
h_0(x) &= 1 \\
h_1(x) &= x \\
h_2(x) &= x^2 - 1 \\
&\cdots \\
h_{n+1}(x) &= x h_n(x) - n h_{n-1}(x) \text{ (recurrence relation).}
\end{aligned}$$

The defining feature of this set of polynomials is that

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} h_i(x) h_j(x) dx = \begin{cases} 0, & \text{if and only if } i \neq j \\ 1, & \text{if and only if } i = j. \end{cases}$$

Note that we have chosen a scaling so that the inner product comes out to be exactly one - some textbook definitions of the Hermite polynomials will not match this. Note also that the inner product is taken with respect to the Gaussian exponential weighting function, which we equate below with the Gaussian pdf. Now the magic comes from the following[2]: a $(2n-1)$'th order polynomial $g(x)$ can be written as

$$g(x) \quad = \quad h_n(x)[a_{n-1}h_{n-1}(x) + a_{n-2}h_{n-2}(x) + \cdots + a_0h_0(x)] + b_{n-1}h_{n-1}(x) + \cdots + b_0h_0(x).$$

This formula, with the $2n$ coefficients a and b, covers the $(2n-1)$'th-order term with $a_{n-1}h_n(x)h_{n-1}(x)$, and the zero'th order term with $b_0h_0(x)$. It can be shown that all the products are linearly independent, so that indeed *all* polynomials up to order $2n-1$ are accounted for.

Returning to the integration problem, for the determination of moments, recall the definition that

$$E(g(x)) = \int_{x \epsilon X} p(x)g(x)dx,$$

where $p(x)$ is the probability density function of random variable x. Employing specifically the Gaussian pdf in $p(x)$ and the Hermite polynomials so as to achieve orthogonality, we can integrate our $g(x)$ as follows:

$$\int_{-\infty}^{\infty} p(x)g(x)dx \quad = \quad b_0 \int_{-\infty}^{\infty} p(x)h_0(x)dx = b_0.$$

Thus, we have only to find b_0! To do this, we cleverly select the abscissas to be the zeros (or roots) of $h_n(x)$ - let's call them $x_1, \cdots, x_n$. We obtain a linear system:

$$\begin{aligned} \left\{ \begin{array}{c} g(x_1) \\ \vdots \\ g(x_N) \end{array} \right\} &= \left[\begin{array}{ccc} h_{n-1}(x_1) & \cdots & h_0(x_1) \\ \vdots & & \vdots \\ h_{n-1}(x_n) & \cdots & h_0(x_n) \end{array} \right] \left\{ \begin{array}{c} b_{n-1} \\ \vdots \\ b_0 \end{array} \right\}, \text{ or} \\ \vec{g} &= H\vec{b} \end{aligned}$$

Notice that the a coefficients do not appear here because the x_i are taken at the roots of $h_{n-1}(x)$. The linear system can be solved easily as $\vec{b} = H^{-1}\vec{g}$, and we have a special interest in the last row, which of course has the solution for b_0. Indeed, the bottom row of H^{-1} is the set of weights w_i, in complete analogy with the weights we defined above for the trapezoid rule:

$$b_0 = \int_{-\infty}^{\infty} p(x)g(x)dx = \sum_{i=1}^{n} w_i g(x_i).$$

[2] Parts of this derivation follow J.R. Hockenberry and B.C. Lesieutre (2004), IEEE Transactions on Power Systems, 19:1483-1491.

8.4 Issues of Cost and Accuracy

The balance of cost versus error for the methods above are summarized as:

- The Monte-Carlo estimator has a variance that decreases with $1/N$, where N is the number of evaluations. There is no dependence on the random dimension d or on the form of the function.
- The trapezoidal rule in one dimension has an error that decreases with n^{-2}.
- Gauss-Hermite quadrature in one dimension matches exactly all polynomials $g(x)$ of order less than or equal to $2n-1$, and makes the best fit possible when it is of higher order. The error is zero if $g(x)$ is a polynomial of order $2n-1$ or less, but if not then the error goes as n^{-r}, where r is the "smoothness" of the function. The smoothness is the same as the number of derivatives that can be taken everywhere in the domain.

