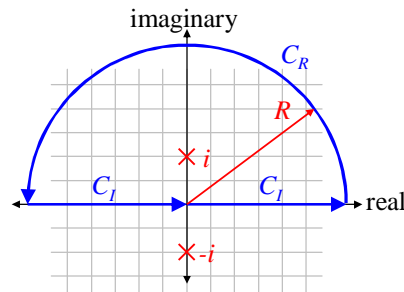
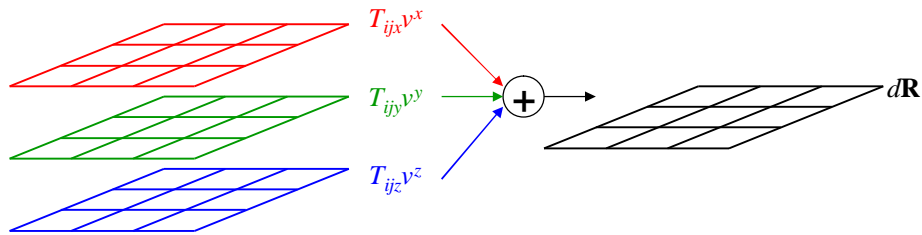


# Funky Mathematical Physics Concepts

The Anti-Textbook\*

A Work In Progress. See [physics.ucsd.edu/~emichels](http://physics.ucsd.edu/~emichels) for the latest versions of the Funky Series.  
Please send me comments.

Eric L. Michelsen



“I study mathematics to learn how to think.  
I study physics to have something to think about.”

“Perhaps the greatest irony of all is not that the square root of two is irrational, but that Pythagoras himself was irrational.”

\* Physical, conceptual, geometric, and pictorial physics that didn't fit in your textbook.

Please do NOT distribute this document. Instead, link to [physics.ucsd.edu/~emichels/FunkyMathPhysics.pdf](http://physics.ucsd.edu/~emichels/FunkyMathPhysics.pdf).  
Please cite as: Michelsen, Eric L., *Funky Mathematical Physics Concepts*, [physics.ucsd.edu/~emichels](http://physics.ucsd.edu/~emichels), 5/7/2015.

2006 values from NIST. For more physical constants, see <http://physics.nist.gov/cuu/Constants/> .

Speed of light in vacuum	$c = 299\,792\,458\text{ m s}^{-1}$ (exact)
Boltzmann constant	$k = 1.380\,6504(24) \times 10^{-23}\text{ J K}^{-1}$
Stefan-Boltzmann constant	$\sigma = 5.670\,400(40) \times 10^{-8}\text{ W m}^{-2}\text{ K}^{-4}$
Relative standard uncertainty	$\pm 7.0 \times 10^{-6}$
Avogadro constant	$N_A, L = 6.022\,141\,79(30) \times 10^{23}\text{ mol}^{-1}$
Relative standard uncertainty	$\pm 5.0 \times 10^{-8}$
Molar gas constant	$R = 8.314\,472(15)\text{ J mol}^{-1}\text{ K}^{-1}$
Electron mass	$m_e = 9.109\,382\,15(45) \times 10^{-31}\text{ kg}$
Proton mass	$m_p = 1.672\,621\,637(83) \times 10^{-27}\text{ kg}$
Proton/electron mass ratio	$m_p/m_e = 1836.152\,672\,47(80)$
Elementary charge	$e = 1.602\,176\,487(40) \times 10^{-19}\text{ C}$
Electron g-factor	$g_e = -2.002\,319\,304\,3622(15)$
Proton g-factor	$g_p = 5.585\,694\,713(46)$
Neutron g-factor	$g_N = -3.826\,085\,45(90)$
Muon mass	$m_\mu = 1.883\,531\,30(11) \times 10^{-28}\text{ kg}$
Inverse fine structure constant	$\alpha^{-1} = 137.035\,999\,679(94)$
Planck constant	$h = 6.626\,068\,96(33) \times 10^{-34}\text{ J s}$
Planck constant over $2\pi$	$\hbar = 1.054\,571\,628(53) \times 10^{-34}\text{ J s}$
Bohr radius	$a_0 = 0.529\,177\,208\,59(36) \times 10^{-10}\text{ m}$
Bohr magneton	$\mu_B = 927.400\,915(23) \times 10^{-26}\text{ J T}^{-1}$

### Reviews

“... most excellent tensor paper.... I feel I have come to a deep and abiding understanding of relativistic tensors.... The best explanation of tensors seen anywhere!” -- physics graduate student

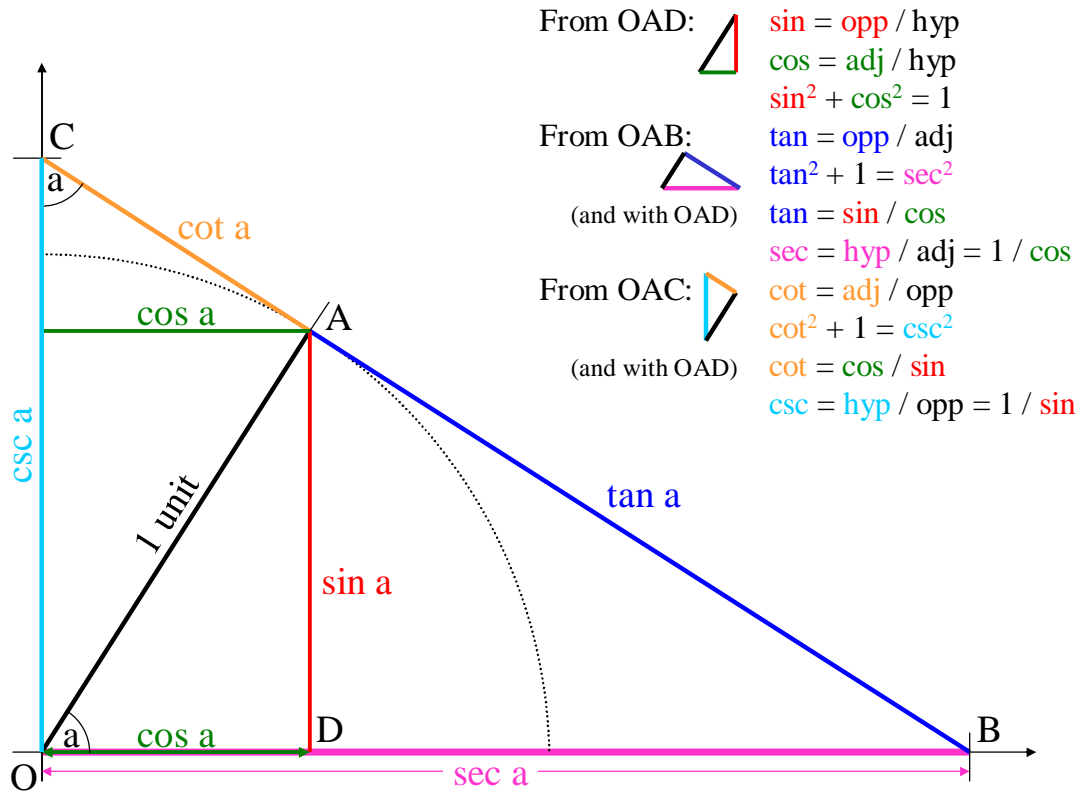
<h1>Contents</h1>
-------------------

<b>1</b>	<b>Introduction .....</b>	<b>8</b>
	Why Funky?.....	8
	How to Use This Document.....	8
	Why Physicists and Mathematicians Dislike Each Other .....	8
	Thank You .....	8
	Scope.....	8
	Notation.....	9
<b>2</b>	<b>Random Topics .....</b>	<b>12</b>
	What’s Hyperbolic About Hyperbolic Sine?.....	12
	Basic Calculus You May Not Know .....	13
	The Product Rule .....	15
	Integration By Pictures .....	15
	Theoretical Importance of IBP .....	16
	Delta Function Surprise.....	17
	Spherical Harmonics Are Not Harmonics .....	19
	The Binomial Theorem for Negative and Fractional Exponents .....	20
	When Does a Divergent Series Converge?.....	21
	Algebra Family Tree .....	22
	Convoluted Thinking.....	23
<b>3</b>	<b>Vectors .....</b>	<b>25</b>
	Small Changes to Vectors.....	25
	Why $(r, \theta, \phi)$ Are <i>Not</i> the Components of a Vector.....	25
	Laplacian’s Place .....	26
	Vector Dot Grad Vector .....	34
<b>4</b>	<b>Green’s Functions.....</b>	<b>36</b>
<b>5</b>	<b>Complex Analytic Functions.....</b>	<b>48</b>
	Residues.....	49
	Contour Integrals.....	50
	Evaluating Integrals.....	50
	Choosing the Right Path: Which Contour?.....	52
	Evaluating Infinite Sums .....	58
	Multi-valued Functions.....	60
<b>6</b>	<b>Conceptual Linear Algebra .....</b>	<b>61</b>
	Matrix Multiplication .....	61
	Determinants.....	62
	Cramer’s Rule.....	63
	Area and Volume as a Determinant.....	64
	The Jacobian Determinant and Change of Variables.....	65
	Expansion by Cofactors.....	67
	Proof That the Determinant Is Unique.....	69
	Getting Determined.....	70
	Advanced Matrices.....	71
	Getting to Home Basis .....	71
	Diagonalizing a Self-Adjoint Matrix.....	72
	Contraction of Matrices.....	74
	Trace of a Product of Matrices.....	74
	Linear Algebra Briefs .....	75
<b>7</b>	<b>Probability, Statistics, and Data Analysis.....</b>	<b>76</b>
	Probability and Random Variables.....	76
	Precise Statement of the Question Is Critical.....	77
	How to Lie With Statistics.....	78
	Choosing Wisely: An Informative Puzzle .....	78
	Multiple Events.....	79

Combining Probabilities .....	80
To B, or To Not B? .....	82
Continuous Random Variables and Distributions .....	83
Population and Samples .....	84
Population Variance .....	84
Population Standard Deviation .....	85
New Random Variables From Old Ones .....	86
Some Distributions Have Infinite Variance, or Infinite Average .....	87
Samples and Parameter Estimation .....	88
Why Do We Use Least Squares, and Least Chi-Squared ( $\chi^2$ )? .....	88
Average, Variance, and Standard Deviation .....	89
Functions of Random Variables.....	92
Statistically Speaking: What Is The Significance of This? .....	92
Predictive Power: Another Way to Be Significant, but Not Important.....	95
Unbiased vs. Maximum-Likelihood Estimators.....	96
Correlation and Dependence.....	98
Independent Random Variables are Uncorrelated.....	99
Statistical Analysis Algebra.....	100
The Average of a Sum: Easy?.....	100
The Average of a Product.....	100
Variance of a Sum.....	100
Covariance Revisited.....	101
Capabilities and Limits of the Sample Variance .....	101
How to Do Statistical Analysis Wrong, and How to Fix It.....	103
Introduction to Data Fitting (Curve Fitting).....	105
Goodness of Fit.....	106
Linear Regression.....	110
Review of Multiple Linear Regression.....	110
We Fit to the Predictors, <i>Not</i> the Independent Variable .....	111
The Sum-of-Squares Identity .....	112
The Raw Sum-of-Squares Identity .....	114
The Geometric View of a Least-Squares Fit.....	115
Algebra and Geometry of the Sum-of-Squares Identity .....	116
The ANOVA Sum-of-Squares Identity .....	117
The Failure of the ANOVA Sum-of-Squares Identity.....	118
Subtracting DC Before Analysis.....	118
Fitting to Orthonormal Functions.....	119
Hypothesis Testing with the Sum of Squares Identity.....	119
Introduction to Analysis of Variance (ANOVA) .....	120
The Temperature of Liberty.....	120
The F-test: The Decider for Zero Mean Gaussian Noise .....	124
Coefficient of Determination and Correlation Coefficient.....	125
Uncertainty Weighted Data.....	127
Be Sure of Your Uncertainty .....	128
Average of Uncertainty Weighted Data.....	128
Variance and Standard Deviation of Uncertainty Weighted Data .....	130
Normalized weights .....	132
Numerically Convenient Weights .....	132
Transformation to Equivalent Homoskedastic Measurements .....	133
Linear Regression with Individual Uncertainties .....	134
Linear Regression With Uncertainties and the Sum-of-Squares Identity .....	136
Hypothesis Testing a Model in Linear Regression with Uncertainties .....	140
Fitting To Histograms.....	140
Guidance Counselor: Practical Considerations for Computer Code to Fit Data.....	144
<b>8 Numerical Analysis.....</b>	<b>147</b>
Round-Off Error, And How to Reduce It .....	147

How To Extend Precision In Sums Without Using Higher Precision Variables .....	148
Numerical Integration.....	149
Sequences of Real Numbers .....	149
Root Finding.....	149
Simple Iteration Equation .....	149
Newton-Raphson Iteration.....	151
Pseudo-Random Numbers .....	153
Generating Gaussian Random Numbers.....	154
Generating Poisson Random Numbers.....	155
Generating Weirder Random Numbers .....	156
Exact Polynomial Fits.....	156
Two's Complement Arithmetic.....	158
How Many Digits Do I Get, 6 or 9? .....	159
How many digits do I need? .....	160
How Far Can I Go? .....	160
Software Engineering.....	160
Object Oriented Programming .....	161
The Best of Times, the Worst of Times.....	162
Matrix Addition .....	162
Memory Consumption vs. Run Time .....	166
Cache Withdrawal: Matrix Multiplication.....	167
Cache Summary.....	169
IEEE Floating Point Formats And Concepts.....	169
Precision in Decimal Representation.....	177
Underflow.....	178
<b>9 Fourier Transforms and Digital Signal Processing.....</b>	<b>184</b>
Model of Digitization and Sampling .....	185
Complex Sequences and Complex Fourier Transform.....	185
Basis Functions and Orthogonality .....	188
Real Sequences .....	189
Normalization and Parseval's Theorem.....	190
Continuous and Discrete, Finite and Infinite .....	191
White Noise and Correlation .....	192
Why Oversampling Does Not Improve Signal-to-Noise Ratio .....	192
Filters TBS?? .....	192
What Happens to a Sine Wave Deferred?.....	193
Nonuniform Sampling and Arbitrary Basis Functions .....	195
Don't Pad Your Data, Even for FFTs.....	196
Two Dimensional Fourier Transforms .....	197
Note on Continuous Fourier Series and Uniform Convergence .....	197
Fourier Transforms, Periodograms, and Lomb-Scargle.....	198
The Discrete Fourier Transform vs. the Periodogram .....	199
Practical Considerations .....	200
The Lomb-Scargle Algorithm.....	201
The Meaning Behind the Math .....	202
Bandwidth Correction (aka Bandwidth Penalty).....	206
Analytic Signals and Hilbert Transforms.....	209
Summary .....	214
<b>10 Tensors, Without the Tension.....</b>	<b>216</b>
Approach .....	216
Two Physical Examples.....	216
Magnetic Susceptibility .....	216
Mechanical Strain .....	220
When Is a Matrix Not a Tensor? .....	222
Heading In the Right Direction .....	222
Some Definitions and Review.....	222

Vector Space Summary .....	223
When Vectors Collide .....	224
“Tensors” vs. “Symbols” .....	225
Notational Nightmare .....	225
Tensors? What Good Are They? .....	225
A Short, Complicated Definition .....	225
Building a Tensor .....	226
Tensors in Action .....	227
Tensor Fields .....	229
Dot Products and Cross Products as Tensors .....	229
The Danger of Matrices .....	230
Reading Tensor Component Equations .....	230
Adding, Subtracting, Differentiating Tensors .....	231
Higher Rank Tensors .....	232
Tensors In General .....	233
Change of Basis: Transformations .....	234
Matrix View of Basis Transformation .....	235
Non-Orthonormal Systems: Contravariance and Covariance .....	235
What Goes Up Can Go Down: Duality of Contravariant and Covariant Vectors .....	238
The <i>Real</i> Summation Convention .....	239
Transformation of Covariant Indexes .....	239
Indefinite Metrics: Relativity .....	239
Is a Transformation Matrix a Tensor? .....	240
How About the Pauli Vector? .....	240
Cartesian Tensors .....	241
The Real Reason Why the Kronecker Delta Is Symmetric .....	242
Tensor Appendices .....	242
Pythagorean Relation for 1-forms .....	242
Geometric Construction Of The Sum Of Two 1-Forms: .....	243
“Fully Anti-symmetric” Symbols Expanded .....	244
Metric? We Don’t Need No Stinking Metric! .....	245
References: .....	247
<b>11 Differential Geometry .....</b>	<b>248</b>
Manifolds .....	248
Coordinate Bases .....	248
Covariant Derivatives .....	250
Christoffel Symbols .....	252
Visualization of n-Forms .....	253
Review of Wedge Products and Exterior Derivative .....	253
1-D .....	253
2-D .....	253
3-D .....	254
<b>12 Math Tricks .....</b>	<b>256</b>
Math Tricks That Come Up A Lot .....	256
The Gaussian Integral .....	256
Math Tricks That Are Fun and Interesting .....	256
Phasors .....	257
Future Funky Mathematical Physics Topics .....	257
<b>13 Appendices .....</b>	<b>258</b>
References .....	258
Glossary .....	258
Formulas .....	262
Index .....	263



Copyright 2001 Inductive Logic. All rights reserved.

# 1 Introduction

## Why Funky?

The purpose of the “Funky” series of documents is to help develop an accurate physical, conceptual, geometric, and pictorial understanding of important physics topics. We focus on areas that don’t seem to be covered well in most texts. The Funky series attempts to clarify those neglected concepts, and others that seem likely to be challenging and unexpected (funky?). The Funky documents are intended for serious students of physics; they are not “popularizations” or oversimplifications.

Physics includes math, and we’re not shy about it, but we also don’t hide behind it.

Without a conceptual understanding, math is gibberish.

This work is one of several aimed at graduate and advanced-undergraduate physics students. Go to <http://physics.ucsd.edu/~emichels> for the latest versions of the Funky Series, and for contact information. We’re looking for feedback, so please let us know what you think.

## How to Use This Document

This work is not a text book.

There are plenty of those, and they cover most of the topics quite well. This work is meant to be used *with* a standard text, to help emphasize those things that are most confusing for new students. When standard presentations don’t make sense, come here.

You should read all of this introduction to familiarize yourself with the notation and contents. After that, this work is meant to be read in the order that most suits you. Each section stands largely alone, though the sections are ordered logically. Simpler material generally appears before more advanced topics. You may read it from beginning to end, or skip around to whatever topic is most interesting. The “Shorts” chapter is a diverse set of very short topics, meant for quick reading.

If you don’t understand something, read it again once, then keep reading.  
Don’t get stuck on one thing. Often, the following discussion will clarify things.

The index is not yet developed, so go to the web page on the front cover, and text-search in this document.

## Why Physicists and Mathematicians Dislike Each Other

Physics goals and mathematics goals are antithetical. Physics seeks to ascribe meaning to mathematics that describe the world, to “understand” it, physically. Mathematics seeks to strip the equations of all physical meaning, and view them in purely abstract terms. These divergent goals set up a natural conflict between the two camps. Each goal has its merits: the value of physics is (or should be) self-evident; the value of mathematical abstraction, separate from any single application, is generality: the results can be used on a wide range of applications.

## Thank You

I owe a big thank you to many professors at both SDSU and UCSD, for their generosity even when I wasn’t a real student: Dr. Herbert Shore, Dr. Peter Salamon, Dr. Arlette Baljon, Dr. Andrew Cooksy, Dr. George Fuller, Dr. Tom O’Neil, Dr. Terry Hwa, and others.

## Scope

### What This Text Covers

This text covers some of the unusual or challenging concepts in graduate mathematical physics. It is also very suitable for upper-division undergraduate level, as well. We expect that you are taking or have



taken such a course, and have a good text book. *Funky Mathematical Physics Concepts* supplements those other sources.

**What This Text Doesn't Cover**

This text is not a mathematical physics course in itself, nor a review of such a course. We do not cover all basic mathematical concepts; only those that are very important, unusual, or especially challenging (funky?).

**What You Already Know**

This text assumes you understand basic integral and differential calculus, and partial differential equations. Further, it assumes you have a mathematical physics text for the bulk of your studies, and are using *Funky Mathematical Physics Concepts* to supplement it.

**Notation**

Sometimes the variables are inadvertently not written in italics, but I hope the meanings are clear.

?? refers to places that need more work.

TBS To be supplied (one hopes) in the future.

Interesting points that you may skip are “asides,” shown in smaller font and narrowed margins. Notes to myself may also be included as asides.

Common misconceptions are sometimes written in dark red dashed-line boxes.

**Formulas:** We write the integral over the entire domain as a subscript “∞”, for any number of dimensions:

$$1\text{-D: } \int_{-\infty}^{\infty} dx \quad 3\text{-D: } \int_{-\infty}^{\infty} d^3x$$

Evaluation between limits: we use the notation  $[function]_a^b$  to denote the evaluation of the function between  $a$  and  $b$ , i.e.,

$$[f(x)]_a^b \equiv f(b) - f(a). \quad \text{For example, } \int_0^1 3x^2 dx = [x^3]_0^1 = 1^3 - 0^3 = 1.$$

We write the probability of an event as “Pr(event).”

**Column vectors:** Since it takes a lot of room to write column vectors, but it is often important to distinguish between column and row vectors, I sometimes save vertical space by using the fact that a column vector is the transpose of a row vector:

$$\begin{pmatrix} a \\ b \\ c \\ d \end{pmatrix} = (a, b, c, d)^T$$

**Random variables:** We use a capital letter, e.g.  $X$ , to represent the *population* from which instances of a random variable,  $x$  (lower case), are observed. In a sense,  $X$  is a representation of the PDF of the random variable,  $\text{pdf}_X(x)$ .

We denote that a random variable  $X$  comes from a population PDF as:  $X \in \text{pdf}_X$ , e.g.:  $X \in \chi^2_n$ . To denote that  $X$  is a constant times a random variable from  $\text{pdf}_Y$ , we write:  $X \in k \text{ pdf}_Y$ , e.g.  $X \in k \chi^2_n$ .