Some simple but important scaling laws show that the Monte Carlo will ultimately outperform any quadrature rule, in high enough dimensions. Recall that the MC estimator G has error - or standard deviation of error - that scales as

$$\epsilon = O(1/\sqrt{N})$$

Now, in contrast, the error of a quadrature rule in one dimension scales as:

$$\epsilon = O(n^{-k}),$$

where k is a parameter of the method used and the function itself, two for the trapezoid rule we described above, and the smoothness r for Gaussian quadrature. Considering the multi-dimensional quadrature case, the error stays the same but now the total number of evaluations N includes *all* the grid points, i.e., $N = n^d$ (assuming $N_1 = N_2 = \cdots = N$). Hence, here we have

$$\epsilon = O((N^{1/d})^{-k}) = O(N^{-k/d})$$

Error in the quadrature rules is affected dramatically by d: Even in two dimensions, the n^{-2} convergence of the trapezoid rule is degraded to n^{-1}! In four dimensions, the error rate of $1/\sqrt{N}$ is the same as for Monte Carlo. The Gaussian quadrature may do better for some functions, but there are plenty of them for which $r = 2$, for example the innocuous $g(x) = x^2|x|$!

Another major factor to consider is that the quadrature rules are inherently trying to make a polynomial approximation of the function, whereas the Monte Carlo technique has no such intention. Discontinuous and otherwise non-smooth functions in particular can cause serious problems for the quadratures, which must have many points in order to cover the sharp spots accurately.

Summarizing, we recommend that you keep the two major classes of integration tools handy - grid-less and grid-based. For lower dimensions and smoother functions, Gaussian quadratures can provide exceptional results, whereas the Monte-Carlo workhorse always comes out on top for high-dimension, difficult functions. The differences are trivial for very cheap evaluations of g, but become very compelling when each evaluation takes a lot of computer time, or involves an actual experiment.

There are substantial extensions to multi-dimensional quadrature rules and to Monte Carlo methods, some of which are available in the following references:

M.H. Kalos and P.A. Whitlock, 1986, *Monte Carlo methods, volume 1: basics*, New York: Wiley.

W.H. Press, S.A. Teukolsky, W.T. Vetterling, and B.P. Flannery, 1992, *Numerical recipes in C*, Cambridge, UK: Cambridge University Press.

9 KINEMATICS OF MOVING FRAMES

An understanding of inertial guidance systems for navigation of vehicles and robots requires some background in kinematics. Central in our discussion is the use of multiple reference frames. Such frames surround us in our daily lives:

- Earth latitude and longitude
- Forward, backward motion relative to current position
- Right, left motion
- Axes printed on an inertial measurement unit
- Vehicle-referenced coordinates, e.g., relative to the centroid

We first describe how to transform vectors through changes in reference frame. Considering differential rotations over differential time elements gives rise to the concept of the rotation vector, which is used in deriving inertial dynamics in a moving body frame.

9.1 Rotation of Reference Frames

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 denote through a subscript the specific reference system of a vector. Let a vector expressed in the inertial (Earth) 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 principal axes, and has a solid link with the intuitive notions of roll, pitch, and yaw. One of the problems with Euler angles, however, is that for certain specific values the transformation exhibits discontinuities (as will be seen below). Quaternions present a more elegant and robust method, but with more abstraction. We will develop the equations of motion here 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, e.g., the walls of the room you are in, 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.

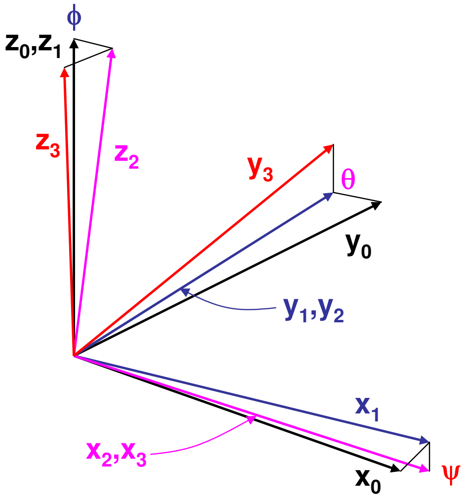


Figure 1: Successive application of three Euler angles transforms the original coordinate frame into an arbitrary orientation.

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

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

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:

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

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

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

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!):

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

All of the transformation matrices, including $R(\phi,\theta,\psi)$, are orthonormal: their inverse is equivalent to their transpose, so that $\vec{x} = R^T\vec{x}_b$. 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

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

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

9.2 Differential Rotations

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

$$\begin{aligned} R &\simeq \begin{bmatrix} 1 & \delta\phi & -\delta\theta \\ -\delta\phi & 1 & \delta\psi \\ \delta\theta & -\delta\psi & 1 \end{bmatrix} \\ &= \begin{bmatrix} 0 & \delta\phi & -\delta\theta \\ -\delta\phi & 0 & \delta\psi \\ \delta\theta & -\delta\psi & 0 \end{bmatrix} + I_{3\times3}, \end{aligned}$$

where $I_{3\times3}$ 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 differential

Euler angles, ordered with the axes $[x, y, z]$. Small rotations are completely decoupled; their order does not matter. Since $R^{-1} = R^T$, we have also $R^{-1} = I_{3\times 3} + \delta \vec{E} \times$;

$$\begin{aligned} \vec{x}_b &= \vec{x} - \delta \vec{E} \times \vec{x} \\ \vec{x} &= \vec{x}_b + \delta \vec{E} \times \vec{x}_b. \end{aligned}$$

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:

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

Dividing by the differential time step gives

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

where the *rotation rate* vector $\vec{\omega} \simeq d\vec{E}/dt$ 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:

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

such that

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

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:

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

Finally, allowing the origin to move as well gives

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

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

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

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 $\vec{f}$:

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

9.3 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 $\vec{\omega} \rightarrow d\vec{E}/dt$.

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:

$$\begin{aligned} \vec{\omega} &= R(\psi)R(\theta)\left\{\begin{array}{c} 0 \\ 0 \\ d\phi/dt \end{array}\right\} + R(\psi)\left\{\begin{array}{c} 0 \\ d\theta/dt \\ 0 \end{array}\right\} + \left\{\begin{array}{c} d\psi/dt \\ 0 \\ 0 \end{array}\right\} \\ &= \left[\begin{array}{ccc} 1 & 0 & -\sin\theta \\ 0 & \cos\psi & \sin\psi\cos\theta \\ 0 & -\sin\psi & \cos\psi\cos\theta \end{array}\right]\left\{\begin{array}{c} d\psi/dt \\ d\theta/dt \\ d\phi/dt \end{array}\right\}. \end{aligned}$$

Taking the inverse gives

$$\begin{aligned} \frac{d\vec{E}}{dt} &= \left[\begin{array}{ccc} 1 & \sin\psi\tan\theta & \cos\psi\tan\theta \\ 0 & \cos\psi & -\sin\psi \\ 0 & \sin\psi/\cos\theta & \cos\psi/\cos\theta \end{array}\right]\vec{\omega} \\ &= \Gamma(\vec{E})\vec{\omega}. \end{aligned}$$

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 many vehicles, the singularity in pitch is acceptable, because a ninety-degree pitch angle is outside the normal operating condition.

9.4 A Practical Example: 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 of this type.

Suppose that the heading is estimated from an angular rate gyro. We use

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

where r is the measured angular rate. As you might expect, long-term errors in this rule will be worse than for the previous, because integration of the rate gyro signal is subject to drift.

Suppose that we have in addition to a sensor for U and r, a sensor for the cross-body velocity V. Our dead-reckoning problem is

$$\begin{aligned}\dot{\phi} &= r \\ \dot{x} &= U\cos\phi - V\sin\phi \\ \dot{y} &= U\sin\phi + V\cos\phi,\end{aligned}$$

10 VEHICLE INERTIAL DYNAMICS

We consider the rigid body dynamics with a coordinate system affixed on the body. We will develop equations useful for the simulation of vehicles, as well as for understanding the signals measured by an inertial measurement unit (IMU).

A common frame for boats, submarines, aircraft, terrestrial wheeled and other 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.

10.1 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

$$\vec{F} = \frac{d}{dt}(m\vec{v})$$

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

$$\vec{F}_i + \vec{R}_i = \frac{d}{dt}(m_i\vec{v}_i),$$

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

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

Then summing up all the particle momentum equations gives

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

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.

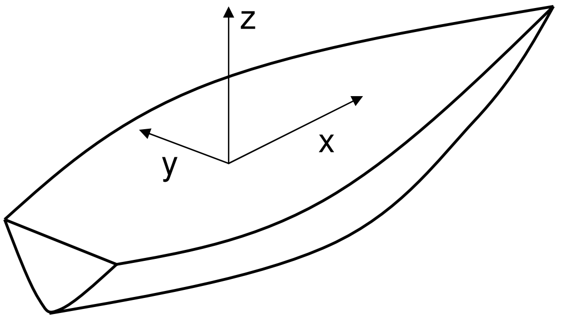


Figure 2: Convention for the body-referenced coordinate system on a vehicle: x is forward, y is sway to the left, and z is heave upwards. Looking forward from the vehicle "helm," roll about the x axis is positive counterclockwise, pitch about the y-axis is positive bow-down, and yaw about the z-axis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

10.2 Linear Momentum in a Moving Frame

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

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

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

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

we have

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

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

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

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

$$\vec{v}_o = \{u, v, w\} \text{ (body-referenced velocity)}$$

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

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

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

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., $\vec{r}_G \neq \vec{0}$.