For Greek letters, pronunciations, and use, see *Funky Quantum Concepts*. Other math symbols:

**Symbol Definition**

∀	for all
∃	there exists

$\ni$	such that
iff	if and only if
$\propto$	proportional to. E.g., $a \propto b$ means “ $a$ is proportional to $b$ ”
$\perp$	perpendicular to
$\therefore$	therefore
$\sim$	of the order of (sometimes used imprecisely as “approximately equals”)
$\equiv$	is defined as; identically equal to (i.e., equal in all cases)
$\Rightarrow$	implies
$\rightarrow$	leads to
$\otimes$	tensor product, aka outer product
$\oplus$	direct sum

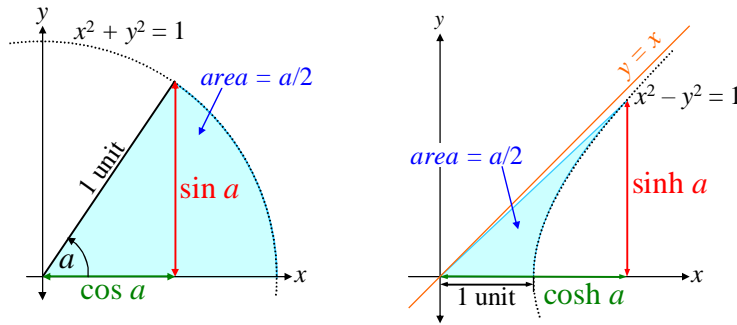
In mostly older texts, German type (font: Fraktur) is used to provide still more variable names:

Latin	German Capital	German Lowercase	Notes
A	Ⓐ	ⓐ	Distinguish capital from U, V
B	Ⓑ	ⓑ	
C	Ⓒ	ⓒ	Distinguish capital from E, G
D	Ⓓ	ⓓ	Distinguish capital from O, Q
E	Ⓔ	ⓔ	Distinguish capital from C, G
F	Ⓕ	ⓕ	
G	Ⓖ	ⓖ	Distinguish capital from C, E
H	Ⓖ	ⓗ	
I	Ⓘ	ⓓ	Capital almost identical to J
J	Ⓙ	ⓓ	Capital almost identical to I
K	Ⓚ	ⓕ	
L	Ⓛ	ⓓ	
M	Ⓜ	ⓓ	Distinguish capital from W
N	Ⓝ	ⓓ	
O	Ⓞ	ⓓ	Distinguish capital from D, Q
P	Ⓟ	ⓓ	
Q	Ⓠ	ⓓ	Distinguish capital from D, O
R	Ⓡ	ⓓ	Distinguish lowercase from x
S	Ⓢ	ⓓ	Distinguish capital from C, G, E

T	$\mathcal{T}$	t	Distinguish capital from I
U	$\mathcal{U}$	u	Distinguish capital from A, V
V	$\mathcal{V}$	v	Distinguish capital from A, U
W	$\mathcal{W}$	w	Distinguish capital from M
X	$\mathcal{X}$	x	Distinguish lowercase from r
Y	$\mathcal{Y}$	y	
Z	$\mathcal{Z}$	z	

## 2 Random Topics

### What's Hyperbolic About Hyperbolic Sine?



From where do the hyperbolic trigonometric functions get their names? By analogy with the circular functions. We usually think of the argument of circular functions as an angle,  $a$ . But in a unit circle, the area covered by the angle  $a$  is  $a/2$  (above left):

$$area = \frac{a}{2\pi} \pi r^2 = \frac{a}{2} \quad (r = 1).$$

Instead of the unit circle,  $x^2 + y^2 = 1$ , we can consider the area bounded by the  $x$ -axis, the ray from the origin, and the unit hyperbola,  $x^2 - y^2 = 1$  (above right). Then the  $x$  and  $y$  coordinates on the curve are called the **hyperbolic cosine** and **hyperbolic sine**, respectively. Notice that the hyperbola equation implies the well-known hyperbolic identity:

$$x = \cosh a, \quad y = \sinh a, \quad x^2 - y^2 = 1 \quad \Rightarrow \quad \cosh^2 - \sinh^2 = 1.$$

Proving that the area bounded by the  $x$ -axis, ray, and hyperbola satisfies the standard definition of the hyperbolic functions requires evaluating an elementary, but tedious, integral: (?? is the following right?)

$$area = \frac{a}{2} = \frac{1}{2} xy - \int_1^x y \, dx \quad \text{Use: } y = \sqrt{x^2 - 1}$$

$$a = x\sqrt{x^2 - 1} - 2 \int_1^x \sqrt{x^2 - 1} \, dx$$

For the integral, let  $x = \sec \theta, \quad dx = \tan \theta \sec \theta \, d\theta \quad \Rightarrow \quad y = \sqrt{\sec^2 \theta - 1} = \tan \theta$

$$\int_1^x \sqrt{x^2 - 1} \, dx = \int_1^x \sqrt{\sec^2 \theta - 1} \tan \theta \sec \theta \, d\theta = \int_1^x \tan^2 \theta \sec \theta \, d\theta = \int_1^x \frac{\sin^2 \theta}{\cos^3 \theta} \, d\theta$$

We try integrating by parts (but fail):

~~$$U = \tan \theta \quad dV = \sec \theta \tan \theta \, d\theta \quad \Rightarrow \quad dU = \sec^2 \theta \, d\theta, \quad V = \sec \theta$$~~

~~$$\int_1^x \tan^2 \theta \sec \theta \, d\theta = UV - \int V \, dU = \sec \theta \tan \theta \Big|_1^x - \int_1^x \sec^3 \theta \, d\theta$$~~

This is too hard, so we try reverting to fundamental functions  $\sin(\ )$  and  $\cos(\ )$ :

$$\begin{aligned}
 U = \sin \theta \quad dV = \cos^{-3} \theta \sin \theta \, d\theta \quad \Rightarrow \quad dU = \cos \theta \, d\theta, \quad V = \frac{1}{2} \cos^{-2} \theta \\
 2 \int_1^x \frac{\sin^2 \theta}{\cos^3 \theta} \, d\theta = 2UV - 2 \int V \, dU = \frac{\sin \theta}{\cos^2 \theta} \Big|_1^x - \int_1^x \cos^{-2} \theta \cos \theta \, d\theta \quad \text{Use: } \frac{\sin \theta}{\cos^2 \theta} \Big|_1^x = \sec \theta \tan \theta = xy \\
 = xy - \int_1^x \sec \theta \, d\theta = xy - (\ln |\sec \theta + \tan \theta|) \Big|_1^x = xy - \left( \ln \left| x + \sqrt{x^2 - 1} \right| \right) \Big|_1^x \\
 = xy - \ln \left| x + \sqrt{x^2 - 1} \right| - \ln 1 \\
 a = xy - xy + \ln \left| x + \sqrt{x^2 - 1} \right| = \ln \left| x + \sqrt{x^2 - 1} \right| \\
 e^a = x + \sqrt{x^2 - 1}
 \end{aligned}$$

Solve for  $x$  in terms of  $a$ , by squaring both sides:

$$e^{2a} = x^2 + 2x\sqrt{x^2 - 1} + x^2 - 1 = 2x \underbrace{\left( x + \sqrt{x^2 - 1} \right)}_{e^a} - 1 = 2xe^a - 1$$

$$e^{2a} + 1 = 2xe^a$$

$$e^a + e^{-a} = 2x \quad \Rightarrow \quad x \equiv \cosh a = \frac{(e^a + e^{-a})}{2}$$

The definition for sinh follows immediately from:

$$\cosh^2 - \sinh^2 = x^2 - y^2 = 1 \Rightarrow y = \sqrt{x^2 - 1}$$

$$\sinh a \equiv y = \sqrt{\left( \frac{e^a + e^{-a}}{2} \right)^2 - 1} = \sqrt{\frac{e^{2a} + 2 + e^{-2a}}{4} - 1} = \sqrt{\frac{e^{2a} - 2 + e^{-2a}}{4}} = \sqrt{\frac{(e^a - e^{-a})^2}{4}} = \frac{e^a - e^{-a}}{2}$$

## Basic Calculus You May Not Know

Amazingly, many calculus courses never provide a precise definition of a “limit,” despite the fact that both of the fundamental concepts of calculus, derivatives and integrals, are defined as limits! So here we go:

Basic calculus relies on 4 major concepts:

1. Functions
2. Limits
3. Derivatives
4. Integrals

**1. Functions:** Briefly, (in real analysis) a **function** takes one or more real values as inputs, and produces one or more real values as outputs. The inputs to a function are called the **arguments**. The simplest case is a real-valued function of a real-valued argument e.g.,  $f(x) = \sin x$ . Mathematicians would write  $(f: R^1 \rightarrow R^1)$ , read “ $f$  is a map (or function) from the real numbers to the real numbers.” A function which produces more than one output may be considered a vector-valued function.

**2. Limits:** Definition of “limit” (for a real-valued function of a single argument,  $f: R^1 \rightarrow R^1$ ):

$L$  is the **limit** of  $f(x)$  as  $x$  approaches  $a$ , iff for every  $\varepsilon > 0$ , there exists a  $\delta (> 0)$  such that  $|f(x) - L| < \varepsilon$  whenever  $0 < |x - a| < \delta$ . In symbols:

$$L = \lim_{x \rightarrow a} f(x) \text{ iff } \forall \varepsilon > 0, \exists \delta \text{ such that } |f(x) - L| < \varepsilon \text{ whenever } 0 < |x - a| < \delta .$$

This says that the value of the function *at*  $a$  doesn't matter; in fact, most often the function is not defined at  $a$ . However, the behavior of the function *near*  $a$  is important. If you can make the function arbitrarily close to some number,  $L$ , by restricting the function's argument to a small neighborhood around  $a$ , then  $L$  is the limit of  $f$  as  $x$  approaches  $a$ .

Surprisingly, this definition also applies to complex functions of complex variables, where the absolute value is the usual complex magnitude.

**Example:** Show that  $\lim_{x \rightarrow 1} \frac{2x^2 - 2}{x - 1} = 4$ .

Solution: We prove the existence of  $\delta$  given any  $\varepsilon$  by computing the necessary  $\delta$  from  $\varepsilon$ . Note that for  $x \neq 1$ ,  $\frac{2x^2 - 2}{x - 1} = 2(x + 1)$ . The definition of a limit requires that

$$\left| \frac{2x^2 - 2}{x - 1} - 4 \right| < \varepsilon \text{ whenever } 0 < |x - 1| < \delta .$$

We solve for  $x$  in terms of  $\varepsilon$ , which will then define  $\delta$  in terms of  $\varepsilon$ . Since we don't care what the function is at  $x = 1$ , we can use the simplified form,  $2(x + 1)$ . When  $x = 1$ , this is 4, so we suspect the limit = 4. Proof:

$$|2(x+1) - 4| < \varepsilon \Rightarrow 2|(x+1) - 2| < \varepsilon \Rightarrow |x - 1| < \frac{\varepsilon}{2} \quad \text{or} \quad 1 - \frac{\varepsilon}{2} < x < 1 + \frac{\varepsilon}{2} .$$

So by setting  $\delta = \varepsilon/2$ , we construct the required  $\delta$  for any given  $\varepsilon$ . Hence, for every  $\varepsilon$ , there exists a  $\delta$  satisfying the definition of a limit.

**3. Derivatives:** Only now that we have defined a limit, can we define a **derivative**:

$$f'(x) \equiv \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} .$$

**4. Integrals:** A simplified definition of an **integral** is an infinite sum of areas under a function divided into equal subintervals (Figure 2.1, left):

$$\int_a^b f(x) dx \equiv \lim_{N \rightarrow \infty} \underbrace{\frac{b-a}{N}}_{\Delta x} \sum_{i=1}^N f\left( (b-a) \frac{i}{N} \right) \quad \text{(simplified definition)} .$$

For practical physics, this definition would be fine. For mathematical preciseness, the actual definition of an integral is the limit over *any possible set* of subintervals, so long as the maximum of the subinterval size goes to zero. This is called "the norm of the subdivision," written as  $\|\Delta x_i\|$ :

$$\int_a^b f(x) dx \equiv \lim_{\|\Delta x_i\| \rightarrow 0} \sum_{i=1}^N f(x_i) \Delta x_i \quad \text{(precise definition)} .$$



**Figure 2.1** (Left) Simplified definition of an integral as the limit of a sum of equally spaced samples. (Right) Precise definition requires convergence for arbitrary, but small, subdivisions.

Why do mathematicians require this more precise definition? It's to avoid bizarre functions, such as:  $f(x)$  is 1 if  $x$  is rational, and zero if irrational. This means  $f(x)$  toggles wildly between 1 and 0 an infinite number of times over any interval. However, with the simplified definition of an integral, the following is well defined:

$$\int_0^{3.14} f(x) dx = 3.14, \quad \text{but} \quad \int_0^{\pi} f(x) dx = 0 \quad (\text{with simplified definition of integral}).$$

But properly, and with the precise definition of an integral, both integrals are undefined. (There are other types of integrals defined, but they are beyond our scope.)

### The Product Rule

Given functions  $U(x)$  and  $V(x)$ , the product rule (aka the **Leibniz rule**) says that for differentials,

$$d(UV) = U dV + V dU .$$

This leads to integration by parts, which is mostly known as an integration tool, but it is also an important theoretical (analytic) tool, and the essence of Legendre transformations.

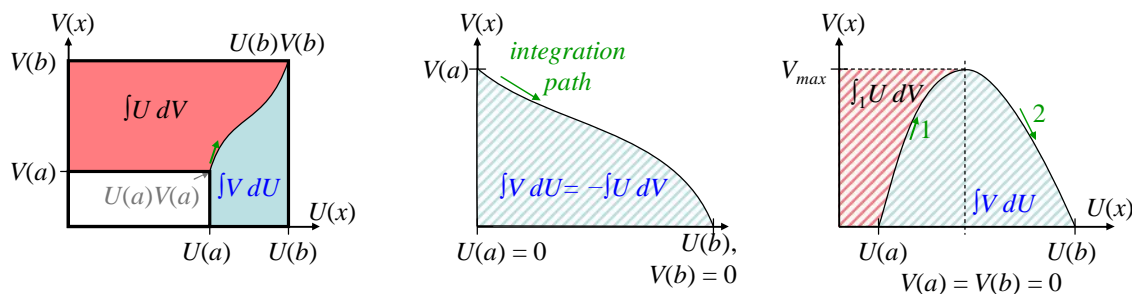
### Integration By Pictures

We assume you are familiar with integration by parts (IBP) as a tool for performing indefinite integrals, usually written as:

$$\int U dV = UV - \int V dU, \quad \text{which really means}$$

$$\int U(x) \underbrace{V'(x) dx}_{dV} = U(x)V(x) - \int V(x) \underbrace{U'(x) dx}_{dU}$$

This comes directly from the product rule above:  $U dV = d(UV) - V dU$ , and integrate both sides. Note that  $x$  is the integration variable (not  $U$  or  $V$ ), and  $x$  is also the parameter to the functions  $U(x)$  and  $V(x)$ .



**Figure 2.2** Three cases of integration by parts: (Left)  $U(x)$  and  $V(x)$  increasing. (Middle)  $V(x)$  decreasing to 0. (Right)  $V(x)$  progressing from zero, to finite, and back to zero.

The diagram above illustrates IBP in three cases. The left is the simplest case where  $U(x)$  and  $V(x)$  are monotonically increasing functions of  $x$  (note that  $x$  is not an axis,  $U$  and  $V$  are the axes, but  $x$  is the integration parameter). IBP says

$$\int_{x=a}^b U dV = \underbrace{[U(x)V(x)]_{x=a}^b}_{\text{boundary term}} - \int_{x=a}^b V dU = [U(b)V(b) - U(a)V(a)] - \int_{x=a}^b V dU .$$

The LHS (left hand side) of the equation is the red shaded area; the term in brackets on the right is the big rectangle minus the white rectangle; the last term is the blue shaded area. The left diagram illustrates IBP visually as areas. The term in brackets is called the **boundary term** (or “surface term”), because in some applications, it represents the part of the integral corresponding to the boundary (or surface) of the region of integration.

The middle diagram illustrates another common case: that in which the surface term  $UV$  is zero. In this case,  $UV = 0$  at  $x = a$  and  $x = b$ , because  $U(a) = 0$  and  $V(b) = 0$ . The shaded area is the integral, but the path of integration means that  $dU > 0$ , but  $dV < 0$ . Therefore  $\int V dU > 0$ , but  $\int U dV < 0$ .

The right diagram shows the case where one of  $U(x)$  or  $V(x)$  starts and ends at 0. For illustration, we chose  $V(a) = V(b) = 0$ . Then the surface term is zero, and we have:

$$[U(x)V(x)]_{x=a}^b = 0 \quad \Rightarrow \quad \int_{x=a}^b U dV = -\int_{x=a}^b V dU .$$

For  $V(x)$  to start and end at zero,  $V(x)$  must grow with  $x$  to some maximum,  $V_{max}$ , and then decrease back to 0. For simplicity, we assume  $U(x)$  is always increasing. The  $V dU$  integral is the blue striped area below the curve; the  $U dV$  integral is the area to the left of the curves. We break the  $dV$  integral into two parts: path 1, leading up to  $V_{max}$ , and path 2, going back down from  $V_{max}$  to zero. The integral from 0 to  $V_{max}$  (path 1) is the red striped area; the integral from  $V_{max}$  back down to 0 (path 2) is the negative of the entire (blue + red) striped area. Then the blue shaded region is the difference: (1) the (red) area to the left of path 1 (where  $dV$  is positive, because  $V(x)$  is increasing), minus (2) the (blue + red) area to the left of path 2, because  $dV$  is negative when  $V(x)$  is decreasing:

$$\begin{aligned} \int_{\text{path1+path2}} U dV &= \int_{V=0}^{V_{max}} U dV + \int_{V=V_{max}}^0 U dV = \int_{V=0}^{V_{max}} U dV - \int_{V=0}^{V_{max}} U dV \\ &= \int_{\text{path1}} U dV - \int_{\text{path2}} U dV \\ &= -\int_{x=a}^b V dU . \end{aligned}$$

### Theoretical Importance of IBP

Besides being an integration tool, an important theoretical consequence of IBP is that the variable of integration is changed, from  $dV$  to  $dU$ . Many times, one differential is unknown, but the other is known:



Under an integral, integration by parts allows one to exchange a derivative that cannot be directly evaluated, even in principle, in favor of one that can.

The classic example of this is deriving the Euler-Lagrange equations of motion from the principle of stationary action. The action of a dynamic system is defined by

$$S \equiv \int L(q(t), \dot{q}(t)) dt .$$

where the lagrangian is a given function of the trajectory  $q(t)$ . **Stationary action** means that the action does not change (to first order) for small changes in the trajectory. I.e., given a small variation in the trajectory,  $\delta q(t)$ :

$$\begin{aligned} \delta S = 0 &= \int L(q + \delta q, \dot{q} + \delta \dot{q}) dt - S = \int \left[ \frac{\partial L}{\partial q} \delta q + \frac{\partial L}{\partial \dot{q}} \delta \dot{q} \right] dt && \text{Use } \delta \dot{q} = \frac{d}{dt} \delta q \\ &= \int \left[ \frac{\partial L}{\partial q} \delta q + \frac{\partial L}{\partial \dot{q}} \frac{d}{dt} \delta q \right] dt . \end{aligned}$$

The quantity in brackets involves both  $\delta q(t)$  and its time derivative,  $\delta \dot{q}$ . We are free to vary  $\delta q(t)$  arbitrarily, but that fully determines  $\delta \dot{q}$ . We cannot vary both  $\delta q$  and  $\delta \dot{q}$  separately. We also know that  $\delta q(t) = 0$  at its endpoints, but  $\delta \dot{q}$  is unconstrained at its endpoints. Therefore, it would be simpler if the quantity in brackets was written entirely in terms of  $\delta q(t)$ , and not in terms of  $\delta \dot{q}$ . IBP allows us to eliminate the time derivative of  $\delta q(t)$  in favor of the time derivative of  $\partial L / \partial \dot{q}$ . Since  $L(q, \dot{q})$  is given, we can easily determine  $\partial L / \partial \dot{q}$ . Therefore, this is a good trade. Integrating the 2<sup>nd</sup> term in brackets by parts gives:

$$\begin{aligned} \text{Let } U &= \frac{\partial L}{\partial \dot{q}}, & dU &= \left( \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} \right) dt, & dV &= \frac{d}{dt} \delta q dt, & V &= \delta q \\ \int \underbrace{\frac{\partial L}{\partial \dot{q}}}_{U} \underbrace{\frac{d}{dt} \delta q}_{V'} dt &= UV - \int V dU = \left[ \frac{\partial L}{\partial \dot{q}} \delta q(t) \right]_{t=0}^{t=f} - \int \underbrace{\delta q}_{V} \underbrace{\left( \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} \right)}_{U'} dt \end{aligned}$$

The boundary term is zero because  $\delta q(t)$  is zero at both endpoints. The variation in action  $\delta S$  is now:

$$\delta S = \int \left[ \frac{\partial L}{\partial q} - \frac{d}{dt} \frac{\partial L}{\partial \dot{q}} \right] \delta q dt = 0 \quad \forall \delta q(t) .$$

The only way  $\delta S = 0$  can be satisfied for *any*  $\delta q(t)$  is if the quantity in brackets is identically 0. Thus IBP has lead us to an important *theoretical* conclusion: the Euler-Lagrange equation of motion.

This fundamental result has nothing to do with evaluating a specific difficult integral. IBP: it's not just for hard integrals any more.

### Delta Function Surprise

Rarely, one needs to consider the 3D  $\delta$ -function in coordinates other than rectangular. The 3D  $\delta$ -function is written  $\delta^3(\mathbf{r} - \mathbf{r}')$ . For example, in 3D Green's functions, whose definition depends on a  $\delta^3$ -function, it may be convenient to use cylindrical or spherical coordinates. In these cases, there are some unexpected consequences [Wyl p280]. This section assumes you understand the basic principle of a 1D and 3D  $\delta$ -function. (See the introduction to the delta function in Funky Quantum Concepts.)

Recall the defining property of  $\delta^3(\mathbf{r} - \mathbf{r}')$ :

$$\int_{\infty} d^3\mathbf{r} \delta^3(\mathbf{r} - \mathbf{r}') = 1 \quad \forall \mathbf{r}' \quad (\forall \equiv \text{"for all"}) \quad \Rightarrow \quad \int_{\infty} d^3\mathbf{r} \delta^3(\mathbf{r} - \mathbf{r}') f(\mathbf{r}) = f(\mathbf{r}') .$$

The above definition is “coordinate free,” i.e. it makes no reference to any choice of coordinates, and is true in *every* coordinate system. As with Green’s functions, it is often helpful to think of the  $\delta$ -function as a function of  $\mathbf{r}$ , which is zero everywhere except for an impulse located at  $\mathbf{r}'$ . As we will see, this means that it is properly a function of  $\mathbf{r}$  and  $\mathbf{r}'$  separately, and should be written as  $\delta^3(\mathbf{r}, \mathbf{r}')$  (like Green’s functions are).

**Rectangular coordinates:** In rectangular coordinates, however, we now show that we *can* simply break up  $\delta^3(x, y, z)$  into 3 components. By writing  $(\mathbf{r} - \mathbf{r}')$  in rectangular coordinates, and using the defining integral above, we get:

$$\begin{aligned} \mathbf{r} - \mathbf{r}' &\equiv (x - x', y - y', z - z') &\Rightarrow &\int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy \int_{-\infty}^{\infty} dz \delta^3(x - x', y - y', z - z') = 1 \\ & &\Rightarrow &\delta^3(x - x', y - y', z - z') = \delta(x - x')\delta(y - y')\delta(z - z'). \end{aligned}$$

In rectangular coordinates, the above shows that we *do* have translation invariance, so we can simply write:

$$\delta^3(x, y, z) = \delta(x)\delta(y)\delta(z).$$

In other coordinates, we do *not* have translation invariance. Recall the 3D infinitesimal volume element in 4 different systems: coordinate-free, rectangular, cylindrical, and spherical coordinates:

$$d^3\mathbf{r} = dx dy dz = r dr d\phi dz = r^2 \sin\theta dr d\theta d\phi.$$

The presence of  $r$  and  $\theta$  imply that when writing the 3D  $\delta$ -function in non-rectangular coordinates, we must include a pre-factor to maintain the defining integral = 1. We now show this explicitly.