10.3 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

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

10.3.1 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.

10.3.2 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:

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

10.3.3 Stationary Frame

A frame fixed in inertial space, and momentarily coincident with the frame on the mass (10.3.1), 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.

10.4 Angular Momentum

For angular momentum, the summed particle equation is

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

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

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

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

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

The second term expands as (using the triple product)

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

Employing the definitions of moments of inertia,

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

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:

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

Letting $\vec{M} = \{K, M, N\}$ be the total moment acting on the body, i.e., the left side of Equation 1, the complete moment equations are

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

10.5 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

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

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.

10.5.1 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.

10.5.2 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 $\delta p(t) = \delta p(0) e^{\sqrt{-\alpha} t}$.

10.5.3 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.

10.6 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 vehicle 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:

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

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.

10.7 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 our equations for linear and angular momentum, with some coupling that generally requires a 6 × 6 matrix inverse. The Cartesian position propagates according to

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

while the Euler angles follow:

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

10.8 What Does an Inertial Measurement Unit Measure?

A common in-the-box assembly of components today is a perpendicular triad of accelerometers (strain-guage typically), along with a triad of angular rate gyros. The six measurements of this inertial measurement unit (IMU) have generally obviated inclinometers, which are functionally equivalent to a pendulum whose angle (following gravity) relative to the housing is measured via a potentiometer.

This combination of sensors within an IMU brings up a fundamental user parameter. First, the accelerometers on a non-accerating frame will point down (gravity); they can be used to estimate pitch and roll, and hence replace inclinometers. When the platform actually does accelerate, however, the measured acceleration vector is the vector sum of the true acceleration and the gravity effect. So the pitch and roll of an IMU during accelerations is critical if we are to separate out the gravity effect from the measured accelerometer signals. The rate gyros possess a different characteristic: they are completely insensitive to linear acceleration (and gravity), but suffer a bias, so that the integration of a measured rate to deduce angle will drift. A typical drift rate for a fiber optic gyro is 72^o/hour, certainly not good enough for a long-term pitch or roll measurement. In the short term, gyros are quite accurate.

The accelerometers and rate gyros are typically taken together to derive a best estimate of pitch and roll. Specifically, the *low-frequency* components of the accelerometer signals are used to eliminate the drift in the angle estimates; the assumption is that a controlled body generally has only short periods of significant linear acceleration. Conversely, the *high-frequency* portion of the the rate gyros' signals are integrated to give a short-term view of attitude. The interesting user parameter is, then, deciding whether what time frame applies to the accelerometer signals, and what time frame applies to the rate gyro signals.

Two additional points can be made about IMU's. First, an IMU with three accelerometers and three rate gyros has no idea what direction is north; hence, an IMU is typically augmented with a magnetic compass. Such a compass has a slower response than the rate gyros and so a frequency division as above can be used. Our second point is that the double integration of measured accelerations is ill-advised in an IMU, due to accumulating biases. A special class of IMU, called an inertial navigation system (INS), however has high quality sensors that make this step possible. Even then, some additional sources of navigation correction are needed for long-term applications.

The three accelerometers measure the total derivative of velocity, in the body frame, plus the projection of gravity onto the sensor axes. Using the above notation, assuming the sensor [x,y,z] is aligned with the body [x,y,z], and assuming that the sensor is located at the vector $\vec{r}_S$, this is

$$\begin{aligned}
\mathrm{acc}_x &= \frac{\partial u}{\partial t} + qw - rv + \frac{dq}{dt}z_S - \frac{dr}{dt}y_S + (qy_S + rz_S)p - (q^2 + r^2)x_S - \sin\theta g \\
\mathrm{acc}_y &= \frac{\partial v}{\partial t} + ru - pw + \frac{dr}{dt}x_S - \frac{dp}{dt}z_S + (rz_S + px_S)q - (r^2 + p^2)y_S + \sin\psi\cos\theta g
\end{aligned}$$

$$\text{acc}_z = \frac{\partial w}{\partial t} + pv - qu + \frac{dp}{dt} y_S - \frac{dq}{dt} x_S + (px_S + qy_S)r - (p^2 + q^2) z_S + \cos\psi \cos\theta g.$$

Here $g = 9.81 m/s^2$, and $[\phi, \theta, \psi]$ are the three Euler angle rotations. The accelerations have some intuitive elements. The first term on the right-hand side captures actual honest-to-goodness linear acceleration. The second and third terms capture centripetal acceleration - e.g., in the y-channel, an acceleration ru is reported, the product of the forward velocity u and the leftward turning rate r. The fourth and fifth terms account for the linear effect of placing the sensor away from the body origin; later terms capture the nonlinear effects. Gravity comes in most naturally in the acceleration in the z-channel: if the roll and pitch Euler angles are zero, then the sensor thinks the vehicle is accelerating upward at one g.

The rate gyros are considerably easier!

$$\begin{aligned} \text{rate}_x &= p \\ \text{rate}_y &= q \\ \text{rate}_z &= r. \end{aligned}$$

The rate gyros measure the body-referenced rotation rates.

11 CONTROL FUNDAMENTALS

11.1 Introduction

11.1.1 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 macro-economics 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 plant 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	inflation, M1
Nuclear reactor	cooling, neutron flux	heat, power level	temp., pressure

11.1.2 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 heuristic 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 other equations.

Beyond prediction of plant behavior based on physics, *system identification* generates a plant model from actual 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.

11.1.3 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).

11.2 Partial Fractions

Partial fractions are presented here, in the context of control systems, as the *fundamental* link between pole locations and stability. Solving linear time-invariant systems by the Laplace Transform method will generally create a signal containing the (factored) form

$$Y(s) = \frac{K(s+z_1)(s+z_2)\cdots(s+z_m)}{(s+p_1)(s+p_2)\cdots(s+p_n)}. \tag{1}$$

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 first- and second-order transform pairs can be used to find the time-domain output signals. We must have $m < n$ for this procedure; if this is not the case, then we have to strip off extra powers of s to solve the problem, and then add them back on at the end.

11.2.1 Partial Fractions: Unique Poles

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

$$Y(s) = \frac{a_1}{s+p_1} + \frac{a_2}{s+p_2} + \cdots \frac{a_n}{s+p_n}, \tag{2}$$

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.

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.

11.2.2 Partial Fractions: Complex-Conjugate Poles

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

$$Y(s) = \frac{b_1 s + b_2}{(s+p_1)(s+p_2)} + \frac{a_3}{s+p_3} + \cdots, \qquad (3)$$

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

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 $\zeta = 0$)

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

11.3 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.*

11.4 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.

11.5 General Stability

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

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

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

11.6 Representing Linear Systems

The transfer function description of linear systems has already been described in the presentation 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 many standard tools from linear algebra.

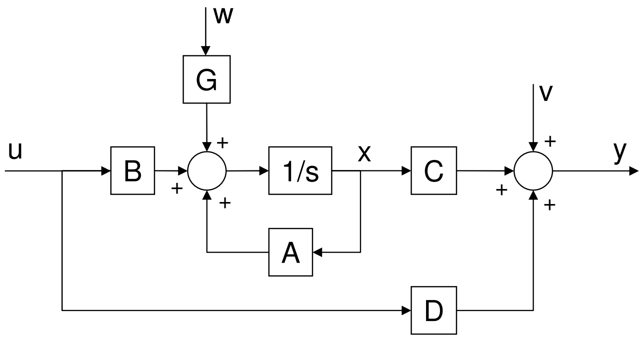
11.6.1 Standard State-Space Form

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

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

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.



11.6.2 Converting a State-Space Model into a Transfer Function

Many different state-space descriptions can create the same transfer function - they are not unique. In the case of no disturbances or noise, the transfer function can be written as

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

where I is the identity matrix with the same size as A. To see that this is true, simply transform the differential equation into frequency space:

$$\begin{aligned} sx(s) &= Ax(s) + Bu(s) \longrightarrow \\ x(s)(sI - A) &= Bu(s) \longrightarrow \\ x(s) &= (sI - A)^{-1}Bu(s) \longrightarrow \\ y(s) &= Cx(s) + Du(s) = C(sI - A)^{-1}Bu(s) + Du(s). \end{aligned}$$

A similar equation holds for $y(s)/w(s)$, and clearly $y(s)/v(s) = 1$.