**Cylindrical coordinates:** In cylindrical coordinates, for  $r > 0$ , we have (using the imprecise notation of [Wyl p280]):

$$\begin{aligned} \mathbf{r} - \mathbf{r}' &= (r - r', \phi - \phi', z - z') &\Rightarrow & \\ \int_0^{\infty} dr \int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz r \delta^3(r - r', \phi - \phi', z - z') &= 1 \\ \Rightarrow \delta^3(r - r', \phi - \phi', z - z') &= \frac{1}{r'} \delta(r - r')\delta(\phi - \phi')\delta(z - z'), r' > 0 \end{aligned}$$

Note the  $1/r'$  pre-factor on the RHS. This may seem unexpected, because the pre-factor depends on the location of  $\delta^3(\ )$  in space (hence, no radial translation invariance). The rectangular coordinate version of  $\delta^3(\ )$  has no such pre-factor. Properly speaking,  $\delta^3(\ )$  isn’t a function of  $r - r'$ ; it is a function of  $r$  and  $r'$  separately.

In non-rectangular coordinates,  $\delta^3(\ )$  does not have translation invariance, and includes a pre-factor which depends on the position of  $\delta^3(\ )$  in space, i.e. depends on  $\mathbf{r}'$ .

At  $r' = 0$ , the pre-factor blows up, so we need a different pre-factor. We’d like the defining integral to be 1, regardless of  $\phi$ , since all values of  $\phi$  are equivalent at the origin. This means we must drop the  $\delta(\phi - \phi')$ , and replace the pre-factor to cancel the constant we get when we integrate out  $\phi$ :

$$\begin{aligned} \int_0^{\infty} dr \int_0^{2\pi} d\phi \int_{-\infty}^{\infty} dz r \delta^3(r - r', \phi - \phi', z - z') &= 1, \quad r' = 0 \\ \Rightarrow \delta^3(r - r', \phi - \phi', z - z') &= \frac{1}{2\pi r} \delta(r)\delta(z - z'), \quad r' = 0, \\ \text{assuming that } \int_0^{\infty} dr \delta(r) &= 1. \end{aligned}$$

This last assumption is somewhat unusual, because the  $\delta$ -function is usually thought of as symmetric about 0, where the above radial integral would only be  $\frac{1}{2}$ . The assumption implies a “right-sided”  $\delta$ -function, whose entire non-zero part is located at  $0^+$ . Furthermore, notice the factor of  $1/r$  in  $\delta(r - 0, z - z')$ . This

factor blows up at  $r = 0$ , and has no effect when  $r \neq 0$ . Nonetheless, it is needed because the volume element  $r dr d\phi dz$  goes to zero as  $r \rightarrow 0$ , and the  $1/r$  in  $\delta(r - 0, z - z')$  compensates for that.

**Spherical coordinates:** In spherical coordinates, we have similar considerations. First, away from the origin,  $r' > 0$ :

$$\int_0^\infty dr \int_0^\pi d\theta \int_0^{2\pi} d\phi r^2 \sin\theta \delta^3(r - r', \theta - \theta', \phi - \phi') = 1 \Rightarrow$$

$$\delta^3(r - r', \theta - \theta', \phi - \phi') = \frac{1}{r'^2 \sin\theta'} \delta(r - r') \delta(\theta - \theta') \delta(\phi - \phi'), \quad r' > 0. \quad [\text{Wyl 8.9.2 p280}]$$

Again, the pre-factor depends on the position in space, and properly speaking,  $\delta^3(\ )$  is a function of  $r, r', \theta,$  and  $\theta'$  separately, not simply a function of  $r - r'$  and  $\theta - \theta'$ . At the origin, we'd like the defining integral to be 1, regardless of  $\phi$  or  $\theta$ . So we drop the  $\delta(\phi - \phi') \delta(\theta - \theta')$ , and replace the pre-factor to cancel the constant we get when we integrate out  $\phi$  and  $\theta$ :

$$\int_0^\infty dr \int_0^\pi d\theta \int_0^{2\pi} d\phi r^2 \sin\theta \delta^3(r - 0, \theta - \theta', \phi - \phi') = 1, \quad r' = 0$$

$$\Rightarrow \delta^3(r - 0, \theta - \theta', \phi - \phi') = \frac{1}{4\pi r^2} \delta(r), \quad r' = 0,$$

assuming that  $\int_0^\infty dr \delta(r) = 1.$

Again, this definition uses the modified  $\delta(r)$ , whose entire non-zero part is located at  $0^+$ . And similar to the cylindrical case, this includes the  $1/r^2$  factor to preserve the integral at  $r = 0$ .

**2D angular coordinates:** For 2D angular coordinates  $\theta$  and  $\phi$ , we have:

$$\int_0^\pi d\theta \int_0^{2\pi} d\phi \sin\theta \delta^2(\theta - \theta', \phi - \phi') = 1, \quad \theta' > 0$$

$$\Rightarrow \delta^2(\theta - \theta', \phi - \phi') = \frac{1}{\sin\theta'} \delta(\theta - \theta') \delta(\phi - \phi'), \quad \theta' > 0.$$

Once again, we have a special case when  $\theta' = 0$ : we must have the defining integral be 1 for any value of  $\phi$ . Hence, we again compensate for the  $2\pi$  from the  $\phi$  integral:

$$\int_0^\pi d\theta \int_0^{2\pi} d\phi \sin\theta \delta^2(\theta - \theta', \phi - \phi') = 1, \quad \theta' = 0$$

$$\Rightarrow \delta^2(\theta - 0, \phi - \phi') = \frac{1}{2\pi \sin\theta} \delta(\theta), \quad \theta' = 0.$$

Similar to the cylindrical and spherical cases, this includes a  $1/(\sin\theta)$  factor to preserve the integral at  $\theta = 0$ .

## Spherical Harmonics Are Not Harmonics

See *Funky Electromagnetic Concepts* for a full discussion of harmonics, Laplace's equation, and its solutions in 1, 2, and 3 dimensions. Here is a brief overview.

Spherical harmonics are the angular parts of solid harmonics, but we will show that they are not truly "harmonics." A **harmonic** is a function which satisfies Laplace's equation:

$$\nabla^2 \Phi(\mathbf{r}) = 0, \quad \text{with } \mathbf{r} \text{ typically in 2 or 3 dimensions.}$$

Solid harmonics are 3D harmonics: they solve Laplace's equation in 3 dimensions. For example, one form of solid harmonics separates into a product of 3 functions in spherical coordinates:

$$\Phi(r, \theta, \phi) = R(r)P(\theta)Q(\phi) = \left( A_l r^l + B_l r^{-(l+1)} \right) P_{lm}(\cos \theta) (C_l \sin m\phi + D_l \cos m\phi)$$

where  $R(r) = A_l r^l + B_l r^{-(l+1)}$  is the radial part,

$P(\theta) = P_{lm}(\cos \theta)$  is the polar angle part, the associated Legendre functions,

$Q(\phi) = (C_l \sin m\phi + D_l \cos m\phi)$  is the azimuthal part .

The spherical harmonics are just the angular ( $\theta, \phi$ ) parts of these solid harmonics. But notice that the angular part alone does not satisfy the 2D Laplace equation (i.e., on a sphere of fixed radius):

$$\begin{aligned} \nabla^2 &= \frac{1}{r^2} \frac{\partial}{\partial r} \left( r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \phi^2}, & \text{but for fixed } r: \\ &= \frac{1}{r^2} \left[ \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right]. \end{aligned}$$

However, direct substitution of spherical harmonics into the above Laplace operator shows that the result is *not* 0 (we let  $r = 1$ ). We proceed in small steps:

$$Q(\phi) = C \sin m\phi + D \cos m\phi \quad \Rightarrow \quad \frac{\partial^2}{\partial \phi^2} Q(\phi) = -m^2 Q(\phi).$$

For integer  $m$ , the **associated Legendre functions**,  $P_{lm}(\cos \theta)$ , satisfy, for given  $l$  and  $m$ :

$$\frac{1}{r^2 \sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial}{\partial \theta} \right) P_{lm}(\cos \theta) = \left( -\frac{l(l+1)}{r^2} + m^2 \right) P_{lm}(\cos \theta).$$

Combining these 2 results ( $r = 1$ ):

$$\begin{aligned} \nabla^2 (P(\theta)Q(\phi)) &= \left[ \frac{1}{\sin \theta} \frac{\partial}{\partial \theta} \left( \sin \theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2 \theta} \frac{\partial^2}{\partial \phi^2} \right] (P(\theta)Q(\phi)) \\ &= \left( -l(l+1) + m^2 \right) P_{lm}(\cos \theta) Q(\theta) - m^2 P_{lm}(\cos \theta) Q(\theta) \\ &= -l(l+1) P_{lm}(\cos \theta) Q(\theta) \end{aligned}$$

Hence, the spherical harmonics are *not* solutions of Laplace's equation, i.e. they are *not* "harmonics."

## The Binomial Theorem for Negative and Fractional Exponents

You may be familiar with the **binomial theorem** for positive integer exponents, but it is very useful to know that the binomial theorem also works for negative and fractional exponents. We can use this fact to easily find series expansions for things like  $\frac{1}{1-x}$  and  $\sqrt{1+x} = (1+x)^{1/2}$ .

First, let's review the simple case of positive integer exponents:

$$(a+b)^n = a^n b^0 + \frac{n}{1} a^{n-1} b^1 + \frac{n(n-1)}{1 \cdot 2} a^{n-2} b^2 + \frac{n(n-1)(n-2)}{1 \cdot 2 \cdot 3} a^{n-3} b^3 + \dots + \frac{n!}{n!} a^0 b^n.$$

[For completeness, we note that we can write the general form of the  $m^{\text{th}}$  term:

$$m^{\text{th}} \text{ term} = \frac{n!}{(n-m)!m!} a^{n-m} b^m, \quad n \text{ integer} > 0; \quad m \text{ integer}, 0 \leq m \leq n.]$$

But we're much more interested in the iterative procedure (recursion relation) for finding the  $(m + 1)^{\text{th}}$  term from the  $m^{\text{th}}$  term, because we use that to generate a power series expansion. The process is this:

1. The first term ( $m = 0$ ) is always  $a^n b^0 = a^n$ , with an implicit coefficient  $C_0 = 1$ .
2. To find  $C_{m+1}$ , multiply  $C_m$  by the power of  $a$  in the  $m^{\text{th}}$  term,  $(n - m)$ ,
3. divide it by  $(m + 1)$ , [the number of the new term we're finding]:
4. lower the power of  $a$  by 1 (to  $n - m$ ), and
5. raise the power of  $b$  by 1 to  $(m + 1)$ .

$$C_{m+1} = \frac{(n-m)}{m+1} C_m$$

This procedure is valid for all  $n$ , even negative and fractional  $n$ . A simple way to remember this is:

For any real  $n$ , we generate the  $(m + 1)^{\text{th}}$  term from the  $m^{\text{th}}$  term by differentiating with respect to  $a$ , and integrating with respect to  $b$ .

The general expansion, for any  $n$ , is then:

$$m^{\text{th}} \text{ term} = \frac{n(n-1)(n-2)\dots(n-m+1)}{m!} a^{n-m} b^m, \quad n \text{ real}; \quad m \text{ integer} \geq 0$$

Notice that for integer  $n > 0$ , there are  $n+1$  terms. For fractional or negative  $n$ , we get an infinite series.

**Example 1:** Find the Taylor series expansion of  $\frac{1}{1-x}$ . Since the Taylor series is unique, any method we use to find a power series expansion will give us the Taylor series. So we can use the binomial theorem, and apply the rules above, with  $a = 1$ ,  $b = (-x)$ :

$$\begin{aligned} \frac{1}{1-x} &= (1+(-x))^{-1} = 1^{-1} + \frac{(-1)}{1} 1^{-2} (-x)^1 + \frac{(-1)(-2)}{1 \cdot 2} 1^{-3} (-x)^2 + \frac{(-1)(-2)(-3)}{1 \cdot 2 \cdot 3} 1^{-4} (-x)^3 + \dots \\ &= 1 + x + x^2 + \dots + x^m + \dots \end{aligned}$$

Notice that all the fractions, all the powers of 1, and all the minus signs cancel.

**Example 2:** Find the Taylor series expansion of  $\sqrt{1+x} = (1+x)^{1/2}$ . The first term is  $a^{1/2} = 1^{1/2}$ :

$$\begin{aligned} (1+x)^{1/2} &= 1^{1/2} + \frac{1}{2} \frac{1}{(1)} 1^{-1/2} x^1 + \frac{1}{2} \left(-\frac{1}{2}\right) \frac{1}{(1 \cdot 2)} 1^{-3/2} x^2 + \frac{1}{2} \left(-\frac{1}{2}\right) \left(-\frac{3}{2}\right) \frac{1}{(1 \cdot 2 \cdot 3)} 1^{-5/2} x^3 + \dots \\ &= 1 + \frac{1}{2} x - \frac{1}{8} x^2 + \frac{3}{48} x^3 - \dots + (-1)^{m+1} \frac{(2m-3)!!}{2^m m!} x^m \end{aligned}$$

where  $p!! \equiv p(p-2)(p-4)\dots(2 \text{ or } 1)$

## When Does a Divergent Series Converge?

Consider the infinite series

$$1 + x + x^2 + \dots + x^n + \dots$$

When is it convergent? Apparently, when  $|x| < 1$ . What is the value of the series when  $x = 2$ ? "Undefined!" you say. But there is a very important sense in which the series converges for  $x = 2$ , and its value is  $-1$ ! How so?

Recall the Taylor expansion (you can use the binomial theorem, see earlier section):

$$\frac{1}{1-x} = (1-x)^{-1} = 1 + x + x^2 + \dots + x^n + \dots$$

It is exactly the original infinite series above. So the series sums to  $1/(1-x)$ . This is defined for all  $x \neq 1$ . And its value for  $x = 2$  is  $-1$ .

Why is this important? There are cases in physics when we use perturbation theory to find an expansion of a number in an infinite series. Sometimes, the series appears to diverge. But by finding the analytic expression corresponding to the series, we can evaluate the analytic expression at values of  $x$  that make the series diverge. In many cases, the analytic expression provides an important and meaningful answer to a perturbation problem. This happens in quantum mechanics, and quantum field theory.

This is an example of *analytic continuation*. A Taylor series is a special case of a Laurent series, and any function with a Laurent expansion is **analytic**. If we know the Laurent series (or if we know the values of an analytic function and all its derivatives at *any one* point), then we know the function everywhere, even for complex values of  $x$ . The original series is analytic around  $x = 0$ , therefore it is analytic everywhere it converges (everywhere it is defined). The process of extending a function which is defined in some small region to be defined in a much larger (even complex) region, is called **analytic continuation** (see Complex Analysis, discussed elsewhere in this document).

TBS: show that the sum of the integers  $1 + 2 + 3 + \dots = -1/12$ . ??

---

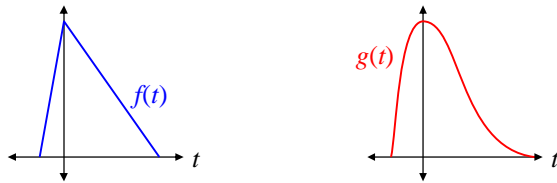
### Algebra Family Tree

Doodad	Properties	Examples
<b>group</b>	Finite or infinite set of elements and operator $(\cdot)$ , with closure, associativity, identity element and inverses. Possibly commutative: $a \cdot b = c$ w/ $a, b, c$ group elements	rotations of a square by $n \times 90^\circ$ continuous rotations of an object
<b>ring</b>	Set of elements and 2 binary operators (+ and *), with: • commutative <i>group</i> under + • left and right distributivity: $a(b + c) = ab + ac, (a + b)c = ac + bc$ • usually multiplicative associativity	integers mod $m$ polynomials $p(x) \text{ mod } m(x)$
<b>integral domain, or domain</b>	A <i>ring</i> , with: • commutative multiplication • multiplicative identity (but no inverses) • no zero divisors ( $\Rightarrow$ cancellation is valid): $ab = 0$ only if $a = 0$ or $b = 0$	integers polynomials, even abstract polynomials, with abstract variable $x$ , and coefficients from a “field”
<b>field</b>	“rings with multiplicative inverses (& identity)” • commutative <i>group</i> under addition • commutative <i>group</i> (excluding 0) under multiplication. • distributivity, multiplicative inverses Allows solving simultaneous linear equations. Field can be finite or infinite	integers with arithmetic modulo 3 (or any prime) real numbers complex numbers

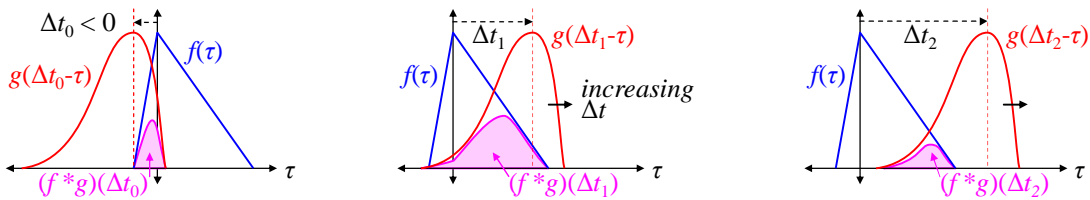
<b>vector space</b>	<ul style="list-style-type: none"> <li>• <i>field</i> of scalars</li> <li>• <i>group</i> of vectors under +.</li> </ul> Allows solving simultaneous vector equations for unknown scalars or vectors. Finite or infinite dimensional.	physical vectors real or complex functions of space: $f(x, y, z)$ kets (and bras)
<b>Hilbert space</b>	<i>vector space</i> over field of complex numbers with: <ul style="list-style-type: none"> <li>• a conjugate-bilinear inner product,  <math>\langle av bw\rangle = (a^*)b\langle v w\rangle</math>,  <math>\langle v w\rangle = \langle w v\rangle^*</math>  <math>a, b</math> scalars, and <math>v, w</math> vectors</li> <li>• Mathematicians require it to be infinite dimensional; physicists don't.</li> </ul>	real or complex functions of space: $f(x, y, z)$ quantum mechanical wave functions

### Convolved Thinking

Convolution arises in many physics, engineering, statistics, and other mathematical areas.



Two functions,  $f(t)$  and  $g(t)$ .



(Left)  $(f * g)(\Delta t_0)$ ,  $\Delta t_0 < 0$ . (Middle)  $(f * g)(\Delta t_1)$ ,  $\Delta t_1 > 0$ . (Right)  $(f * g)(\Delta t_2)$ ,  $\Delta t_2 > \Delta t_1$ . The convolution is the magenta shaded area.

Given two functions,  $f(t)$  and  $g(t)$ , the convolution of  $f(t)$  and  $g(t)$  is a function of a time-displacement,  $\Delta t$ , defined by (see diagram above):

$$(f * g)(\Delta t) \equiv \int_{-\infty}^{\infty} d\tau f(\tau)g(\Delta t - \tau) \text{ where the integral covers some domain of interest}$$

When  $\Delta t < 0$ , the two functions are “backing into each other” (above left). When  $\Delta t > 0$ , the two functions are “backing away from each other” (above middle and right).

Of course, we don't require functions of time. Convolution is useful with a variety of independent variables. E.g., for functions of space,  $f(x)$  and  $g(x)$ ,  $f * g(\Delta x)$  is a function of spatial displacement,  $\Delta x$ .

Notice that convolution is

(1) commutative:  $f * g = g * f$

(2) linear in each of the two functions:

$$f * kg = k(f * g) = (kf) * g, \quad \text{and}$$
$$f * (g + h) = f * g + f * h.$$

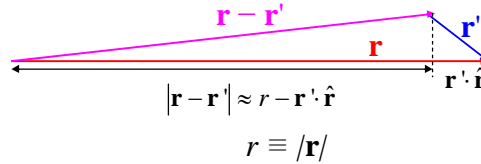
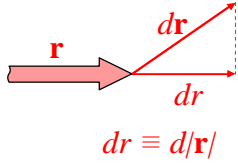
The verb “to convolve” means “to form the convolution of.” We convolve  $f$  and  $g$  to form the convolution  $f * g$ .



### 3 Vectors

#### Small Changes to Vectors

##### Projection of a Small Change to a Vector Onto the Vector



(Left) A small change to a vector, and its projection onto the vector.  
 (Right) Approximate magnitude of the difference between a big and small vector.

It is sometimes useful (in orbital mechanics, for example) to relate the change in a vector to the change in the vector's magnitude. The diagram above (left) leads to a somewhat unexpected result:

$$\mathbf{r} \cdot \hat{\mathbf{r}} = dr \quad \text{or} \quad (\text{multiplying both sides by } r \text{ and using } \mathbf{r} = r\hat{\mathbf{r}})$$

$$\mathbf{r} \cdot d\mathbf{r} = r dr$$

And since this is true for any small change, it is also true for any rate of change (just divide by  $dt$ ):

$$\mathbf{r} \cdot \dot{\mathbf{r}} = r \dot{r}$$

#### Vector Difference Approximation

It is sometimes useful to approximate the magnitude of a large vector minus a small one. (In electromagnetics, for example, this is used to compute the far-field from a small charge or current distribution.) The diagram above (right) shows that:

$$|\mathbf{r} - \mathbf{r}'| \approx |\mathbf{r}| - \mathbf{r}' \cdot \hat{\mathbf{r}}, \quad |\mathbf{r}| \gg |\mathbf{r}'|$$

#### Why $(r, \theta, \phi)$ Are Not the Components of a Vector

$(r, \theta, \phi)$  are *parameters* of a vector, but not *components*. That is, the parameters  $(r, \theta, \phi)$  uniquely define the vector, but they are not components, because you can't add them. This is important in much physics, e.g. involving magnetic dipoles (ref Jac problem on mag dipole field). **Components** of a vector are *defined* as coefficients of basis vectors. For example, the components  $\mathbf{v} = (x, y, z)$  can multiply the basis vectors to construct  $\mathbf{v}$ :

$$\mathbf{v} = x\hat{\mathbf{x}} + y\hat{\mathbf{y}} + z\hat{\mathbf{z}}$$

There is no similar equation we can write to construct  $\mathbf{v}$  from its spherical components  $(r, \theta, \phi)$ . Position vectors are displacements from the origin, and there are no  $\hat{\mathbf{r}}, \hat{\theta}, \hat{\phi}$  defined at the origin.

Put another way, you can always add the components of two vectors to get the vector sum:

Let  $\mathbf{w} = (a, b, c)$  rectangular components. Then  $\mathbf{v} + \mathbf{w} = (a + x)\hat{\mathbf{x}} + (b + y)\hat{\mathbf{y}} + (c + z)\hat{\mathbf{z}}$

We can't do this in spherical coordinates:

Let  $\mathbf{w} = (r_w, \theta_w, \phi_w)$  spherical components. Then  $\mathbf{v} + \mathbf{w} \neq (r_v + r_w, \theta_v + \theta_w, \phi_v + \phi_w)$

However, at a point off the origin, the basis vectors  $\hat{\mathbf{r}}, \hat{\theta}, \hat{\phi}$  are well defined, and can be used as a basis for general vectors. [In differential geometry, vectors referenced to a point in space are called **tangent vectors**, because they are "tangent" to the space, in a higher dimensional sense. See Differential Geometry elsewhere in this document.]