11.6.3 Converting a Transfer Function into a State-Space Model

Because state-space models are not unique, there are many different ways to create them from a transfer function. In the simplest case, 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:

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

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

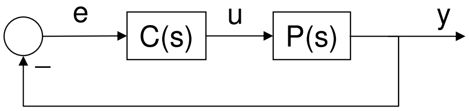
$$\begin{aligned} \frac{d\vec{x}}{dt} &= \begin{bmatrix} -b/m & -k/m \\ 1 & 0 \end{bmatrix} \vec{x} + \begin{bmatrix} 1/m \\ 0 \end{bmatrix} u \\ y &= [1 \;\; 1]\, \vec{x} \end{aligned}$$

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.

11.7 Block Diagrams and Transfer Functions of Feedback Systems

11.7.1 Block Diagrams: Fundamental Form

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, a simple mass driven by a controlled force 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 (see below) 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 $u(s) = -(k_p + k_d s)y(s)$. This feedback loop in block diagram form is shown below.



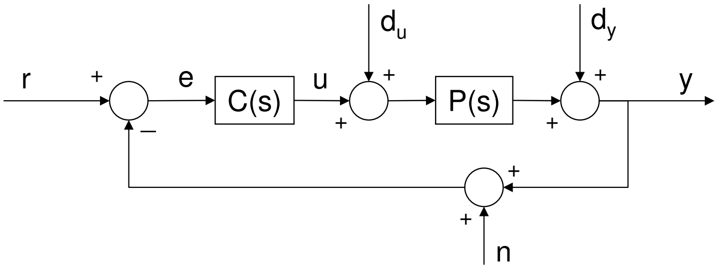
11.7.2 Block Diagrams: General Case

The simple feedback system above is augmented in practice by three external inputs. The first is a process disturbance we call d, 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 n, 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.

Note that the disturbances d_u and d_y, and the noise n are generalizations on the unknown disturbance and sensor noise we discussed at the beginning of this section.



11.7.3 Transfer Functions

Some algebra applied to the above figure (and neglecting the Laplace variable s) shows that

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

Let us derive the first of these. Working directly from the figure, we have

$$\begin{aligned}e(s) &= r(s) - y(s) \\ e(s) &= r(s) - P(s)u(s) \\ e(s) &= r(s) - P(s)C(s)e(s) \\ (1+P(s)C(s))e(s) &= r(s) \\ \frac{e(s)}{r(s)} &= \frac{1}{1+P(s)C(s)}.\end{aligned}$$

The fact that we are able to make this kind of construction is a direct consequence of the frequency-domain representation of the system, and namely that we can freely multiply and divide system impulse responses and signals, so as to represent convolutions in the time-domain.

Now $e/r = S$ 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 certain frequencies, so that the non-dimensional tracking error e/r 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 some others:

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

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

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 be the procedure when we address loopshaping.

11.8 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

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

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. This will retard motion toward zero error, which helps to reduce overshoot.

The common variations are: P, PD, PI, PID.

11.9 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 $y'(0) = 0$.

11.9.1 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.

11.9.2 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 tolerated 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.

11.9.3 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. Note that for the case of the PD control, it was enough to select k_p positive and k_d positive because these terms represent spring and dashpot-type forces. The use of k_i complicates the stability however, and it is not enough in general to set all three gains positive - stability should be checked explicitly.

11.10 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:

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

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.

12 CONTROL SYSTEMS – LOOPSHAPING

12.1 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 linear quadratic Gaussian (LQG) control and the loop transfer recovery (LTR).

12.2 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 feedback 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 $s = j\omega$, $\omega = [-\infty, \infty]$. Remarkably, 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, also the `rlocus()` command in MATLAB).

12.2.1 Mapping Theorem

To give some understanding of the Nyquist plot, we begin by imposing 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.

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

Every zero of $F(s)$ that is enclosed by a path in the s-plane generates exactly one CW encirclement of the origin in the $F(s)$-plane. Conversely, every pole of $F(s)$ that is enclosed by a path 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 $Z - P = CW$.

So it will be possible to relate poles and zeros in the $F(s)$-plane to encirclements of the origin in the s-plane. Since we get to design the path in the s-plane, the trick is to 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:

$$P = CCW.$$

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.

12.2.2 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 $F(s)$*:

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

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:

- From the formula, if neither the plant nor the controller have unstable poles, then the loci of $P(s)C(s)$ must not encircle the critical point at all, for closed-loop stability. If the plant and the controller comprise q unstable poles, then the loci of $P(s)C(s)$ must encircle the critical point q times in the CCW dirction.

- 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.