## Laplacian's Place

What is the physical meaning of the Laplacian operator? And how can I remember the Laplacian operator in any coordinates? These questions are related because understanding the physical meaning allows you to quickly derive in your head the Laplacian operator in any of the common coordinates.

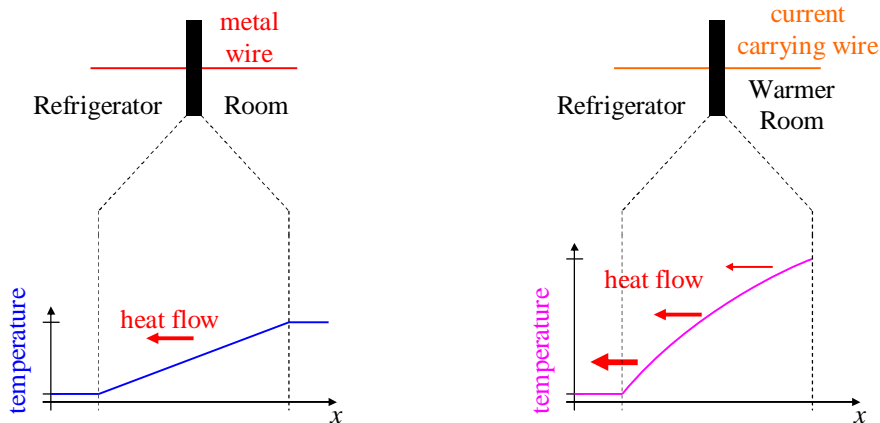
Let's take a step-by-step look at the action of the Laplacian, first in 1D, then on a 3D differential volume element, with physical examples at each step. After rectangular, we go to spherical coordinates, because they illustrate all the principles involved. Finally, we apply the concepts to cylindrical coordinates, as well. We follow this outline:

1. Overview of the Laplacian operator
2. 1D examples of heat flow
3. 3D heat flow in rectangular coordinates
4. Examples of physical scalar fields [temperature, pressure, electric potential (2 ways)]
5. 3D differential volume elements in other coordinates
6. Description of the physical meaning of Laplacian operator terms, such as

$$\nabla T, \quad \frac{\partial T}{\partial r}, \quad r^2 \frac{\partial T}{\partial r}, \quad \frac{\partial}{\partial r} \left( r^2 \frac{\partial T}{\partial r} \right), \quad r^2 \frac{\partial}{\partial r} \left( r^2 \frac{\partial T}{\partial r} \right).$$

**Overview of Laplacian operator:** Let the Laplacian act on a scalar field  $T(\mathbf{r})$ , a physical function of space, e.g. temperature. Usually, the Laplacian represents the net outflow per unit volume of some physical quantity: something/volume, e.g., something/m<sup>3</sup>. The Laplacian operator itself involves spatial second-derivatives, and so carries units of inverse area, say m<sup>-2</sup>.

**1D Example: Heat Flow:** Consider a temperature gradient along a line. It could be a perpendicular wire through the wall of a refrigerator (below left). It is a 1D system, i.e. only the gradient *along* the wire matters.



Let the left and right sides of the wire be in thermal equilibrium with the refrigerator and room, at 2 C and 27 C, respectively. The wire is passive, and can neither generate nor dissipate heat; it can only conduct it. Let the 1D thermal conductivity be  $k = 100 \text{ mW-cm/C}$ . Consider the part of the wire inside the insulated wall, 4 cm thick. How much heat (power, J/s or W) flows through the wire?

$$P = k \frac{dT}{dx} = (100 \text{ mW-cm/C}) \frac{25 \text{ C}}{4 \text{ cm}} = 625 \text{ mW} .$$

There is no heat generated or dissipated in the wire, so the heat that flows into the right side of any segment of the wire (differential or finite) must later flow out the left side. Thus, the heat flow must be

constant along the wire. Since heat flow is proportional to  $dT/dx$ ,  $dT/dx$  must be constant, and the temperature profile is linear. In other words, (1) since no heat is created or lost in the wire, heat-in = heat-out; (2) but heat flow  $\sim dT/dx$ ; so (3) the change in the temperature gradient is zero:

$$\frac{d}{dx} \left( \frac{dT}{dx} \right) = 0 = \frac{d^2T}{dx^2}.$$

(At the edges of the wall, the 1D approximation breaks down, and the inevitable nonlinearity of the temperature profile in the  $x$  direction is offset by heat flow out the sides of the wire.)

Now consider a current carrying wire which generates heat all along its length from its resistance (diagram above, right). The heat that flows into the wire from the room is added to the heat generated in the wire, and the sum of the two flows into the refrigerator. The heat generated in a length  $dx$  of wire is

$$P_{gen} = I^2 \rho dx \quad \text{where} \quad \rho \equiv \text{resistance per unit length, and } I^2 \rho = \text{const}.$$

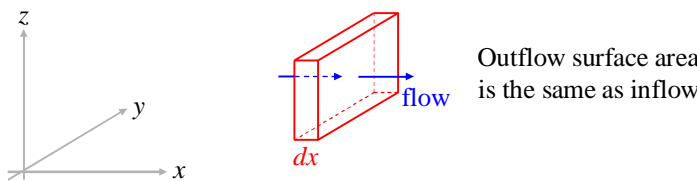
In steady state, the net outflow of heat from a segment of wire must equal the heat generated in that segment. In an infinitesimal segment of length  $dx$ , we have heat-out = heat-in + heat-generated:

$$\begin{aligned} P_{out} = P_{in} + P_{gen} &\Rightarrow \frac{dT}{dx} \Big|_a = \frac{dT}{dx} \Big|_{a+dx} + I^2 \rho dx \\ \frac{dT}{dx} \Big|_{a+dx} - \frac{dT}{dx} \Big|_a &= -I^2 \rho dx \\ \frac{d}{dx} \left( \frac{dT}{dx} \right) dx &= -I^2 \rho dx \Rightarrow \frac{d^2T}{dx^2} = -I^2 \rho \end{aligned}$$

The negative sign means that when the temperature gradient is positive (increasing to the right), the heat flow is negative (to the left), i.e. the heat flow is opposite the gradient. Many physical systems have a similar negative sign. Thus the 2<sup>nd</sup> derivative of the temperature is the negative of heat outflow (net inflow) from a segment, per unit length of the segment. Longer segments have more net outflow (generate more heat).

### 3D Rectangular Volume Element

Now consider a 3D bulk resistive material, carrying some current. The current generates heat in each volume element of material. Consider the heat flow in the  $x$  direction, with this volume element:



The temperature gradient normal to the  $y$ - $z$  face drives a heat flow per unit area, in  $W/m^2$ . For a net flow to the right, the temperature gradient must be increasing in magnitude (becoming more negative) as we move to the right. The change in gradient is proportional to  $dx$ , and the heat outflow flow is proportional to the area, and the change in gradient:

$$P_{out} - P_{in} = -k \frac{d}{dx} \left( \frac{dT}{dx} \right) dx dy dz \Rightarrow \frac{P_{out} - P_{in}}{dx dy dz} = -k \frac{d^2T}{dx^2}.$$

Thus the net heat outflow per unit volume, due to the  $x$  contribution, goes like the 2<sup>nd</sup> derivative of  $T$ . Clearly, a similar argument applies to the  $y$  and  $z$  directions, each also contributing net heat outflow per unit volume. Therefore, the *total* heat outflow per unit volume from all 3 directions is simply the sum of the heat flows in each direction:

$$\frac{P_{out} - P_{in}}{dx dy dz} = -k \left( \frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2} \right).$$

We see that in all cases, the

$$\text{net outflow of flux per unit volume} = \text{change in (flux per unit area), per unit distance}$$

We will use this fact to derive the Laplacian operator in spherical and cylindrical coordinates.

### General Laplacian

We now generalize. For the Laplacian to mean anything, it must act on a scalar field whose gradient drives a flow of some physical thing.

**Example 1:** My favorite is  $T(\mathbf{r}) =$  temperature. Then  $\nabla T(\mathbf{r})$  drives heat (energy) flow, heat per unit time, per unit area:

$$\frac{\text{heat} / t}{\text{area}} \equiv \mathbf{q} = -k \nabla T(\mathbf{r}) \quad \text{where } k \equiv \text{thermal conductivity}$$

$$\mathbf{q} \equiv \text{heat flow vector}$$

Then  $\frac{\partial T}{\partial r} \sim q_r =$  radial component of heat flow

**Example 2:**  $T(\mathbf{r}) =$  pressure of an incompressible viscous fluid (e.g. honey). Then  $\nabla T(\mathbf{r})$  drives fluid mass (or volume) flow, mass per unit time, per unit area:

$$\frac{\text{mass} / t}{\text{area}} \equiv \mathbf{j} = -k \nabla T(\mathbf{r}) \quad \text{where } k \equiv \text{some viscous friction coefficient}$$

$$\mathbf{j} \equiv \text{mass flow density vector}$$

Then  $\frac{\partial T}{\partial r} \sim j_r =$  radial component of mass flow

**Example 3:**  $T(\mathbf{r}) =$  electric potential in a resistive material. Then  $\nabla T(\mathbf{r})$  drives charge flow, charge per unit time, per unit area:

$$\frac{\text{charge} / t}{\text{area}} \equiv \mathbf{j} = -\sigma \nabla T(\mathbf{r}) \quad \text{where } \sigma \equiv \text{electrical conductivity}$$

$$\mathbf{j} \equiv \text{current density vector}$$

Then  $\frac{\partial T}{\partial r} \sim j_r =$  radial component of current density.

**Example 4:** Here we abstract a little more, to add meaning to the common equations of electromagnetics. Let  $T(\mathbf{r}) =$  electric potential in a vacuum. Then  $\nabla T(\mathbf{r})$  measures the energy per unit distance, per unit area, required to push a fixed charge density  $\rho$  through a surface, by a distance of  $dn$ , normal to the surface:

$$\frac{\text{energy/distance}}{\text{area}} \equiv \rho \nabla T(\mathbf{r}) \quad \text{where } \rho \equiv \text{electric charge volume density}.$$

Then  $\partial T / \partial r \sim$  net energy per unit radius, per unit area, to push charges of density  $\rho$  out the same distance through both surfaces.

In the first 3 examples, we use the word “flow” to mean the flow in time of some physical quantity, per unit area. In the last example, the “flow” is energy expenditure per unit distance, per unit area. The requirement of “per unit area” is essential, as we soon show.

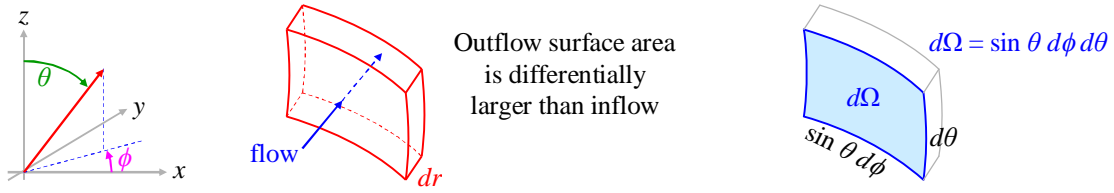
### Laplacian In Spherical Coordinates

To understand the Laplacian operator terms in other coordinates, we need to take into account two effects:

1. The outflow surface area may be different than the inflow surface area
2. The derivatives with respect to angles ( $\theta$  or  $\phi$ ) need to be converted to rate-of-change *per unit distance*.

We'll see how these two effects come into play as we develop the spherical terms of the Laplacian operator. The cylindrical terms are simplifications of the spherical terms.

**Spherical radial contribution:** We first consider the radial contribution to the spherical Laplacian operator, from this volume element:



The differential volume element has thickness  $dr$ , which can be made arbitrarily small compared to the lengths of the sides. The inner surface of the element has area  $r^2 d\Omega$ . The outer surface has infinitesimally more area. Thus the radial contribution includes both the “surface-area” effect, but not the “converting-derivatives” effect.

The increased area of the outflow surface means that for the same flux-density (flow) on inner and outer surfaces, there would be a net outflow of flux, since flux = (flux-density)(area). Therefore, we must take the derivative of the flux itself, not the flux density, and then convert the result back to per-unit-volume. We do this in 3 steps:

$$flux = (area)(flux-density) \sim (r^2 d\Omega) \left( \frac{\partial}{\partial r} \right)$$

$$\frac{d(flux)}{dr} = \frac{\partial}{\partial r} (r^2 d\Omega) \left( \frac{\partial}{\partial r} \right)$$

$$\frac{outflow}{volume} = \frac{d(flux)}{(area)dr} = \frac{1}{r^2 d\Omega} \frac{\partial}{\partial r} (r^2 d\Omega) \left( \frac{\partial}{\partial r} \right) = \frac{1}{r^2} \frac{\partial}{\partial r} (r^2) \left( \frac{\partial}{\partial r} \right)$$

The constant  $d\Omega$  factor from the area cancels when converting to flux, and back to flux-density. In other words, we can think of the fluxes as per-steradian.

We summarize the stages of the spherical radial Laplacian operator as follows:

$$\nabla^2_r T(\mathbf{r}) = \frac{1}{r^2} \frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} T(\mathbf{r})$$

$\frac{\partial}{\partial r} T$  = radial flux per unit area

$r^2 \frac{\partial}{\partial r} T$  = radial flux, per unit solid-angle =  $\frac{(area)(flow\ per\ unit\ area)}{d\Omega}$

$\frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} T$  = change in radial flux per unit length, per unit solid-angle;      positive is increasing flux

$\frac{1}{r^2} \frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} T$  = change in radial flux per unit length, per unit area

= net outflow of flux per unit volume

$$\frac{1}{r^2} \frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} T$$

$\underbrace{\quad \quad \quad}_{\text{radial flow per unit area}}$   
 $\underbrace{\quad \quad \quad}_{\text{radial flux per steradian}}$   
 $\underbrace{\quad \quad \quad}_{\text{change in radial flux per unit length per steradian}}$   
 $\underbrace{\quad \quad \quad}_{\text{change in radial flux per unit length, per unit area}}$

Following the steps in the example of heat flow, let  $T(\mathbf{r})$  = temperature. Then

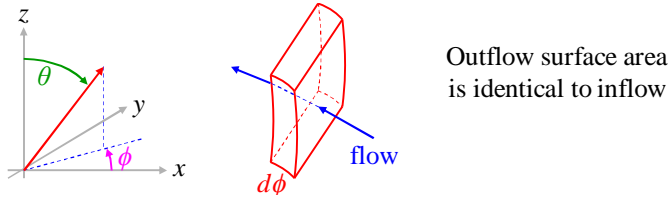
$$\frac{\partial}{\partial r} T = \text{radial heat flow per unit area, W/m}^2$$

$$r^2 \frac{\partial}{\partial r} T = \text{radial heat flux, W/solid-angle} = \frac{\text{Watts}}{\text{steradian}}$$

$$\frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} T = \text{change in radial heat flux per unit length, per unit solid-angle}$$

$$\frac{1}{r^2} \frac{\partial}{\partial r} r^2 \frac{\partial}{\partial r} T = \text{net outflow of heat flux per unit volume}$$

**Spherical azimuthal contribution:** The spherical  $\phi$  contribution to the Laplacian has no area-change, but does require converting derivatives. Consider the volume element:



The inflow and outflow surface areas are the same, and therefore area-change contributes nothing to the derivatives.

However, we must convert the derivatives with respect to  $\phi$  into rates-of-change with respect to distance, because physically, the flow is driven by a derivative with respect to distance. In the spherical  $\phi$  case, the effective radius for the arc-length along the flow is  $r \sin \theta$ , because we must project the position vector into the plane of rotation. Thus,  $(\partial/\partial \phi)$  is the rate-of-change per  $(r \sin \theta)$  meters. Therefore,

$$\text{rate-of-change-per-meter} = \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi}$$

Performing the two derivative conversions, we get

$$\nabla^2_{\phi} T(\mathbf{r}) = \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} T(\mathbf{r})$$

$$\frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} T = \text{azimuthal flux per unit area}$$

$$\frac{\partial}{\partial \phi} \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} T = \text{change in (azimuthal flux per unit area) per radian}$$

$$\frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} T = \text{change in (azimuthal flux per unit area) per unit distance}$$

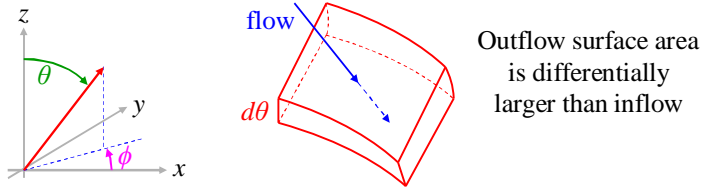
= net azimuthal outflow of flux per unit volume

$$\frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} \underbrace{\frac{1}{r \sin \theta} \frac{\partial}{\partial \phi} T}_{\substack{\text{azimuthal flux} \\ \text{per unit area}}} = \frac{1}{r^2 \sin^2 \theta} \frac{\partial^2}{\partial \phi^2} T$$

$\underbrace{\hspace{10em}}_{\substack{\text{change in (azimuthal flux} \\ \text{per unit area) per radian}}}$   
 $\underbrace{\hspace{10em}}_{\substack{\text{change in (azimuthal flux per} \\ \text{unit area) per unit distance}}}$

Notice that the  $r^2 \sin^2 \theta$  in the denominator is not a physical area; it comes from two derivative conversions.

**Spherical polar angle contribution:**



The volume element is like a wedge of an orange: it gets wider (in the northern hemisphere) as  $\theta$  increases. Therefore the outflow area is differentially larger than the inflow area (in the northern hemisphere). In particular,  $area = (r \sin \theta) dr$ , but we only need to keep the  $\theta$  dependence, because the factors of  $r$  cancel, just like  $d\Omega$  did in the spherical radial contribution. So we have

$$area \propto \sin \theta .$$

In addition, we must convert the  $\partial/\partial \theta$  to a rate-of-change with distance. Thus the spherical polar angle contribution has *both* area-change and derivative-conversion.

Following the steps of converting to flux, taking the derivative, then converting back to flux-density, we get

$$\nabla^2_{\theta} T(\mathbf{r}) = \frac{1}{\sin\theta} \frac{1}{r} \frac{\partial}{\partial\theta} \sin\theta \frac{1}{r} \frac{\partial}{\partial\theta} T(\mathbf{r})$$

$$\frac{1}{r} \frac{\partial}{\partial\theta} T = \hat{\theta}\text{-flux per unit area}$$

$$\sin\theta \frac{1}{r} \frac{\partial}{\partial\theta} T = \hat{\theta}\text{-flux, per unit radius} = \frac{(\text{area})(\text{flux per unit area})}{dr}$$

$$\frac{\partial}{\partial\theta} \sin\theta \frac{1}{r} \frac{\partial}{\partial\theta} T = \text{change in } (\hat{\theta}\text{-flux per unit radius}), \text{ per radian}$$

$$\frac{1}{r} \frac{\partial}{\partial\theta} \sin\theta \frac{1}{r} \frac{\partial}{\partial\theta} T = \text{change in } (\hat{\theta}\text{-flux per unit radius}), \text{ per unit distance}$$

$$\frac{1}{\sin\theta} \frac{1}{r} \frac{\partial}{\partial\theta} \sin\theta \frac{1}{r} \frac{\partial}{\partial\theta} T = \text{change in } (\hat{\theta}\text{-flux per unit area}), \text{ per unit distance}$$

= net outflow of flux per unit volume

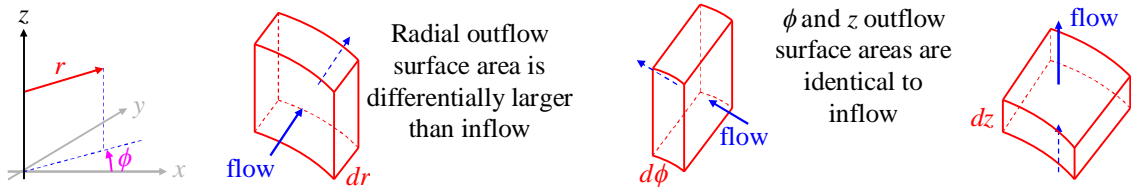
$$\frac{1}{\sin\theta} \frac{1}{r} \frac{\partial}{\partial\theta} \underbrace{\sin\theta \frac{1}{r} \frac{\partial}{\partial\theta} T}_{\substack{\hat{\theta}\text{-flux per} \\ \text{unit area}}} = \frac{1}{r^2} \frac{\partial}{\partial\theta} \sin\theta \frac{\partial}{\partial\theta} T$$

$\underbrace{\hspace{10em}}_{\substack{\hat{\theta}\text{-flux, per} \\ \text{unit radius}}}$   
 $\underbrace{\hspace{10em}}_{\substack{\text{change in } (\hat{\theta}\text{-flux per} \\ \text{unit radius}), \text{ per radian}}}$   
 $\underbrace{\hspace{10em}}_{\substack{\text{change in } (\hat{\theta}\text{-flux per unit} \\ \text{radius}), \text{ per unit distance}}}$   
 $\underbrace{\hspace{10em}}_{\substack{\text{change in } (\hat{\theta}\text{-flux per unit} \\ \text{area}), \text{ per unit distance}}}$

Notice that the  $r^2$  in the denominator is not a physical area; it comes from two derivative conversions.

### Cylindrical Coordinates

The cylindrical terms are simplifications of the spherical terms.