12.2.3 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 $P(s)C(s) = -1 = 1\angle 180^\circ$, and how close the loci come to it can be expressed in terms of magnitude and angle:

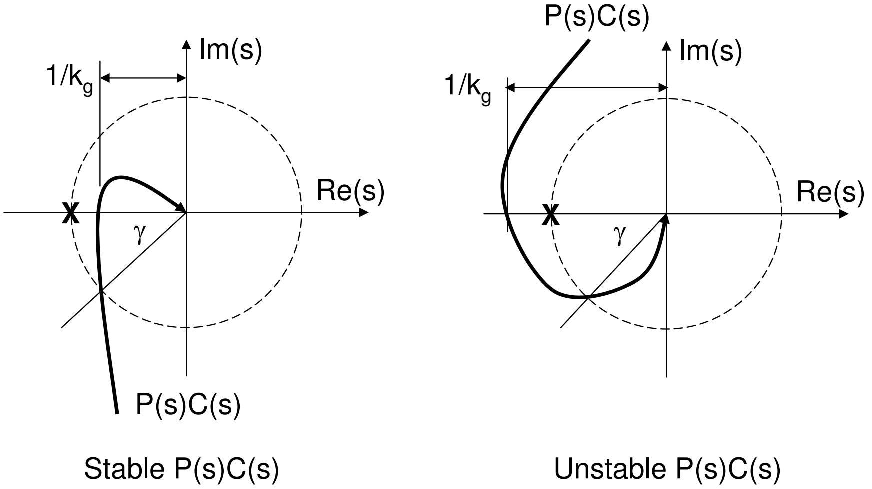
- When the angle of $P(s)C(s)$ is -180°, the magnitude $|P(s)C(s)|$ should not be near one.

- 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, we desire $k_g \geq 2$ and $\gamma \geq 30^\circ$; for an unstable plant, $k_g \leq 0.5$ and $\gamma \geq 30^\circ$. As defined, these conditions will maintain stability even if the gain is increased by a factor of two for the stable open-loop system, or decreased by a factor of two for the unstable OL system. In both cases, the phase angle can be in error by thirty degrees without losing stability. Note that the system behavior in the closed-loop, while technically stable through these perturbations, might be very poor from the performance point of view. The following two sections outline how to manage robustness and performance simultaneously using the Nyquist plot.

12.3 Design for Nominal Performance

Performance requirements of a feedback controller, using the nominal plant model, can be cast in terms of the Nyquist plot. 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

$$|W_1(s)S(s)| < 1,$$

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.

12.4 Design for Robustness

It is ubiquitous 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 to 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:

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

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 $P = k_0/(s-1)$, augmented 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 $W_2 = 3/7$. In particular, $k_0 \pm W_2 = [2, 5]$

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

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

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

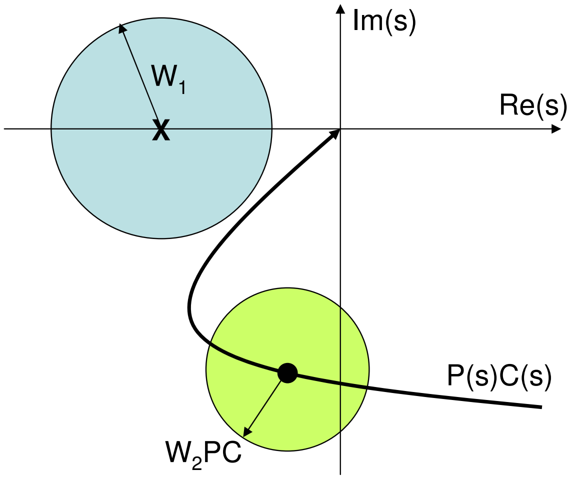
12.5 Robust Performance

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

$$|W_1S| + |W_2T| < 1.$$

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.



12.6 Implications of Bode's Integral

The loop transfer function PC cannot roll off too rapidly in the crossover region, and this limits how "dramatic" can be the loop shapes that we create to achieve robustness, nominal performance, or both. 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

$$\angle H(j\omega_0) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{d}{d\nu}\left[\log(|H(j\omega)|) \cdot \log(\coth(|\nu|/2))\right] d\nu,$$

where $\nu = \log(\omega/\omega_0)$, and coth() is the hyperbolic cotangent. 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 $\log(\coth(|\nu|/2))$ is nonzero only near $\nu = 0$, implying that the angle depends only on the local slope $d(\log|H|)/d\nu$. 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 $\log|H| = -n\log(\omega/\omega_0)$, so that $d(\log|H|)/d\nu = -n$. Then we have just

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

This integral is easy to look up or compute. Each pole at the origin 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 add 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 ($-20 dB/decade$, equivalent to a single pole), and may be second-order ($-40 dB/decade$) if an adequate phase angle can be maintained near crossover.