**Cylindrical radial contribution:** The picture of the cylindrical radial contribution is essentially the same as the spherical, but the “height” of the slab is exactly constant. We still face the issues of varying inflow and outflow surface areas, and converting derivatives to rate of change per unit distance. The change in area is due only to the arc length  $r d\phi$ , with the  $z$  (height) fixed. Thus we write the radial result directly:



$$\nabla_r^2 T(\mathbf{r}) = \frac{1}{r} \frac{\partial}{\partial r} r \frac{\partial}{\partial r} T(\mathbf{r}) \quad (\text{Cylindrical Coordinates})$$

$$\frac{\partial}{\partial r} T = \text{radial flow per unit area}$$

$$r \frac{\partial}{\partial r} T = \text{radial flux per unit angle} = \frac{(\text{flow per unit area})(\text{area})}{d\phi dz}$$

$$\frac{\partial}{\partial r} r \frac{\partial}{\partial r} T = \text{change in (radial flux per unit angle), per unit radius}$$

$$\frac{1}{r} \frac{\partial}{\partial r} r \frac{\partial}{\partial r} T = \text{change in (radial flux per unit area), per unit radius}$$

= net outflow of flux per unit volume

$$\frac{1}{r} \frac{\partial}{\partial r} r \frac{\partial}{\partial r} T$$

$\underbrace{\hspace{1.5cm}}$   
 radial flow  
per unit area  
 $\underbrace{\hspace{1.5cm}}$   
 radial flux  
per radian  
 $\underbrace{\hspace{1.5cm}}$   
 change in radial flux per  
unit length per radian  
 $\underbrace{\hspace{1.5cm}}$   
 change in (radial flux per  
unit area), per unit radius

**Cylindrical azimuthal contribution:** Like the spherical case, the inflow and outflow surfaces have identical areas. Therefore, the  $\phi$  contribution is similar to the spherical case, except there is no  $\sin \theta$  factor;  $r$  contributes directly to the arc-length and rate-of-change per unit distance:

$$\nabla_\phi^2 T(\mathbf{r}) = \frac{1}{r} \frac{\partial}{\partial \phi} \frac{1}{r} \frac{\partial}{\partial \phi} T(\mathbf{r})$$

$$\frac{1}{r} \frac{\partial}{\partial \phi} T = \text{azimuthal flux per unit area}$$

$$\frac{\partial}{\partial \phi} \frac{1}{r} \frac{\partial}{\partial \phi} T = \text{change in (azimuthal flux per unit area) per radian}$$

$$\frac{1}{r} \frac{\partial}{\partial \phi} \frac{1}{r} \frac{\partial}{\partial \phi} T = \text{change in (azimuthal flux per unit area) per unit distance}$$

= net azimuthal outflow of flux per unit volume

$$\frac{1}{r} \frac{\partial}{\partial \phi} \frac{1}{r} \frac{\partial}{\partial \phi} T = \frac{1}{r^2} \frac{\partial^2}{\partial \phi^2} T$$

$\underbrace{\hspace{1.5cm}}$   
 azimuthal flow  
per unit area  
 $\underbrace{\hspace{1.5cm}}$   
 change in azimuthal  
flow per radian  
 $\underbrace{\hspace{1.5cm}}$   
 change in (azimuthal flux per  
unit area) per unit distance

**Cylindrical z contribution:** This is identical to the rectangular case: the inflow and outflow areas are the same, and the derivative is already per unit distance, ergo: (add cylindrical volume element picture??)

$$\nabla^2_z T(\mathbf{r}) = \frac{\partial}{\partial z} \frac{\partial}{\partial z} T(\mathbf{r})$$

$$\frac{\partial}{\partial z} T = \text{vertical flux per unit area}$$

$$\frac{\partial}{\partial z} \frac{\partial}{\partial z} T = \text{change in (vertical flux per unit area) per unit distance}$$

$$= \text{net outflow of flux per unit volume}$$

$$\underbrace{\frac{\partial}{\partial z} \underbrace{\frac{\partial}{\partial z} T}_{\substack{\text{vertical flux} \\ \text{per unit area}}}}_{\substack{\text{change in (vertical flux per} \\ \text{unit area) per unit distance}}} = \frac{\partial^2}{\partial z^2} T$$

### Vector Dot Grad Vector

In electromagnetic propagation, and elsewhere, one encounters the “dot product” of a vector field with the gradient operator, acting on a vector field. What is this  $\mathbf{v} \cdot \nabla$  operator? Here,  $\mathbf{v}(\mathbf{r})$  is a given vector field. The simple view is that  $\mathbf{v}(\mathbf{r}) \cdot \nabla$  is just a notational shorthand for

$$\mathbf{v}(\mathbf{r}) \cdot \nabla \equiv \left( v_x \frac{\partial}{\partial x} + v_y \frac{\partial}{\partial y} + v_z \frac{\partial}{\partial z} \right),$$

$$\text{because } \mathbf{v}(\mathbf{r}) \cdot \nabla = (v_x \hat{\mathbf{x}} + v_y \hat{\mathbf{y}} + v_z \hat{\mathbf{z}}) \cdot \left( \frac{\partial}{\partial x} \hat{\mathbf{x}} + \frac{\partial}{\partial y} \hat{\mathbf{y}} + \frac{\partial}{\partial z} \hat{\mathbf{z}} \right) = \left( v_x \frac{\partial}{\partial x} + v_y \frac{\partial}{\partial y} + v_z \frac{\partial}{\partial z} \right)$$

by the usual rules for a dot product in rectangular coordinates.

There is a deeper meaning, though, which is an important bridge to the topics of tensors and differential geometry.

We can view the  $\mathbf{v} \cdot \nabla$  operator as simply the dot product of the vector field  $\mathbf{v}(r)$  with the gradient of a vector field.

You may think of the gradient operator as acting on a *scalar* field, to produce a vector field. But the gradient operator can also act on a vector field, to produce a tensor field. Here’s how it works: You are probably familiar with derivatives of a vector field:

Let  $\mathbf{A}(x, y, z)$  be a vector field. Then  $\frac{\partial \mathbf{A}}{\partial x} = \left( \frac{\partial A_x}{\partial x} \hat{\mathbf{x}} + \frac{\partial A_y}{\partial x} \hat{\mathbf{y}} + \frac{\partial A_z}{\partial x} \hat{\mathbf{z}} \right)$  is a vector field.

Writing spatial vectors as column vectors,  $A = \begin{pmatrix} A_x \\ A_y \\ A_z \end{pmatrix}$ , and  $\frac{\partial \mathbf{A}}{\partial x} = \begin{pmatrix} \frac{\partial A_x}{\partial x} \\ \frac{\partial A_y}{\partial x} \\ \frac{\partial A_z}{\partial x} \end{pmatrix}$

Similarly,  $\frac{\partial \mathbf{A}}{\partial y}$  and  $\frac{\partial \mathbf{A}}{\partial z}$  are also vector fields.

By the rule for total derivatives, for a small displacement  $(dx, dy, dz)$ ,

$$d\mathbf{A} \equiv \begin{pmatrix} dA_x \\ dA_y \\ dA_z \end{pmatrix} = \frac{\partial \mathbf{A}}{\partial x} dx + \frac{\partial \mathbf{A}}{\partial y} dy + \frac{\partial \mathbf{A}}{\partial z} dz = \begin{pmatrix} \frac{\partial A_x}{\partial x} & \frac{\partial A_x}{\partial y} & \frac{\partial A_x}{\partial z} \\ \frac{\partial A_y}{\partial x} & \frac{\partial A_y}{\partial y} & \frac{\partial A_y}{\partial z} \\ \frac{\partial A_z}{\partial x} & \frac{\partial A_z}{\partial y} & \frac{\partial A_z}{\partial z} \end{pmatrix} \begin{pmatrix} dx \\ dy \\ dz \end{pmatrix} = \begin{pmatrix} \frac{\partial A_x}{\partial x} \\ \frac{\partial A_y}{\partial x} \\ \frac{\partial A_z}{\partial x} \end{pmatrix} dx + \begin{pmatrix} \frac{\partial A_x}{\partial y} \\ \frac{\partial A_y}{\partial y} \\ \frac{\partial A_z}{\partial y} \end{pmatrix} dy + \begin{pmatrix} \frac{\partial A_x}{\partial z} \\ \frac{\partial A_y}{\partial z} \\ \frac{\partial A_z}{\partial z} \end{pmatrix} dz .$$

This says that the vector  $d\mathbf{A}$  is a linear combination of 3 column vectors  $\partial\mathbf{A}/\partial x$ ,  $\partial\mathbf{A}/\partial y$ , and  $\partial\mathbf{A}/\partial z$ , weighted respectively by the displacements  $dx$ ,  $dy$ , and  $dz$ . The 3 x 3 matrix above is the gradient of the vector field  $\mathbf{A}(\mathbf{r})$ . It is the natural extension of the gradient (of a scalar field) to a vector field. It is a rank-2 tensor, which means that given a vector  $(dx, dy, dz)$ , it produces a vector ( $d\mathbf{A}$ ) which is a linear combination of 3 (column) vectors  $(\nabla\mathbf{A})$ , each weighted by the components of the given vector  $(dx, dy, dz)$ .

Note that  $\nabla\mathbf{A}$  and  $\nabla\cdot\mathbf{A}$  are very different: the former is a rank-2 tensor field, the latter is a scalar field.

This concept extends further to derivatives of rank-2 tensors, which are rank-3 tensors: 3 x 3 x 3 cubes of numbers, producing a linear combination of 3 x 3 arrays, weighted by the components of a given vector  $(dx, dy, dz)$ . And so on.

Note that in other coordinates (e.g., cylindrical or spherical),  $\nabla\mathbf{A}$  is *not* given by the derivative of its components with respect to the 3 coordinates. The components interact, because the basis vectors also change through space. That leads to the subject of differential geometry, discussed elsewhere in this document.

## 4 Green's Functions

Green's functions are a method of solving inhomogeneous linear differential equations (or other linear operator equations):

$$\mathcal{L}\{f(x)\} = s(x), \quad \text{where } \mathcal{L}\{ \} \text{ is a linear operator .}$$

We use them when other methods are hard, or to make a useful approximation (the Born approximation). Sometimes, the Green's function itself can be given physical meaning, as in Quantum Field Theory. Green's functions can generate particular (i.e. inhomogeneous) solutions, and solutions matching boundary conditions. They don't generate homogeneous solutions (i.e., where the right hand side is zero). We explore Green's functions through the following steps:

1. Extremely brief review of the  $\delta$ -function.
2. The tired, but inevitable, electromagnetic example.
3. Linear differential equations of one variable (1-dimensional), with sources.
4. Delta function expansions.
5. Green's functions of two variables (but 1 dimension).
6. When you can collapse a Green's function to one variable ("portable Green's functions": translational invariance)
7. Dealing with boundary conditions: at least 5 (6??) kinds of BC
8. Green-like methods: the Born approximation

You will find *no* references to "Green's Theorem" or "self-adjoint" until we get to non-homogeneous boundary conditions, because those topics are unnecessary and confusing before then. We will see that:

The biggest hurdle in understanding Green's functions is the boundary conditions.

### Dirac Delta Function

Recall that the Dirac  $\delta$ -function is an "impulse," an infinitely narrow, tall spike function, *defined* as

$$\delta(x) = 0, \quad \text{for } x \neq 0, \quad \text{and} \quad \int_{-a}^a \delta(x) dx = 1, \quad \forall a > 0 \quad (\text{the area under the d-function is } 1) .$$

The linearity of integration implies the delta function can be offset, and weighted, so that

$$\int_{b-a}^{b+a} w\delta(x-b) dx = w \quad \forall a > 0 .$$

Since the  $\delta$ -function is infinitely narrow, it can "pick out" a single value from a function:

$$\int_{b-a}^{b+a} \delta(x-b)f(x) dx = f(b) \quad \forall a > 0 .$$

[It also implies  $\delta(0) \rightarrow \infty$ , but we don't focus on that here.]

(See *Funky Quantum Concepts* for more on the delta function).

### The Tired, But Inevitable, Electromagnetic Example

You probably have seen Poisson's equation relating the electrostatic potential at a point to a charge distribution creating the potential (in gaussian units):

$$(1) \quad -\nabla^2\phi(\mathbf{r}) = 4\pi\rho(\mathbf{r}) \quad \text{where } \phi \equiv \text{electrostatic potential, } \rho \equiv \text{charge density .}$$

We solved this by noting three things: (1a) electrostatic potential,  $\phi$ , obeys “superposition:” the potential due to multiple charges is the sum of the potentials of the individual charges; (1b) the potential is proportional to the source charge; and (2) the potential due to a point charge is:

$$\phi(\mathbf{r}) = q \frac{1}{r} \quad (\text{point charge at origin}).$$

The properties (1a) and (1b) above, taken together, define a **linear** relationship:

$$\begin{aligned} \text{Given } \rho_1(\mathbf{r}) \rightarrow \phi_1(\mathbf{r}), \quad \text{and} \quad \rho_2(\mathbf{r}) \rightarrow \phi_2(\mathbf{r}) \\ \text{Then } a\rho_1(\mathbf{r}) + \rho_2(\mathbf{r}) \rightarrow \phi_{total}(\mathbf{r}) = a\phi_1(\mathbf{r}) + \phi_2(\mathbf{r}) \end{aligned}$$

To solve Eq (1), we break up the source charge distribution into an infinite number of little point charges spread out over space, each of charge  $\rho d^3r$ . The solution for  $\phi$  is the sum of potential from all the point charges, and the infinite sum is an integral, so we find  $\phi$  as

$$\phi(\mathbf{r}) = \int \rho(\mathbf{r}') d^3r' \frac{1}{|\mathbf{r} - \mathbf{r}'|}.$$

Note that the charge “distribution” for a point charge is a  $\delta$ -function: infinite charge density, but finite total charge. [We have also implicitly used the fact that the potential is translationally invariant, and depends only on the distance from the source. We will remove this restriction later.]

But all of this followed from simple mathematical properties of Eq (1) that have nothing to do with electromagnetics. All we used to solve for  $\phi$  was that the left-hand side is a linear operator on  $\phi$  (so superposition applies), and we have a known solution when the right-hand side is a delta function:

$$\underbrace{-\nabla^2}_{\text{linear operator}} \underbrace{\phi(\mathbf{r})}_{\text{unknown function}} = \underbrace{4\pi\rho(\mathbf{r})}_{\text{given "source" function}} \quad \text{and} \quad \underbrace{-\nabla^2}_{\text{linear operator}} \underbrace{\frac{1}{|\mathbf{r} - \mathbf{r}'|}}_{\text{known solution}} = \underbrace{\delta(\mathbf{r} - \mathbf{r}')}_{\text{given point "source" at } \mathbf{r}'}$$

The solution for a given  $\rho$  is a sum of delta-function solutions. Now we generalize all this to arbitrary (for now, 1D) linear operator equations by letting  $\mathbf{r} \rightarrow x$ ,  $\phi \rightarrow f$ ,  $-\nabla^2 \rightarrow \mathcal{L}$ ,  $\rho \rightarrow s$ , and call the known  $\delta$ -function solution  $G(x)$ :

$$\text{Given } \mathcal{L}\{f(x)\} = s(x) \quad \text{and} \quad \mathcal{L}\{G(x)\} = \delta(x), \quad \text{then} \quad f(x) = \int s(x') dx' G(x - x').$$

assuming, as above, that our linear operator, and any boundary conditions, are translationally invariant.

**A Fresh, New Signal Processing Example**

If this example doesn’t make sense to you, just skip it. Signal processing folk have long used a Green’s function concept, but with different words. A time-invariant linear system (TILS) produces an output which is a linear operation on its input:

$$o(t) = \mathcal{M}\{i(t)\} \quad \text{where } \mathcal{M}\{ \} \text{ is a linear operation taking input to output}$$

In this case, we aren’t given  $\mathcal{M}\{ \}$ , and we don’t solve for it. However, we *are* given a measurement (or computation) of the system’s **impulse response**, called  $h(t)$  (not to be confused with a homogeneous solution to anything). If you poke the system with a very short spike (i.e., if you feed an impulse into the system), it responds with  $h(t)$ .

$$h(t) = \mathcal{M}\{\delta(t)\} \quad \text{where } h(t) \text{ is the system's impulse response.}$$

Note that the impulse response is spread out over time, and usually of (theoretically) infinite duration.  $h(t)$  fully characterizes the system, because we can approximate any input function as a series of impulses, and sum up all the responses. Therefore, we find the output for any input,  $i(t)$ , with:

$$o(t) = \int_{-\infty}^{\infty} i(t')h(t-t') dt'$$

$h(t)$  acts like a Green's function, giving the system response at time  $t$  to a delta function at  $t = 0$ .

**Linear differential equations of one variable, with sources**

We wish to solve for  $f(x)$ , given  $s(x)$ :

$$\mathcal{L}\{f(x)\} = s(x), \quad \text{where } \mathcal{L}\{ \} \text{ is a linear operator}$$

$s(x)$  is called the "source," or forcing function

E.g., 
$$\left( \frac{d^2}{dx^2} + \omega^2 \right) f(x) \equiv \frac{d^2}{dx^2} f(x) + \omega^2 f(x) = s(x)$$

We ignore boundary conditions for now (to be dealt with later). The differential equations often have 3D space as their domain. Note that we are *not* differentiating  $s(x)$ , which will be important when we get to the delta-function expansion of  $s(x)$ .

Green's functions solve the above equation by first solving a related equation: if we can find a function (i.e., a "Green's function") such that

$$\mathcal{L}\{G(x)\} = \delta(x), \quad \text{where } \delta(x) \text{ is the Dirac delta function}$$

E.g., 
$$\left( \frac{d^2}{dx^2} + \omega^2 \right) G(x) = \delta(x)$$

then we can use that Green's function to solve our original equation.

This might seem weird, because  $\delta(0) \rightarrow \infty$ , but it just means that Green's functions often have discontinuities in them or their derivatives. For example, suppose  $G(x)$  is a step function:

$$G(x) = \begin{cases} = 0, & x < 0 \\ = 1, & x > 0 \end{cases} \quad \text{Then } \frac{d}{dx} G(x) = \delta(x).$$

Now suppose our source isn't centered at the origin, i.e.,  $s(x) = \delta(x - a)$ . If  $\mathcal{L}\{ \}$  is translation invariant [along with any boundary conditions], then  $G( )$  can still solve the equation by translation:

$$\mathcal{L}\{f(x)\} = s(x) = \delta(x - a), \quad \Rightarrow \quad f(x) = G(x - a) \text{ is a solution.}$$

If  $s(x)$  is a weighted sum of delta functions at different places, then because  $\mathcal{L}\{ \}$  is linear, the solution is immediate; we just add up the solutions from all the  $\delta$ -functions:

$$\mathcal{L}\{f(x)\} = s(x) = \sum_i w_i \delta(x - x_i) \quad \Rightarrow \quad f(x) = \sum_i w_i G(x - x_i).$$

Usually the source  $s(x)$  is continuous. Then we can use  $\delta$ -functions as a basis to expand  $s(x)$  as an infinite sum of delta functions (described in a moment). The summation goes over to an integral, and a solution is

$$\mathcal{L}\{f(x)\} = s(x) = \sum_{i=1}^{\infty} w_i \delta(x - x_i) \quad \begin{matrix} x_i \rightarrow x' \\ w_i \rightarrow s(x') dx' \\ \rightarrow \end{matrix}$$

$$\mathcal{L}\{f(x)\} = s(x) = \int dx' s(x') \delta(x - x') \quad \text{and} \quad f(x) = \int dx' s(x') G(x - x')$$

We can show directly that  $f(x)$  is a solution of the original equation by plugging it in, and noting that  $\mathcal{L}\{ \}$  acts in the  $x$  domain, and "goes through" (i.e., commutes with) any operation in  $x'$ :

$$\begin{aligned} \mathcal{L}\{f(x)\} &= \mathcal{L}\left\{\int dx' s(x')G(x-x')\right\} \\ &= \int dx' s(x')\mathcal{L}\{G(x-x')\} && \text{moving } \mathcal{L}\{ \} \text{ inside the integral} \\ &= \int dx' s(x')\delta(x-x') = s(x) && \delta() \text{ picks out the value of } s(x). \text{ QED.} \end{aligned}$$

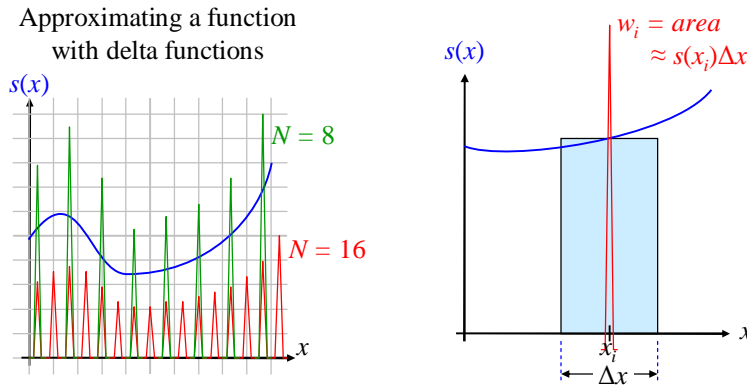
We now digress for a moment to understand the  $\delta$ -function expansion.

**Delta Function Expansion**

As in the EM example, it is frequently quite useful to expand a given function  $s(x)$  as a sum of  $\delta$ -functions:

$$s(x) \approx \sum_{i=1}^N w_i \delta(x-x_i), \quad \text{where } w_i \text{ are the weights of the basis delta functions.}$$

[This same expansion is used to characterize the “impulse-response” of linear systems.]



On the left, we approximate  $s(x)$  first with  $N = 8$   $\delta$ -functions (green), then with  $N = 16$   $\delta$ -functions (red). As we double  $N$ , the weight of each  $\delta$ -function is roughly cut in half, but there are twice as many of them. Hence, the integral of the  $\delta$ -function approximation remains about the same. Of course, the approximation gets better as  $N$  increases. As usual, we let the number of  $\delta$ -functions go to infinity:  $N \rightarrow \infty$ .

On the right above, we show how to choose the weight of each  $\delta$ -function: its weight is such that its integral approximates the integral of the given function,  $s(x)$ , over the interval “covered” by the  $\delta$ -function. In the limit of  $N \rightarrow \infty$ , the approximation becomes arbitrarily good.

In what sense is the  $\delta$ -function series an approximation to  $s(x)$ ? Certainly, if we need the derivative  $s'(x)$ , the delta-function expansion is *terrible*. However, if we want the integral of  $s(x)$ , or any integral operator, such as an inner product or a convolution, then the delta-function series is a good approximation:

$$\begin{aligned} \text{For } \int s(x) dx \quad \text{or} \quad \int f^*(x)s(x) dx \quad \text{or} \quad \int f(x'-x)s(x) dx, \\ \text{then } s(x) \approx \sum_{i=1}^N w_i \delta(x-x_i) \quad \text{where } w_i = s(x_i)\Delta x \end{aligned}$$

As  $N \rightarrow \infty$ , we expand  $s(x)$  in an infinite sum (an integral) of  $\delta$ -functions:

$$s(x) = \sum_i w_i \delta(x-x_i) \quad \begin{matrix} x_i \rightarrow x' \\ \Delta x \rightarrow dx' \\ w_i \rightarrow s(x')dx' \\ \rightarrow \end{matrix} \quad s(x) = \int dx' s(x')\delta(x-x'),$$

which if you think about it, follows directly from the definition of  $\delta(x)$ .

[Aside: Delta-functions are a continuous set of orthonormal basis functions, much like sinusoids from quantum mechanics and Fourier transforms. They satisfy all the usual orthonormal conditions for a continuous basis, i.e. they are **orthogonal** and **normalized**:

$$\int_{-\infty}^{\infty} dx \delta(x-a)\delta(x-b) = \delta(a-b) .]$$

Note that in the final solution of the prior section, we integrate  $s(x)$  times other stuff:

$$f(x) = \int dx' s(x')G(x-x') .$$

and integrating  $s(x)$  is what makes the  $\delta$ -function expansion of  $s(x)$  valid.

### Introduction to Boundary Conditions

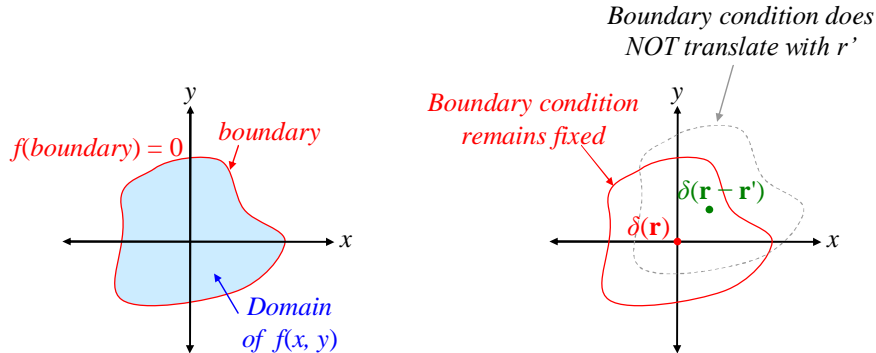
We now incorporate a simple boundary condition. Consider a 2D problem in the plane:

$$\begin{aligned} \mathcal{L}\{f(x,y)\} &= s(x,y) && \text{inside the boundary} \\ f(\text{boundary}) &= 0, && \text{where the boundary is given.} \end{aligned}$$

We define the vector  $\mathbf{r} \equiv (x, y)$ , and recall that

$$\delta(\mathbf{r}) \equiv \delta(x)\delta(y), \quad \text{so that} \quad \delta(\mathbf{r}-\mathbf{r}') = \delta(x-x')\delta(y-y') .$$

[Some references use the notation  $\delta^{(2)}(\mathbf{r})$  for a 2D  $\delta$ -function.]



(Left) The domain of interest, and its boundary. (Right) A solution meeting the BC for the source at  $(0, 0)$  does *not* translate to another point and still meet the BC.

The boundary condition removes the translation invariance of the problem. The delta-function response of  $\mathcal{L}\{G(\mathbf{r})\}$  translates, but the boundary condition does *not*. I.e., a solution of

$$\begin{aligned} \mathcal{L}\{G(\mathbf{r})\} = \delta(\mathbf{r}), \text{ and } G(\text{boundary}) = 0 &\Rightarrow \mathcal{L}\{G(\mathbf{r}-\mathbf{r}')\} = \delta(\mathbf{r}-\mathbf{r}') \\ \text{BUT does NOT} &\Rightarrow G(\text{boundary}-\mathbf{r}') = 0 . \end{aligned}$$

With boundary conditions, for each source point  $\mathbf{r}'$ , we need a different Green's function!

The Green's function for a source point  $\mathbf{r}'$ , call it  $G_{\mathbf{r}'}(\mathbf{r})$ , must satisfy *both*:

$$\mathcal{L}\{G_{\mathbf{r}'}(\mathbf{r})\} = \delta(\mathbf{r}-\mathbf{r}') \quad \text{and} \quad G_{\mathbf{r}'}(\text{boundary}) = 0 .$$

We can think of this as a Green's function of two arguments,  $r$  and  $r'$ , but really,  $r$  is the argument, and  $r'$  is a parameter. In other words, we have a *family* of Green's functions,  $G_{r'}(r)$ , labeled by the location of the delta-function,  $r'$ .



**Example:** Returning to a 1D example in  $r$ : Find the Green's function for the equation

$$\frac{d^2}{dr^2} f(r) = s(r), \text{ on the interval } [0,1], \text{ subject to } f(0) = f(1) = 0.$$

**Solution:** The Green's function equation replaces the source  $s(r)$  with  $\delta(r - r')$ :

$$\frac{d^2}{dr^2} G_{r'}(r) = \delta(r - r').$$

Note that  $G_{r'}(r)$  satisfies the *homogeneous* equation on either side of  $r'$ :

$$\frac{d^2}{dr^2} G_{r'}(r \neq r') = 0.$$

The full Green's function simply matches two homogeneous solutions, one to the left of  $r'$ , and another to the right of  $r'$ , such that the discontinuity at  $r'$  creates the required  $\delta$ -function there. First we find the homogeneous solutions:

$$\frac{d^2}{dr^2} h(r) = 0 \quad \text{Integrate both sides:}$$

$$\frac{d}{dr} h(r) = C \quad \text{where } C \text{ is an integration constant. Integrate again:}$$

$$h(r) = Cr + D \quad \text{where } C, D \text{ are arbitrary constants}$$

There are now 2 cases: (left)  $r < r'$ , and (right)  $r > r'$ . Each solution requires its own set of integration constants.

$$\text{Left case: } r < r' \Rightarrow G_{r'}(r) = Cr + D$$

$$\text{Only the left boundary condition applies to } r < r': G_{r'}(0) = 0 \Rightarrow D = 0$$

$$\text{Right case: } r > r' \Rightarrow G_{r'}(r) = Er + F$$

$$\text{Only the right boundary condition applies to } r > r': G_{r'}(1) = 0 \Rightarrow E + F = 0, F = -E$$

So far, we have:

$$\text{Left case: } G(r < r') = Cr \quad \text{Right case: } G(r > r') = Er - E.$$

The integration constants  $C$  and  $E$  are as-yet unknown. Now we must match the two solutions at  $r = r'$ , and introduce a delta function there. The  $\delta$ -function must come from the highest derivative in  $\mathcal{L}\{ \}$ , in this case the 2<sup>nd</sup> derivative, because if  $G'(r)$  had a delta function, then the 2<sup>nd</sup> derivative  $G''(r)$  would have the derivative of a  $\delta$ -function, which cannot be canceled by any other term in  $\mathcal{L}\{ \}$ . Since the derivative of a step (discontinuity) is a  $\delta$ -function,  $G'(r)$  must have a discontinuity, so that  $G''(r)$  has a  $\delta$ -function. And finally, if  $G'(r)$  has a discontinuity, then  $G(r)$  has a cusp (aka "kink" or sharp point).

We can find  $G(r)$  to satisfy all this by matching  $G(r)$  and  $G'(r)$  of the left and right Green's functions, at the point where they meet,  $r = r'$ :

$$\text{Left: } \frac{d}{dr} G_{r'}(r < r') = C \quad \text{Right: } \frac{d}{dr} G_{r'}(r > r') = E$$

There must be a unit step in the derivative across  $r = r'$ :

$$C + 1 = E$$

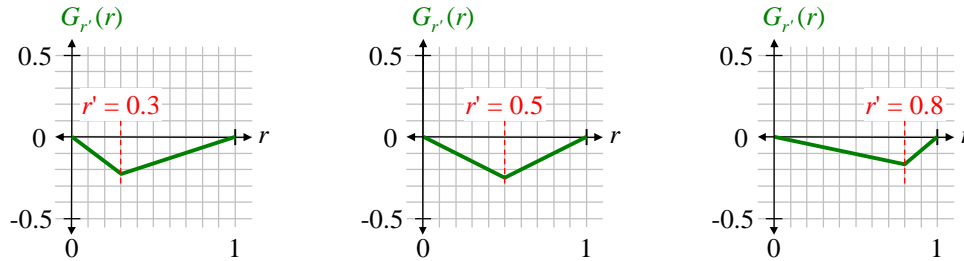
So we eliminate  $E$  in favor of  $C$ . Also,  $G(r)$  must be continuous (or else  $G'(r)$  would have a  $\delta$ -function), which means

$$G_{r'}(r = r'_{-}) = G_{r'}(r = r'_{+}) \Rightarrow Cr' = (C + 1)r' - C - 1, \quad C = r' - 1.$$

yielding the final Green's function for the given differential equation:

$$G_{r'}(r < r') = (r' - 1)r, \quad G_{r'}(r > r') = r'r - r' = r'(r - 1).$$

Here's a plot of these Green's functions for different values of  $r'$ :



To find the solution  $f(x)$ , we need to integrate over  $r'$ ; therefore, it is convenient to write the Green's function as a true function of two variables:

$$G(r; r') \equiv G_{r'}(r) \Rightarrow \mathcal{L}\{G(r; r')\} = \delta(r - r'), \quad \text{and} \quad G(\text{boundary}; r') = 0,$$

where the “;” between  $r$  and  $r'$  emphasizes that  $G(r; r')$  is a function of  $r$ , parameterized by  $r'$ . I.e., we can still think of  $G(r; r')$  as a family of functions of  $r$ , where each family member is labeled by  $r'$ , and each family member satisfies the homogeneous boundary condition.

It is important here that the boundary condition is zero, so that any sum of Green's functions still satisfies the boundary condition.

Our particular solution to the original equation, which now satisfies the homogeneous boundary condition, is

$$f(r) = \int_0^1 dr' s(r') G(r; r') = \int_0^r dr' s(r') \underbrace{r'(r-1)}_{G(r;r'), r>r'} + \int_r^1 dr' s(r') \underbrace{(r'-1)r}_{G(r;r'), r<r'}$$

which satisfies  $f(\text{boundary}) = 0$

**Summary:** To solve  $\mathcal{L}\{G_{x'}(x)\} = \delta(x - x')$ , we break  $G(x)$  into left- and right- sides of  $x'$ . Each side satisfies the homogeneous equation,  $\mathcal{L}\{G_{x'}(x)\} = 0$ , with arbitrary constants. We use the matching conditions to achieve the  $\delta$ -function at  $x'$ , which generates a set of simultaneous equations for the unknown constants in the homogeneous solutions. We solve for the constants, yielding the left-of- $x'$  and right-of- $x'$  pieces of the complete Green's function,  $G_{x'}(x)$ .

Aside: It is amusing to notice that we use solutions to the *homogeneous* equation to construct the Green's function. We then use the Green's function to construct the *particular* solution to the given (inhomogeneous) equation. So we are ultimately constructing a particular solution from a homogeneous solution. That's not like anything we learned in undergraduate differential equations.

**When Can You Collapse a Green's Function to One Variable?**

**“Portable” Green's Functions:** When we first introduced the Green's function, we ignored boundary conditions, and our Green's function was a function of one variable,  $r$ . If our source wasn't at the origin, we just shifted our Green's function, and it was a function of just  $(r - r')$ . Then we saw that with (certain) boundary conditions, shifting doesn't work, and the Green's function is a function of two variables,  $r$  and  $r'$ . In general, then, under what conditions can we write a Green's function in the simpler form, as a function of just  $(r - r')$ ?

When both the differential operator and the boundary conditions are translation-invariant, the Green's function is also translation-invariant.

We can say it's "portable." This is fairly common: differential operators are translation-invariant (i.e., they do not explicitly depend on position), and BCs at infinity are translation-invariant. For example, in E&M it is common to have equations such as

$$-\nabla^2\phi(\mathbf{r}) = \rho(\mathbf{r}), \quad \text{with boundary condition } \phi(\infty) = 0.$$

Because both the operator  $-\nabla^2$  and the boundary conditions are translation invariant, we don't need to introduce  $r'$  explicitly as a parameter in  $G(r)$ . As we did when introducing Green's functions, we can take the origin as the location of the delta-function to find  $G(r)$ , and use translation invariance to "move around" the delta function:

$$G(r; r') \equiv G_{r'}(r) = G(r - r') \quad \text{and} \quad \mathcal{L}\{G(r - r')\} = \delta(r - r')$$

with BC  $G(\infty) = 0$

**Non-homogeneous Boundary Conditions**

So far, we've dealt with homogeneous boundary conditions by requiring  $G_{r'}(r) \equiv G(r; r')$  to be zero on the boundary. There are different kinds of boundary conditions, and different ways of dealing with each kind.

[Note that in general, constraint conditions don't have to be specified at the boundary of anything. They are really just "constraints" or "conditions." For example, one constraint is often that the solution be a "normalized" function, which is not a statement about any boundaries. But in most physical problems, at least one condition *does* occur at a boundary, so we defer to this, and limit ourselves to boundary conditions.]

**Boundary Conditions Specifying Only Values of the Solution**

In one common case, we are given a general (inhomogeneous) boundary condition,  $m(r)$  along the boundary of the region of interest. Our problem is now to find the complete solution  $c(r)$  such that

$$\mathcal{L}\{c(r)\} = s(r), \quad \text{and} \quad c(\text{boundary}) = m(\text{boundary}).$$

One approach to find  $c(r)$  is from elementary differential equations: we find a particular solution  $f(x)$  to the given equation, that doesn't necessarily meet the boundary conditions. Then we add a linear combination of *homogeneous* solutions to achieve the boundary conditions, while preserving the solution of the non-homogeneous equation. Therefore, we (1) first solve for  $f(r)$ , as above, such that

$$\mathcal{L}\{f(r)\} = s(r), \quad \text{and} \quad f(\text{boundary}) = 0, \quad \text{using a Green's function satisfying}$$

$$\mathcal{L}\{G(r; r')\} = \delta(r - r') \quad \text{and} \quad G(\text{boundary}; r') = 0$$

(2) We then find homogeneous solutions  $h_i(r)$  which are non-zero on the boundary, using ordinary methods (see any differential equations text):

$$\mathcal{L}\{h_i(r)\} = 0, \quad \text{and} \quad h_i(\text{boundary}) \neq 0.$$

Recall that in finding the Green's function, we already had to find homogeneous solutions, since every Green's function *is* a homogeneous solution everywhere except at the  $\delta$ -function position,  $r'$ .

(3) Finally, we add a linear combination of homogeneous solutions to the particular solution to yield a complete solution which satisfies both the differential equation and the boundary conditions:

$$A_1h_1(r) + A_2h_2(r) + \dots = m(r), \quad \mathcal{L}\{A_1h_1(r) + A_2h_2(r) + \dots\} = 0 \quad \text{by superposition}$$

$$c(r) = f(r) + A_1h_1(r) + A_2h_2(r) + \dots \quad \text{Therefore,}$$

$$\mathcal{L}\{c(r)\} = \mathcal{L}\{f(r) + A_1h_1(r) + A_2h_2(r) + \dots\}$$

$$= \mathcal{L}\{f(r)\} = s(r) \quad \text{and} \quad c(\text{boundary}) = m(\text{boundary})$$

**Continuing Example:** In our 1D example above, we have:

$$\mathbb{L}\{ \} = \frac{\partial^2}{\partial r^2} \quad \text{and} \quad G_{r'}(r < r') = (r' - 1)r, \quad G_{r'}(r > r') = r'(r - 1),$$

$$\text{satisfying BC: } G_{r'}(0) = G_{r'}(1) = 0 \quad \Rightarrow \quad f(0) = f(1) = 0, \quad \forall s(r)$$

We now add boundary conditions to the original problem. We must satisfy  $c(0) = 2$ , and  $c(1) = 3$ , in addition to the original problem. Our linearly independent homogeneous solutions are:

$$h_1(r) = A_1 r \quad h_0(r) = A_0 \quad (\text{a constant}).$$

To satisfy the BC, we need

$$h_1(0) + h_0(0) = 2 \quad \Rightarrow \quad A_0 = 2$$

$$h_1(1) + h_0(1) = 3 \quad \Rightarrow \quad A_1 = 1$$

and our complete solution is

$$c(r) = \left[ \int_0^1 dr' s(r') G(r; r') \right] + r + 2.$$

### Boundary Conditions Specifying a Value and a Derivative

Another common kind of boundary conditions specifies a value and a derivative for our complete solution. For example, in 1D:

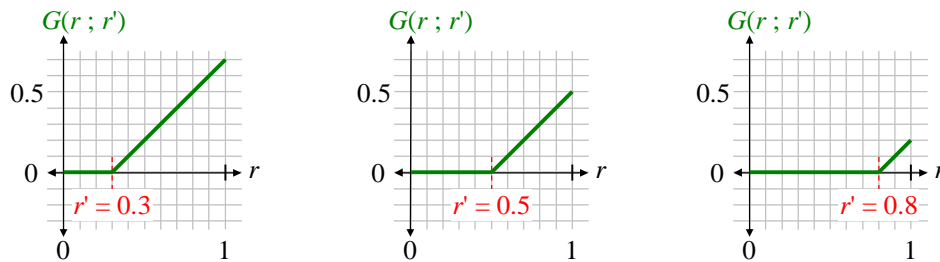
$$c(0) = 1 \quad \text{and} \quad c'(0) = 5.$$

But recall that our Green's function does not have any particular derivative at zero. When we find the particular solution,  $f(x)$ , we have no idea what its derivative at zero,  $f'(0)$ , will be. And in particular, different source functions,  $s(r)$ , will produce different  $f(r)$ , with different values of  $f'(0)$ . This is bad. In the previous case of BC,  $f(r)$  was zero at the boundaries for *any*  $s(r)$ . What we need with our new BC is  $f(0) = 0$  and  $f'(0) = 0$  for any  $s(r)$ . We can easily achieve this by using a *different Green's function!* We subjected our first Green's function to the boundary conditions  $G(0; r') = 0$  and  $G(1; r') = 0$  specifically to give the same BC to  $f(r)$ , so we could add our homogeneous solutions *independently* of  $s(r)$ . Therefore, we now choose our Green's function BC to be:

$$G(0; r') = 0 \quad \text{and} \quad G'(0; r') = 0, \quad \text{with} \quad \mathcal{L}\{G(r; r')\} = \delta(r - r').$$

We can see by inspection that this leads to a new Green's function:

$$G(r; r') = 0 \quad r < r', \quad \text{and} \quad G(r; r') = r - r' \quad r > r'.$$



The 2<sup>nd</sup> derivative of  $G(r; r')$  is everywhere 0, and the first derivative changes from 0 to 1 at  $r'$ . Therefore, our new particular solution  $f(r)$  also satisfies:

$$f(r) = \int_0^1 dr' s(r') G(r; r') \quad \text{and} \quad f(0) = 0, \quad f'(0) = 0, \quad \forall s(r).$$

We now construct the complete solution using our homogeneous solutions to meet the BC:

$$\begin{aligned}
 h_1(r) &= A_1 r & h_0(r) &= A_0 \quad (\text{a constant}) \\
 h_1(0) + h_0(0) &= 1 \Rightarrow & A_0 &= 1 \\
 h_1'(0) + h_0'(0) &= 5 \Rightarrow & A_1 &= 5. \quad \text{Then} \\
 c(r) &= \left[ \int_0^1 dr' s(r') G(r; r') \right] + 5r + 1
 \end{aligned}$$

In general, the Green's function depends not only on the particular operator, but also on the kind of boundary conditions specified.

**Boundary Conditions Specifying Ratios of Derivatives and Values**

Another kind of boundary conditions specifies a ratio of the solution to its derivative, or equivalently, specifies a linear combination of the solution and its derivative be zero. This is equivalent to a homogeneous boundary condition:

$$\frac{c'(0)}{c(0)} = \alpha \quad \text{or equivalently, if } c(0) \neq 0 \quad c'(0) - \alpha c(0) = 0.$$

This BC arises, for example, in some quantum mechanics problems where the normalization of the wave-function is not yet known; the ratio cancels any normalization factor, so the solution can proceed without knowing the ultimate normalization. Note that this is only a single BC. If our differential operator is 2<sup>nd</sup> order, there is one more degree of freedom that can be used to achieve normalization, or some other condition. (This BC is sometimes given as  $\beta c'(0) - \alpha c(0) = 0$ , but this simply multiplies both sides by a constant, and fundamentally changes nothing.)

Also, this condition is homogeneous: a linear combination of functions which satisfy the BC also satisfies the BC. This is most easily seen from the form given above, right:

$$\begin{aligned}
 \text{If} \quad d'(0) - \alpha d(0) &= 0, & \text{and} \quad e'(0) - \alpha e(0) &= 0, \\
 \text{then} \quad c(r) = Ad(r) + Be(r) & \text{satisfies } c'(0) - \alpha c(0) = 0 \\
 \text{because } c'(0) - \alpha c(0) &= A(d'(0) - \alpha d(0)) + B(e'(0) - \alpha e(0))
 \end{aligned}$$

Therefore, if we choose a Green's function which satisfies the given BC, our particular solution  $f(r)$  will also satisfy the BC. There is no need to add any homogeneous solutions.

**Continuing Example:** In our 1D example above, with  $\mathcal{L} = d^2/dr^2$ , we now specify BC:

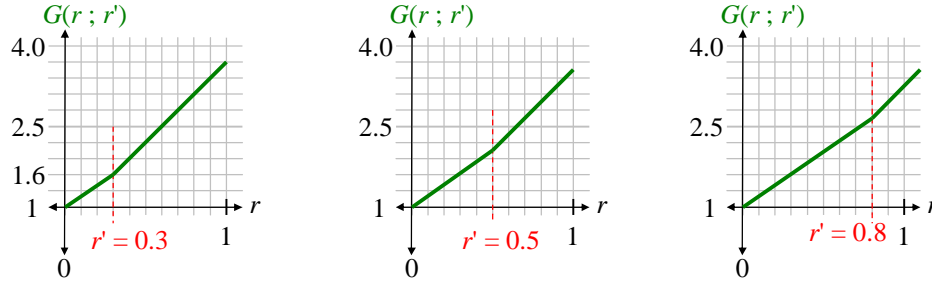
$$c'(0) - 2c(0) = 0.$$

Since our Green's functions for this operator are always two connected line segments (because their 2<sup>nd</sup> derivatives are zero), we have

$$\begin{aligned}
 r < r': \quad G(r; r') &= Cr + D, & D &\neq 0 \text{ so that } c(0) \neq 0 \\
 r > r': \quad G(r; r') &= Er + F \\
 \text{BC at } 0: & \quad C - 2D = 0
 \end{aligned}$$

With this BC, we have an unused degree of freedom, so we choose  $D = 1$ , implying  $C = 2$ . We must find  $E$  and  $F$  so that  $G(r; r')$  is continuous, and  $G'(r; r')$  has a unit step at  $r'$ . The latter condition requires that  $E = 3$ , and then continuity requires

$$\begin{aligned}
 Cr' + D = Er' + F \Rightarrow \quad 2r' + 1 &= 3r' + F, \quad F = -r' + 1. \quad \text{So} \\
 r < r': \quad G(r; r') &= 2r + 1 \quad \text{and} \quad r > r': \quad G(r; r') = 3r - r' + 1
 \end{aligned}$$



and our complete solution is just

$$c(r) = f(r) = \int_0^1 dr' s(r')G(r; r') .$$

**Boundary Conditions Specifying Only Derivatives (Neumann BC)**

Another common kind of BC specifies derivatives at points of the solution. For example, we might have

$$c'(0) = 0 \quad \text{and} \quad c'(1) = 1 .$$

Then, analogous to the BC specifying two values for  $c(\cdot)$ , we choose a Green's function which has zeros for its derivatives at 0 and 1:

$$\frac{d}{dr}G(r = 0; r') = 0 \quad \text{and} \quad \frac{d}{dr}G(r = 1; r') = 0 .$$

Then the sum (or integral) of any number of such Green's functions also satisfies the zero BC:

$$f(r) = \int_0^1 dr' s(r')G(r; r') \quad \text{satisfies} \quad f'(0) = 0 \quad \text{and} \quad f'(1) = 0 .$$

We can now form the complete solution, by adding *homogeneous* solutions that satisfy the given BC:

$$c(r) = f(r) + A_1 h_1'(r) + A_2 h_2'(r) \quad \text{where} \quad A_1 h_1'(0) + A_2 h_2'(0) = 0$$

$$\text{and} \quad A_1 h_1'(1) + A_2 h_2'(1) = 1$$

**Example:** We cannot use our previous example where  $\mathcal{L}\{ \} = d^2/dr^2$ , because there is no solution to

$$\frac{d^2}{dr^2}G(r; r') = \delta(r - r') \quad \text{with} \quad \frac{d}{dr}G(r = 0; r') = \frac{d}{dr}G(r = 1; r') = 0 .$$

This is because the homogenous solutions are straight line segments; therefore, any solution with a zero derivative at any point must be a flat line. So we choose another operator as our example:

**3D Boundary Conditions: Yet Another Method**

More TBS: why self-adjoint. Ref Canadian web site.