12.7 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, $L = PC$. It suffices to just pick

$$C = L/P.$$

This simple step involves a plant inversion: the 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 transfer function.

- When the plant is stable and has stable zeros (minimum-phase), the division can be made directly.
- One caveat for the stable-plant procedure is that lightly-damped poles or zeros should not be canceled 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 the peak 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.

13 MATH FACTS

13.1 Vectors

13.1.1 Definition

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:

$$\begin{aligned} \vec{a} + \vec{b} &= \vec{c} \\ \left\{ \begin{array}{c} 2 \\ 1 \\ 7 \end{array} \right\} + \left\{ \begin{array}{c} 3 \\ 3 \\ 2 \end{array} \right\} &= \left\{ \begin{array}{c} 5 \\ 4 \\ 9 \end{array} \right\}. \end{aligned}$$

   Graphically, addition is stringing the vectors together head to tail.

2. Scalar multiplication:

$$-2 \times \left\{ \begin{array}{c} 2 \\ 1 \\ 7 \end{array} \right\} = \left\{ \begin{array}{r} -4 \\ -2 \\ -14 \end{array} \right\}.$$

13.1.2 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.

13.1.3 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.

13.1.4 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,

$$\vec{x} \times \vec{y} \;\; = \begin{vmatrix} \hat{i} & \hat{j} & \hat{k} \\ x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{vmatrix} = \left\{ \begin{array}{c} (x_2 y_3 - x_3 y_2)\hat{i} \\ (x_3 y_1 - x_1 y_3)\hat{j} \\ (x_1 y_2 - x_2 y_1)\hat{k} \end{array} \right\}.$$

13.2 Matrices

13.2.1 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.

13.2.2 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 $\vec{v}$.

13.2.3 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.

13.2.4 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.

13.2.5 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,

$$\begin{aligned} \vec{a} = & \left\{ \begin{array}{c} 1 \\ 2 \\ 3 \end{array} \right\} \longrightarrow \vec{a}^{\,T} = & \{1 \;\, 2 \;\, 3\} \\ A = & \begin{bmatrix} 1 & 2 \\ 4 & 5 \\ 8 & 9 \end{bmatrix} \longrightarrow A^T = & \begin{bmatrix} 1 & 4 & 8 \\ 2 & 5 & 9 \end{bmatrix}. \end{aligned}$$

A very useful property of the transpose is

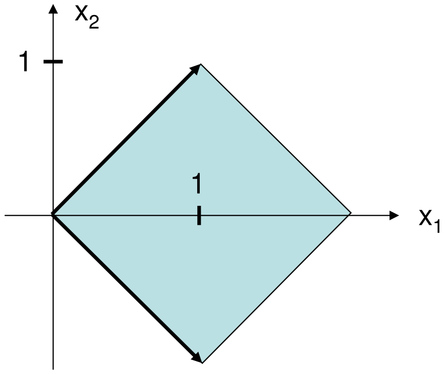
$$(AB)^T = B^T A^T.$$

13.2.6 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:

$$det(A) \;=\; A_{11}A_{22} - A_{21}A_{12}$$

$$det\left(\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\right) \quad = \quad 1+1=2.$$



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.

13.2.7 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

$$A\vec{x} = \vec{b} \longrightarrow \vec{x} = A^{-1}\vec{b}.$$

13.2.8 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:

$$\begin{aligned}
(A^T - \lambda_1 I)\vec{y}_1 &= \vec{0} \longrightarrow \\
\begin{bmatrix} 5 & 2 \\ -5 & -2 \end{bmatrix} \vec{y}_1 &= \vec{0} \longrightarrow \\
\vec{y}_1 &= \left\{ 2\sqrt{29}/29, \ \ -5\sqrt{29}/29 \right\}^T
\end{aligned}$$

$$\begin{aligned}
(A^T - \lambda_2 I)\vec{y}_2 &= \vec{0} \longrightarrow \\
\begin{bmatrix} 2 & 2 \\ -5 & -5 \end{bmatrix} \vec{y}_2 &= \vec{0} \longrightarrow \\
\vec{y}_2 &= \left\{ \sqrt{2}/2, \ \ -\sqrt{2}/2 \right\}^T .
\end{aligned}$$

13.2.9 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.[3]

[3] 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.