TBS: Using Green's theorem.

**Green-Like Methods: The Born Approximation**

In the Born approximation, and similar problems, we have our unknown function, now called  $\psi(x)$ , on both sides of the equation:

$$(1) \quad \mathcal{L}\{\psi(x)\} = \psi(x) .$$

The theory of Green's functions still works, so that

$$\psi(x) = \int \psi(x')G(x;x') dx'$$

but this doesn't solve the equation, because we still have  $\psi$  on both sides of the equation. We could try rearranging Eq (1):

$$\begin{aligned} \mathcal{L}\{\psi(x)\} - \psi(x) &= 0 && \text{which is the same as} \\ \mathcal{L}'\{\psi(x)\} &= 0, && \text{with } \mathcal{L}'\{\psi(x)\} \equiv \mathcal{L}\{\psi(x)\} - \psi(x) \end{aligned}$$

But recall that Green's functions require a source function,  $s(x)$  on the right-hand side. The method of Green's functions can't solve homogeneous equations, because it yields

$$\mathcal{L}\{\psi(x)\} = s(x) = 0 \quad \rightarrow \quad \psi(x) = \int s(x')G(x;x') dx' = \int 0 dx' = 0.$$

which is a solution, but not very useful. So Green's functions don't work when  $\psi(x)$  appears on both sides. However, under the right conditions, we can make a useful approximation. If we have an approximate solution,

$$\mathcal{L}\{\psi^{(0)}(x)\} \approx \psi^{(0)}(x),$$

then we can expand

$$\begin{aligned} \psi(x) &= \psi^{(0)}(x) + \psi^{(1)}(x) + \psi^{(2)}(x) + \dots \\ &\text{where } \psi^{(1)} \text{ is 1}^{\text{st}} \text{ order perturbation, } \psi^{(2)} \text{ is 2}^{\text{nd}} \text{ order, } \dots \end{aligned}$$

Now we can use  $\psi^{(0)}(x)$  as the source term, and use the method of Green's functions, to get a better approximation to  $\psi(x)$ :

$$\begin{aligned} \mathcal{L}\{\psi(x)\} = \psi(x) &\quad \Rightarrow \quad \psi^{(1)}(x) = \int \psi^{(0)}(x')G(x;x') dx' \\ \text{where } G(x;x') &\text{ is the Green's function for } \mathcal{L}, \text{ i.e. } \quad \mathcal{L}\{G(x;x')\} = \delta(x-x'). \end{aligned}$$

$\psi^{(0)}(x) + \psi^{(1)}(x)$  is called the **first Born approximation** of  $\psi(x)$ . Of course, this process can be repeated to arbitrarily high accuracy:

$$\psi^{(2)}(x) = \int \psi^{(1)}(x')G(x;x') dx' \quad \dots \quad \psi^{(n+1)}(x) = \int \psi^{(n)}(x')G(x;x') dx'.$$

This process assumes that the Green's function is "small" enough to produce a converging sequence. The first Born approximation is valid when  $\psi^{(1)}(x) \ll \psi^{(0)}(x)$  everywhere, and in many other, less stringent but harder to quantify, conditions. The extension to higher order approximations is straightforward: the Born approximation is valid when  $\psi^{(n)}(x) \ll \psi^{(0)}(x)$ .

TBS: a real QM example.

## 5 Complex Analytic Functions

For a review of complex numbers and arithmetic, see *Funky Quantum Concepts*.

**Notation:** In this chapter,  $z, w$  are *always* complex variables;  $x, y, r, \theta$  are *always* real variables. Other variables are defined as used.

A complex function of a complex variable  $f(z)$  is **analytic** over some domain if it has an infinite number of continuous derivatives in that domain. It turns out, if  $f(z)$  is once differentiable on a domain, then it is infinitely differentiable, and therefore analytic on that domain.

A *necessary* condition for analyticity of  $f(z) = u(x, y) + iv(x, y)$  near  $z_0$  is that the Cauchy-Riemann equations hold, to wit:

$$\frac{\partial f}{\partial x} = -i \frac{\partial f}{\partial y} \quad \Rightarrow \quad \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} = -i \left( \frac{\partial u}{\partial y} + i \frac{\partial v}{\partial y} \right) = -i \frac{\partial u}{\partial y} + \frac{\partial v}{\partial y} \quad \Rightarrow \quad \frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \text{ and } \frac{\partial v}{\partial x} = -\frac{\partial u}{\partial y}.$$

A *sufficient* condition for analyticity of  $f(z) = u(x, y) + iv(x, y)$  near  $z_0$  is that the Cauchy-Riemann equations hold, and the first partial derivatives of  $f$  exist and are continuous in a neighborhood of  $z_0$ . Note that if the first derivative of a complex function is continuous, then all derivatives are continuous, and the function is analytic. This condition implies

$$\nabla^2 u = \nabla^2 v = 0$$

$$\nabla u \cdot \nabla v = 0 \quad \Rightarrow \quad \text{"level lines" are perpendicular}$$

$$\int_{z_1}^{z_2} f(z) dz \text{ is countour independent if } f(z) \text{ is single-valued}$$

Note that a function can be analytic in some regions, but not others. **Singular** points, or **singularities**, are not in the domain of analyticity of the function, but border the domain [Det def 4.5.2 p156]. E.g.,  $\sqrt{z}$  is singular at 0, because it is not differentiable, but it is continuous at 0. **Poles** are singularities near which the function is unbounded (infinite), but can be made finite by multiplication by  $(z - z_0)^k$  for some finite  $k$  [Det p165]. This implies  $f(z)$  can be written as:

$$f(z) = a_k (z - z_0)^{-k} + a_{k-1} (z - z_0)^{-k+1} + \dots + a_{-1} (z - z_0)^{-1} + a_0 + a_1 (z - z_0)^1 + \dots$$

The value  $k$  is called the **order** of the pole. All poles are singularities. Some singularities are like "poles" of infinite order, because the function is unbounded near the singularity, but it is not a pole because it cannot be made finite by multiplication by any  $(z - z_0)^k$ , for example  $e^{1/z}$ . Such a singularity is called an **essential singularity**.

A Laurent series expansion of a function is similar to a Taylor series expansion, but the sum runs from  $-\infty$  to  $+\infty$ , instead of from 1 to  $\infty$ . In both cases, an expansion is about some point,  $z_0$ :

Taylor series:  $f(z) = f(z_0) + \sum_{n=1}^{\infty} b_n (z - z_0)^n$       where  $b_n = \frac{f^{(n)}(z_0)}{n!}$

Laurent series:  $f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n$ ,      where  $a_n = \frac{1}{2\pi i} \oint_{\text{around } z_0} \frac{f(z)}{(z - z_0)^{k+1}} dz$

[Det thm 4.6.1 p163] Analytic functions have Taylor series expansions about every point in the domain. Taylor series can be thought of as special cases of Laurent series. But analytic functions also have Laurent expansions about isolated singular points, i.e. the expansion point is not even in the domain of analyticity! The Laurent series is valid in some annulus around the singularity, but not across branch cuts. Note that in general, the  $a_k$  and  $b_k$  could be complex, but in practice, they are often real.

Properties of analytic functions:



1. If it is differentiable once, it is infinitely differentiable.
2. The Taylor and Laurent expansions are unique. This means you may use any of several methods to find them for a given function.
3. If you know a function and all its derivatives at any point, then you know the function everywhere in its domain of analyticity. This follows from the fact that every analytic function has a Laurent power series expansion. It implies that the value throughout a region is completely determined by its values at a boundary.
4. An analytic function cannot have a local extremum of absolute value. (Why not??)

## Residues

Mostly, we use complex contour integrals to evaluate difficult real integrals, and to sum infinite series. To evaluate contour integrals, we need to evaluate residues. Here, we introduce residues. The **residue** of a complex function at a complex point  $z_0$  is the  $a_{-1}$  coefficient of the Laurent expansion about the point  $z_0$ . Residues of singular points are the only ones that interest us. (In fact, residues of branch points are not defined [See sec 13.1].)

### Common ways to evaluate residues

1. The residue of a removable singularity is zero. This is because the function is bounded near the singularity, and thus  $a_{-1}$  must be zero (or else the function would blow up at  $z_0$ ):

$$\text{For } a_{-1} \neq 0, \text{ as } z \rightarrow z_0, \quad a_{-1} \frac{1}{z - z_0} \rightarrow \infty \quad \Rightarrow \quad a_{-1} = 0.$$

2. The residue of a **simple pole** at  $z_0$  (i.e., a pole of order 1) is

$$a_{-1} = \lim_{z \rightarrow z_0} (z - z_0) f(z).$$

3. Extending the previous method: the residue of a pole at  $z_0$  of order  $k$  is

$$a_{-1} = \frac{1}{(k-1)!} \lim_{z \rightarrow z_0} \frac{d^{k-1}}{dz^{k-1}} (z - z_0)^k f(z),$$

which follows by substitution of the Laurent series for  $f(z)$ , and direct differentiation. We show it here, noting that poles of order  $m$  imply that  $a_k = 0$  for  $k < -m$ , so we get:

$$f(z) = a_k (z - z_0)^{-k} + a_{k-1} (z - z_0)^{-k+1} + \dots + a_{-1} (z - z_0)^{-1} + a_0 + a_1 (z - z_0)^1 + \dots$$

$$(z - z_0)^k f(z) = a_k + a_{k-1} (z - z_0)^1 + \dots + a_{-1} (z - z_0)^{k-1} + a_0 (z - z_0)^k + a_1 (z - z_0)^{k+1} + \dots$$

$$\frac{d^{k-1}}{dz^{k-1}} (z - z_0)^k f(z) = (k-1)! a_{-1} (z - z_0)^{k-1} + \frac{k!}{1!} a_0 (z - z_0)^k + \frac{(k+1)!}{2!} a_1 (z - z_0)^{k+1} + \dots$$

$$\lim_{z \rightarrow z_0} \frac{d^{k-1}}{dz^{k-1}} (z - z_0)^k f(z) = (k-1)! a_{-1}$$

$$\Rightarrow \quad \boxed{a_{-1} = \frac{1}{(k-1)!} \lim_{z \rightarrow z_0} \frac{d^{k-1}}{dz^{k-1}} (z - z_0)^k f(z)}$$

4. If  $f(z)$  can be written as  $f(z) = \frac{P(z)}{Q(z)}$ , where  $P$  is continuous at  $z_0$ , and  $Q'(z_0) \neq 0$  (and is continuous at  $z_0$ ), then  $f(z)$  has a simple pole at  $z_0$ , and

$$\text{Res}_{z=z_0} f(z) = \frac{P(z_0)}{\left. \frac{d}{dz} Q(z) \right|_{z_0}} \equiv \frac{P(z_0)}{Q'(z_0)}. \quad \text{Why? Near } z_0, \quad Q(z) \approx (z - z_0)Q'(z_0).$$

$$\text{Then: } \text{Res}_{z=z_0} f(z) = \lim_{z \rightarrow z_0} (z - z_0) f(z) = \lim_{z \rightarrow z_0} (z - z_0) \frac{P(z_0)}{(z - z_0)Q'(z_0)} = \frac{P(z_0)}{Q'(z_0)}.$$

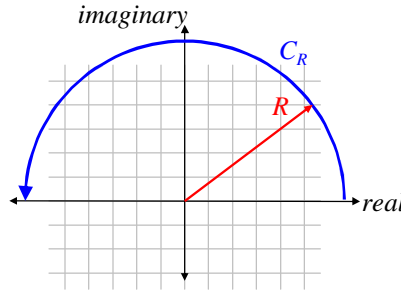
- Find the Laurent series, and hence its coefficient of  $(z - z_0)^{-1}$ . This is sometimes easy if  $f(z)$  is given in terms of functions with well-known power series expansions. See the sum of series example later.

We will include real-life examples of most of these as we go.

## Contour Integrals

Contour integration is an invaluable tool for evaluating both real and complex-valued integrals. Contour integrals are used all over advanced physics, and we could not do physics as we know it today without them. Contour integrals are mostly useful for evaluating difficult ordinary (real-valued) integrals, and sums of series. In many cases, a function is analytic except at a set of distinct points. In this case, a contour integral may enclose, or pass near, some points of non-analyticity, i.e. **singular** points. It is these singular points that allow us to evaluate the integral.

You often let the radius of the contour integral go to  $\infty$  for some part of the contour:



Any arc where

$$\lim_{R \rightarrow \infty} |f(z)| \rightarrow \sim \frac{1}{|z|^{1+\epsilon}}, \quad \epsilon > 0.$$

has an integral of 0 over the arc.

Beware that this is often stated incorrectly as “any function which goes to zero *faster* than  $1/|z|$  has a contour integral of 0.” The problem is that it has to have an *exponent*  $< -1$ ; it is *not* sufficient to be simply smaller than  $1/|z|$ . E.g.  $\frac{1}{|z|+1} < \frac{1}{|z|}$ , but the contour integral still diverges.

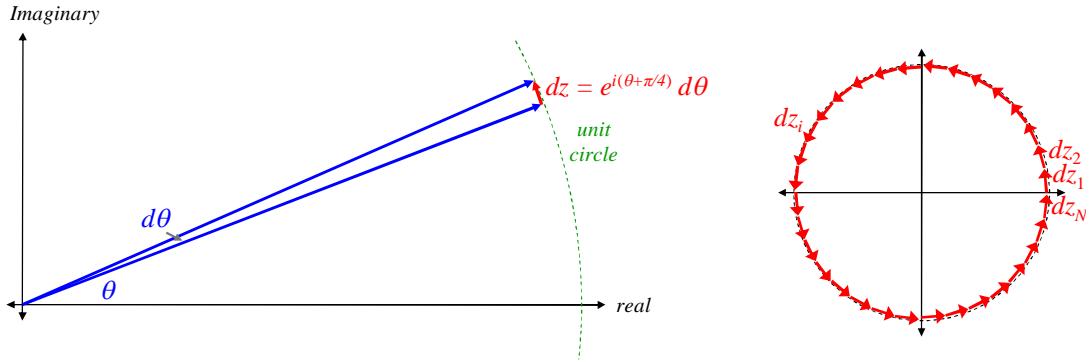
Jordan’s lemma: ??.

## Evaluating Integrals

Surprisingly, we can use complex contour integrals to evaluate difficult real integrals. The main point is to find a contour which (a) includes some known (possibly complex) multiple of the desired (real) integral, (b) includes other segments whose values are zero, and (c) includes a known set of poles whose residues can be found. Then you simply plug into the residue theorem:

$$\oint_C f(z) dz = 2\pi i \sum_{n \text{ residues } z_n} \text{Res}_{z_n} f(z), \text{ where } z_n \text{ are the finite set of isolated singularities.}$$

We can see this by considering the contour integral around the unit circle for each term in the Laurent series expanded about 0. First, consider the  $z^0$  term (the constant term). We seek the value of  $\oint_0 dz$ .  $dz$  is a small complex number, representable as a vector in the complex plane. The diagram below (left) shows the geometric meaning of  $dz$ . Below (right) shows the geometric approximation to the desired integral.



(Left) Geometric description of  $dz$ .

(Right) Approximation of  $\oint_0 dz$  as a sum of 32 small complex terms (vectors).

We see that all the tiny  $dz$  elements add up to zero: the vectors add head-to-tail, and circle back to the starting point. The sum vector (displacement from start) is zero. This is true for any large number of  $dz$ , so we have

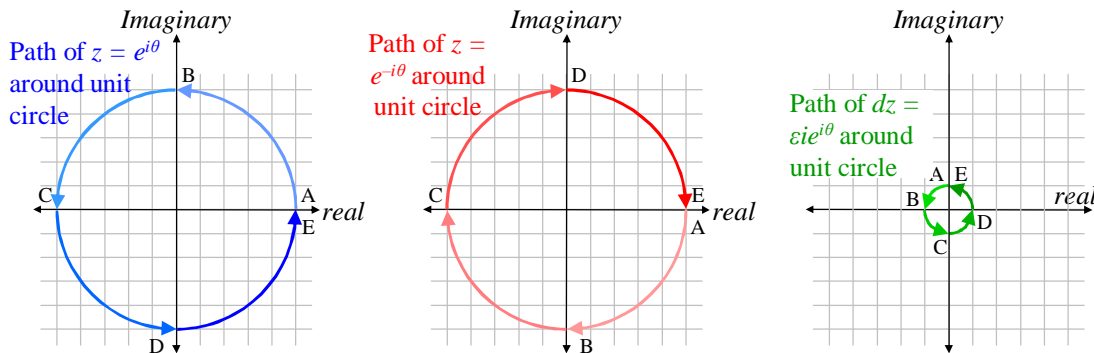
$$\oint_0 dz = 0.$$

Next, consider the  $z^{-1}$  term,  $\oint_0 \left(\frac{1}{z}\right) dz$ , and a change of integration variable to  $\theta$ :

$$\text{Let } z = e^{i\theta}, dz = ie^{i\theta} d\theta: \quad \oint_0 \left(\frac{1}{z}\right) dz = \int_0^{2\pi} e^{-i\theta} ie^{i\theta} d\theta = \int_0^{2\pi} id\theta = 2\pi i.$$

The change of variable maps the complex contour and  $z$  into an ordinary integral of a real variable.

Geometrically, as  $z$  goes positively (counter-clockwise) around the unit circle (below left),  $z^{-1}$  goes around the unit circle in the negative (clockwise) direction (below middle). Its complex angle,  $\arg(1/z) = -\theta$ , where  $z = e^{i\theta}$ . As  $z$  goes around the unit circle,  $dz$  has infinitesimal magnitude  $\epsilon = d\theta$ , and argument  $\theta + \pi/4$ . Hence, the product of  $(1/z) dz$  always has argument of  $-\theta + \theta + \pi/4 = \pi/4$ ; it is always purely imaginary.



Paths of  $z$ ,  $1/z$ , and  $dz$  in the complex plane

The magnitude of  $(1/z) dz = d\theta$ ; thus the integral around the circle is  $2\pi i$ . Multiplying the integrand by some constant,  $a_{-1}$  (the residue), just multiplies the integral by that constant. And any contour integral that encloses the pole  $1/z$  and no other singularity has the same value. Hence, for any contour around the origin

$$\oint_O a_{-1} z^{-1} dz = 2\pi i(a_{-1}) \Rightarrow a_{-1} = \frac{\oint_O a_{-1} z^{-1} dz}{2\pi i}.$$

Now consider the other terms of the Laurent expansion of  $f(z)$ . We already showed that the  $a_0 z^0$  term, which on integration gives the product  $a_0 dz$ , rotates uniformly about all directions, in the positive (counter-clockwise) sense, and sums to zero. Hence the  $a_0$  term contributes nothing to the contour integral.

The  $a_1 z^1 dz$  product rotates uniformly *twice* around all directions in the positive sense, and of course, still sums to zero. Higher powers of  $z$  simply rotate more times, but always an integer number of times around the circle, and hence always sum to zero.

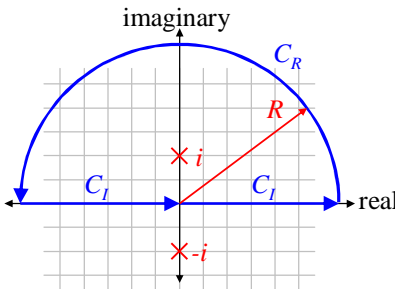
Similarly,  $a_{-2} z^{-2}$ , and all more negative powers, rotate uniformly about all directions, but in the *negative* (clockwise) sense. Hence, all these terms contribute nothing to the contour integral.

So in the end:

The only term of the Laurent expansion about 0 that contributes to the contour integral is the residue term,  $a_{-1} z^{-1}$ .

**The simplest contour integral:** Evaluate  $I = \int_0^\infty \frac{1}{x^2 + 1} dx$ .

We know from elementary calculus (let  $x = \tan u$ ) that  $I = \pi/2$ . We can find this easily from the residue theorem, using the following contour:



“C” denotes a contour, and “I” denotes the integral over that contour. We let the radius of the arc go to infinity, and we see that the closed contour integral  $I_C = I + I + I_R$ . But  $I_R = 0$ , because  $f(R \rightarrow \infty) < 1/R^2$ . Then  $I = I_C / 2$ .  $f(z)$  has poles at  $\pm i$ . The contour encloses one pole at  $i$ . Its residue is

$$\text{Res } f(i) = \frac{1}{\left. \frac{d}{dz}(z^2 + 1) \right|_{z=i}} = \frac{1}{2z|_{z=i}} = \frac{1}{2i}.$$

$$I_C = 2\pi i \sum_n \text{Res } f(z_n) = 2\pi i \frac{1}{2i} = \pi$$

$$I = \frac{I_C}{2} = \frac{\pi}{2}$$

Note that when evaluating a real integral with complex functions and contour integrals, the  $i$ 's always cancel, and you get a real result, as you must. It's a good check to make sure this happens.

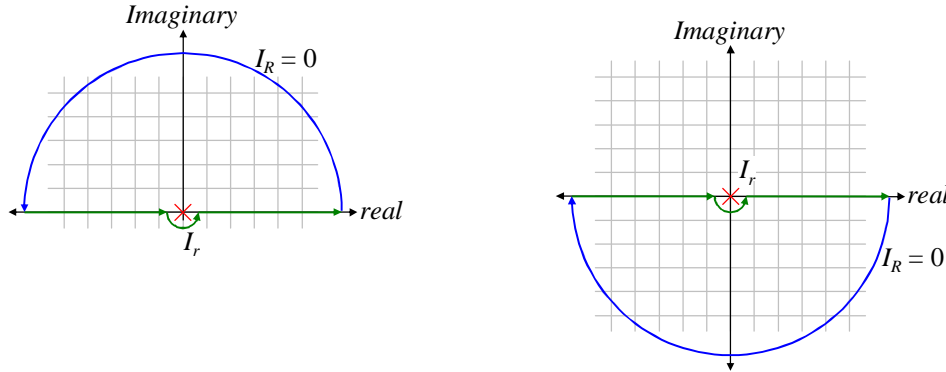
### Choosing the Right Path: Which Contour?

The path of integration is fraught with perils. How will I know which path to choose? There is no universal answer. Often, many paths lead to the same truth. Still, many paths lead nowhere. All we can do is use wisdom as our guide, and take one step in a new direction. If we end up where we started, we are grateful for what we learned, and we start anew.

We here examine several useful and general, but oft neglected, methods of contour integration. We use a some sample problems to illustrate these tools. This section assumes a familiarity with contour integration, and its use in evaluating definite integrals, including the residue theorem.

**Example:** Evaluate  $I = \int_{-\infty}^{\infty} \frac{\sin^2 x}{x^2} dx$ .

The integrand is everywhere nonnegative, and somewhere positive, and it is in the positive direction, so  $I$  must be positive. We observe that the given integrand has no poles. It has only a removable singularity at  $x = 0$ . If we are to use contour integrals, we must somehow create a pole (or a few), to use the residue theorem. Simple poles (i.e. 1<sup>st</sup>-order) are sometimes best, because then we can also use the indented contour theorem.



Contours for the two exponential integrals: (left) positive (counter-clockwise)  $\exp(2z)$ ; (right) negative (clockwise)  $\exp(-2z)$

To use a contour integral (which, a priori, may or may not be a good idea), we must do two things: (1) create a pole; and (2) close the contour. The same method does both: expand the  $\sin(\ )$  in terms of exponentials:

$$I = \int_{-\infty}^{\infty} \frac{\sin^2 x}{x^2} dx = \int_{-\infty}^{\infty} \frac{(e^{iz} - e^{-iz})^2}{(2i)^2 z^2} dz = -\frac{1}{4} \left[ \int_{-\infty}^{\infty} \frac{e^{i2z}}{z^2} dz - \int_{-\infty}^{\infty} \frac{2}{z^2} dz + \int_{-\infty}^{\infty} \frac{e^{-i2z}}{z^2} dz \right].$$

All three integrals have poles at  $z = 0$ . If we indent the contour underneath the origin, then since the function is bounded near there, the limit as  $r \rightarrow 0$  leaves the original integral unchanged (above left). The first integral must be closed in the upper half-plane, to keep the exponential small. The second integral can be closed in either half-plane, since it  $\sim 1/z^2$ . The third integral must be closed in the lower half-plane, again to keep the exponential small (above right). Note that all three contours must use an indentation that preserves the value of the original integral. An easy way to insure this is to use the same indentation on all three.

Now the third integral encloses no poles, so is zero. The 2<sup>nd</sup> integral, by inspection of its Laurent series, has a residue of zero, so is also zero. Only the first integral contributes. By expanding the exponential in a Taylor series, and dividing by  $z^2$ , we find its residue is  $2i$ . Using the residue theorem, we have:

$$I = \int_{-\infty}^{\infty} \frac{\sin^2 x}{x^2} dx = -\frac{1}{4} [2\pi i (2i)] = \pi.$$

**Example:** Evaluate  $I = \int_0^{\infty} \frac{\cos(ax) - \cos(bx)}{x^2} dx$  [B&C p?? Q1].

This innocent looking problem has a number of funky aspects:

- The integrand is two terms. Separately, each term diverges. Together, they converge.
- The integrand is even, so if we choose a contour that includes the whole real line, the contour integral includes twice the integral we seek (twice  $I$ ).
- The integrand has no poles. How can we use any residue theorems if there are no poles? Amazingly, we can create a useful pole.
- A typical contour includes an arc at infinity, but  $\cos(z)$  is ill-behaved for  $z$  far off the real-axis. How can we tame it?
- We will see that this integral leads to the indented contour theorem, which can only be applied to simple poles, i.e., first order poles (unlike the residue theorem, which applies to all poles).

Each of these funky features is important, and each arises in practical real-world integrals. Let us consider each funkiness in turn.

**1. The integrand is two terms. Separately, each term diverges. Together, they converge.**

Near zero,  $\cos(x) \approx 1$ . Therefore, the zero endpoint of either term of the integral looks like

$$\int_0^{\text{anywhere}} \frac{\cos ax}{x^2} dx \sim \int_0^{\text{anywhere}} \frac{1}{x^2} dx = -\frac{1}{x} \Big|_0^{\text{anywhere}} \rightarrow +\infty.$$

Thus each term, separately, diverges. However, the difference is finite. We see this by power series expanding  $\cos(x)$ :

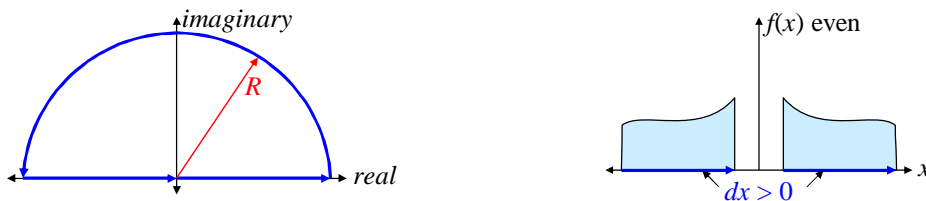
$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \dots \Rightarrow \cos(ax) - \cos(bx) = -\frac{a^2x^2}{2} + \frac{b^2x^2}{2} + O(x^4) \quad \text{and}$$

$$\frac{\cos(ax) - \cos(bx)}{x^2} = -\frac{a^2}{2} + \frac{b^2}{2} + O(x^2) = \frac{b^2 - a^2}{2} + O(x^2) \Rightarrow$$

$$\int_0^{\text{anywhere}} \frac{\cos(ax) - \cos(bx)}{x^2} dx \sim \frac{b^2 - a^2}{2} \quad \text{which is to say, is finite.}$$

**2. The integrand is even, so if we choose a contour that includes the whole real line, the contour integral includes twice the integral we seek (twice  $I$ ).**

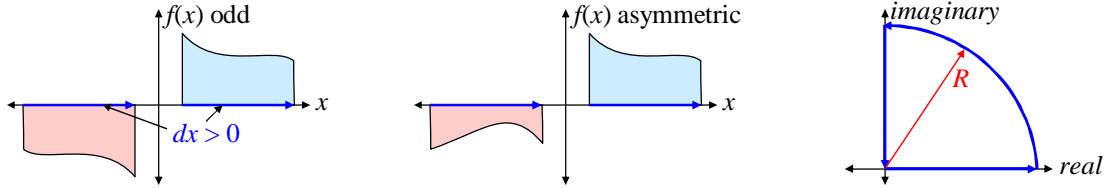
Perhaps the most common integration contour (below left) covers the real line, and an infinitely distant arc from  $+\infty$  back to  $-\infty$ . When our real integral ( $I$  in this case) is only from 0 to  $\infty$ , the contour integral includes more than we want on the real axis. If our integrand is even, the contour integral includes twice the integral we seek (twice  $I$ ). This may seem trivial, but the point to notice is that when integrating from  $-\infty$  to 0,  $dx$  is still positive (below middle).



(Left) A common contour.

(Right) An even function has integral over the real-line twice that of 0 to infinity.

Note that if the integrand is odd (below left), choosing this contour cancels out the original (real) integral from our contour integral, and the contour is of no use. Or if the integrand has no even/odd symmetry (below middle), then this contour tells us nothing about our desired integral. In these cases, a different contour may work, for example, one which only includes the positive real axis (below right).



(Left) An odd function has zero integral over the real line. (Middle) An asymmetric function has unknown integral over the real line. (Right) A contour containing only the desired real integral.

**3. The integrand has no poles. How can we use any residue theorems if there are no poles? Amazingly, we can create a useful pole.**

This is the funkiest aspect of this problem, but illustrates a standard tool. We are given a real-valued integral with no poles. Contour integration is usually useless without a pole, and a residue, to help us evaluate the contour integral. Our integrand contains  $\cos(x)$ , and that is related to  $\exp(ix)$ . We could try replacing cosines with exponentials,

$$\cos z = \frac{\exp(iz) + \exp(-iz)}{2} \quad (\text{does no good}).$$

but this only rearranges the algebra; fundamentally, it buys us nothing. The trick here is to notice that we can often add a made-up imaginary term to our original integrand, perform a contour integration, and then simply take the real part of our result:

$$\text{Given } I = \int_a^b g(x) dx, \text{ let } f(z) \equiv g(z) + ih(z). \text{ Then } I = \text{Re} \left\{ \int_a^b f(z) dz \right\}.$$

For this trick to work,  $ih(z)$  must have no real-valued contribution over the contour we choose, so it doesn't mess up the integral we seek. Often, we satisfy this requirement by choosing  $ih(z)$  to be purely imaginary on the real axis, and having zero contribution elsewhere on the contour. Given an integrand containing  $\cos(x)$ , as in our example, a natural choice for  $ih(z)$  is  $i \sin(z)$ , because then we can write the new integrand as a simple exponential:

$$\cos(x) \rightarrow f(z) = \cos(z) + i \sin(z) = \exp(iz).$$

In our example, the corresponding substitution yields

$$I = \int_0^\infty \frac{\cos ax - \cos bx}{x^2} dx \rightarrow I = \text{Re} \left\{ \int_0^\infty \frac{\exp(iax) - \exp(ibx)}{x^2} dx \right\}.$$

Examining this substitution more closely, we find a wonderful consequence: this substitution introduced a pole! Recall that

$$\sin z = z - \frac{z^3}{3!} + \dots \Rightarrow \frac{i \sin z}{z^2} = i \left( \frac{1}{z} - \frac{z}{3!} + \dots \right).$$

We now have a simple pole at  $z = 0$ , with residue  $i$ .

By choosing to add an imaginary term to the integrand, we now have a pole that we can work with to evaluate a contour integral!

It's like magic. In our example integral, our residue is:

$$\frac{i \sin az - i \sin bz}{z^2} = i \left( \frac{a-b}{z} + \dots \right), \text{ and } \text{residue} = i(a-b).$$

Note that if our original integrand contained  $\sin(x)$  instead of  $\cos(x)$ , we would have made a similar substitution, but taken the *imaginary* part of the result:

Given  $I = \int_a^b \sin(x) dx$ , let  $f(z) \equiv \cos(z) + i \sin(z)$ . Then  $I = \text{Im} \left\{ \int_a^b f(z) dz \right\}$ .

**4. A typical contour includes an arc at infinity, but  $\cos(z)$  is ill-behaved for  $z$  far off the real axis. How can we tame it?**

This is related to the previous funkiness. We're used to thinking of  $\cos(x)$  as a nice, bounded, well-behaved function, but this is only true when  $x$  is real.

When integrating  $\cos(z)$  over a contour, we must remember that  $\cos(z)$  blows up rapidly off the real axis.

In fact,  $\cos(z) \sim \exp(\text{Im}\{z\})$ , so it blows up extremely quickly off the real axis. If we're going to evaluate a contour integral with  $\cos(z)$  in it, we must cancel its divergence off the real axis. There is only one function which can exactly cancel the divergence of  $\cos(z)$ , and that is  $\pm i \sin(z)$ . The plus sign cancels the divergence *above* the real axis; the minus sign cancels it *below*. There is nothing that cancels it everywhere. We show this cancellation simply:

Let  $z \equiv x + iy$   
 $\cos z + i \sin z = \exp(iz) = \exp(i(x + iy)) = \exp(ix) \exp(-y)$  and  
 $|\exp(ix) \exp(-y)| = |\exp(ix)| \cdot |\exp(-y)| = \exp(-y)$

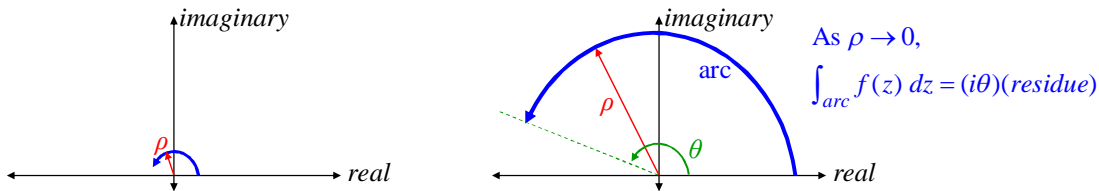
For  $z$  above the real axis, this shrinks rapidly. Recall that in the previous step, we added  $i \sin(x)$  to our integrand to give us a pole to work with. We see now that we also need the *same* additional term to tame the divergence of  $\cos(z)$  off the real axis. For the contour we've chosen, no other term will work.

**5. We will see that this integral leads to the indented contour theorem, which can only be applied to simple poles, i.e., first order poles (unlike the residue theorem, which applies to all poles).**

We're now at the final step. We have a pole at  $z = 0$ , but it is *right on* our contour, not inside it. If the pole were inside the contour, we would use the residue theorem to evaluate the contour integral, and from there, we'd find the integral on the real axis, cut it in half, and take the real part. That is the integral we seek.

But the pole is not inside the contour; it is *on* the contour. The indented contour theorem allows us to work with poles on the contour. We explain the theorem geometrically in the next section, but state it briefly here:

Indented contour theorem: For a simple pole, the integral of an arc of tiny radius around the pole, of angle  $\theta$ , equals  $(i\theta)(\text{residue})$ . See diagram below.



(Left) A tiny arc around a simple pole. (Right) A magnified view; we let  $\rho \rightarrow 0$ .

Note that if we encircle the pole completely,  $\theta = 2\pi$ , and we have the special case of the residue theorem for a simple pole:

$$\oint f(z) dz = 2\pi i (\text{residue}).$$

However, the residue theorem is true for *all* poles, not just simple ones (see The Residue Theorem earlier).

**Putting it all together:** We now solve the original integral using all of the above methods. First, we add  $i \sin(z)$  to the integrand, which is equivalent to replacing  $\cos(z)$  with  $\exp(iz)$ :



$$I = \int_0^\infty \frac{\cos ax - \cos bx}{x^2} dx \rightarrow I = \text{Re} \left\{ \int_0^\infty \frac{\exp(iax) - \exp(ibx)}{x^2} dx \right\}$$

Define  $J \equiv \int_0^\infty \frac{\exp(iax) - \exp(ibx)}{x^2} dx$ , so  $I = \text{Re}\{J\}$

We choose the contour shown below left, with  $R \rightarrow \infty$ , and  $\rho \rightarrow 0$ .



There are no poles enclosed, so the contour integral is zero. The contour includes twice the desired integral, so define:

$$f(z) \equiv \frac{\exp(iaz) - \exp(ibz)}{z^2}. \quad \text{Then} \quad \oint f(z) dz = \int_{C_R} f(z) dz + 2J + \int_{C_\rho} f(z) dz = 0. \quad (5.1)$$

For  $C_R$ ,  $|f(z)| < 1/R^2$ , so as  $R \rightarrow \infty$ , the integral goes to 0. For  $C_\rho$ , the residue is  $i(a - b)$ , and the arc is  $\pi$  radians in the negative direction, so the indented contour theorem says:

$$\lim_{\rho \rightarrow 0} \int_{C_\rho} f(z) dz = -(\pi i) i(a - b) = \pi(a - b).$$

Plugging into (5.1), we finally get

$$2J + \pi(a - b) = 0 \quad \Rightarrow \quad I = \text{Re}\{J\} = \frac{\pi}{2}(b - a).$$

In this example, the contour integral  $J$  happened to be real, so taking  $I = \text{Re}\{J\}$  is trivial, but in general, there's no reason why  $J$  must be real. It could well be complex, and we would need to take the real part of it.

To illustrate this and more, we evaluate the integral again, now with the alternate contour shown above right. Again, there are no poles enclosed, so the contour integral is zero. Again, the integral over  $C_R = 0$ . We then have:

$$\oint f(z) dz = \int_{C_R} f(z) dz + \int_{C_2} f(z) dz + J + \int_{C_\rho} f(z) dz = 0$$

$$\text{And} \quad \lim_{\rho \rightarrow 0} \int_{C_\rho} f(z) dz = -(i\pi/2) i(a - b) = \frac{\pi}{2}(a - b)$$

The integral over  $C_2$  is down the imaginary axis:

$$\text{Let } z = x + iy = 0 + iy = iy, \quad \text{then } dz = i dy$$

$$\int_{C_2} f(z) dz = \int_{C_2} \frac{\exp(iaz) - \exp(ibz)}{z^2} dz = \int_\infty^0 \frac{\exp(-ay) - \exp(-by)}{-y^2} i dy$$

We don't know what this integral is, but *we don't care!* In fact, it is divergent, but we see that it is purely imaginary, so will contribute only to the imaginary part of  $J$ . But we seek  $I = \text{Re}\{J\}$ , and therefore

$$I = \lim_{\rho \rightarrow 0} \text{Re}\{J\} \quad \text{is well-defined.}$$

Therefore we ignore the divergent imaginary contribution from  $C_2$ . We then have

$$i(\text{something}) + J + \frac{\pi}{2}(a - b) = 0 \quad \Rightarrow \quad I = \text{Re}\{J\} = \frac{\pi}{2}(b - a).$$

as before.

## Evaluating Infinite Sums

Perhaps the simplest infinite sum in the world is  $S = \sum_{n=1}^{\infty} \frac{1}{n^2}$ . The general method for using contour integrals is to find an countably infinite set of residues whose values are the terms of the sum, and whose contour integral can be evaluated by other means. Then

$$I_C = 2\pi i \sum_{n=1}^{\infty} \text{Res } f(z_n) = 2\pi i S \quad \Rightarrow \quad S = \frac{I_C}{2\pi i}.$$

The hard part is finding the function  $f(z)$  that has the right residues. Such a function must first have poles at all the integers, and then also have residues at those poles equal to the terms of the series.

To find such a function, consider the complex function  $\pi \cot(\pi z)$ . Clearly, this has poles at all real integer  $z$ , due to the  $\sin(\pi z)$  function in the denominator of  $\cot(z)$ . Hence,

$$\text{For } z_n = n \text{ (integer),} \quad \text{Res}[\pi \cot(z_n)] = \text{Res}\left[\pi \frac{\cos(\pi z_n)}{\sin(\pi z_n)}\right] = \pi \frac{\cos(\pi z_n)}{\pi \cos(\pi z_n)} = 1,$$

where in the last step we used if  $Q(z) = 0$  then  $\text{Res}_{z=z_0} \frac{P(z)}{Q(z)} = \frac{P(z)}{Q'(z_0)}$ , if this is defined.

Thus  $\pi \cot(\pi z)$  can be used to generate lots of infinite sums, by simply multiplying it by a continuous function of  $z$  that equals the terms of the infinite series when  $z$  is integer. For example, for the sum above,

$$S = \sum_{n=1}^{\infty} \frac{1}{n^2}, \text{ we simply define:}$$

$$f(z) = \frac{1}{z^2} \pi \cot(\pi z), \quad \text{and its residues are} \quad \text{Res } f(z_n) = \frac{1}{n^2}, \quad n \neq 0.$$

[In general, to find  $\sum_{n=1}^{\infty} s(n)$ , define

$$f(z) = s(z)[\pi \cot(\pi z)], \quad \text{and its residues are} \quad \text{Res}_{z=n} f(z) = s(n).$$

However, now you may have to deal with the residues for  $n \leq 0$ .]

Continuing our example, now we need the residue at  $n = 0$ . Since  $\cot(z)$  has a simple pole at zero,  $\cot(z)/z^2$  has a 3<sup>rd</sup> order pole at zero. We optimistically try tedious brute force for an  $m^{\text{th}}$  order pole with  $m = 3$ , only to find that it fails:

$$\begin{aligned} \text{Res}_{z=0} \frac{\pi \cot \pi z}{z^2} &= \lim_{z \rightarrow 0} \left[ \frac{1}{2!} \frac{d^2}{dz^2} z^3 \frac{\pi \cot \pi z}{z^2} \right] = \lim_{z \rightarrow 0} \left[ \frac{1}{2!} \frac{d^2}{dz^2} \pi z \cot \pi z \right] \\ &= \frac{\pi}{2} \lim_{z \rightarrow 0} \frac{d}{dz} \left[ \cot \pi z - \pi z \csc^2 \pi z \right] = \frac{\pi}{2} \lim_{z \rightarrow 0} \frac{d}{dz} \left[ \frac{\cos \pi z \sin \pi z - \pi z}{\sin^2 \pi z} \right] = \frac{\pi}{2} \lim_{z \rightarrow 0} \frac{d}{dz} \left[ \frac{\frac{1}{2} \sin 2\pi z - \pi z}{\sin^2 \pi z} \right] \end{aligned}$$

Use  $d\frac{U}{V} = \frac{VdU - UdV}{V^2}$ :

$$\begin{aligned} \operatorname{Res}_{z=0} \frac{\pi \cot \pi z}{z^2} &= \frac{\pi}{2} \lim_{z \rightarrow 0} \frac{\sin^2 \pi z (\pi \cos 2\pi z - \pi) - \left(\frac{1}{2} \sin 2\pi z - \pi z\right) 2\pi \sin \pi z \cos \pi z}{\sin^4 \pi z} \\ &= \frac{\pi}{2} \lim_{z \rightarrow 0} \frac{\sin \pi z (\pi \cos 2\pi z - \pi) - \left(\frac{1}{2} \sin 2\pi z - \pi z\right) 2\pi \cos \pi z}{\sin^3 \pi z} \end{aligned}$$

Use L'Hopital's rule:

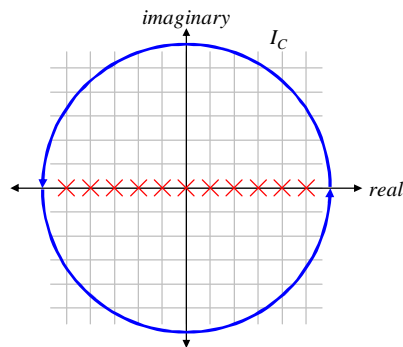
$$\begin{aligned} \operatorname{Res}_{z=0} \frac{\pi \cot \pi z}{z^2} &= \frac{\pi}{2} \lim_{z \rightarrow 0} \frac{1}{3\pi \sin^2 \pi z \cos \pi z} \left[ \pi \cos \pi z (\pi \cos 2\pi z - \pi) + \sin \pi z (-2\pi \sin 2\pi z - 1) \right. \\ &\quad \left. - (\pi \cos 2\pi z - \pi) 2\pi \cos \pi z - \left(\frac{1}{2} \sin 2\pi z - \pi z\right) 2\pi^2 \sin \pi z \right] \\ &= \frac{\pi}{2} \lim_{z \rightarrow 0} \frac{-\pi^2 \cos \pi z (\cos 2\pi z - 1) + \sin \pi z (-2\pi \sin 2\pi z - 1) - 2\pi^2 \left(\frac{1}{2} \sin 2\pi z - \pi z\right) \sin \pi z}{3\pi \sin^2 \pi z \cos \pi z} \end{aligned}$$

At this point, we give up on brute force, because we see from the denominator that we'll have to use L'Hopital's rule twice more to eliminate the zero there, and the derivatives will get untenably complicated.

But in 2 lines, we can find the  $a_{-1}$  term of the Laurent series from the series expansions of  $\sin$  and  $\cos$ . The  $z^1$  coefficient of  $\cot(z)$  becomes the  $z^1$  coefficient of  $f(z) = \cot(z)/z^2$ :

$$\begin{aligned} \cot z &= \frac{\cos z}{\sin z} \approx \frac{1 - z^2/2 + \dots}{z - z^3/6 + \dots} = \left(\frac{1}{z}\right) \frac{1 - z^2/2}{1 - z^2/6} \approx \left(\frac{1}{z}\right) (1 - z^2/2)(1 + z^2/6) \approx \left(\frac{1}{z}\right) (1 - z^2/3) = \frac{1}{z} - \frac{z}{3} \\ \cot \pi z &\approx \frac{1}{\pi z} - \frac{\pi z}{3} \quad \Rightarrow \quad \operatorname{Res}_{z=0} \pi \frac{\cot \pi z}{z^2} = -\frac{\pi^2}{3} \end{aligned}$$

Now we take a contour integral over a circle centered at the origin: (no good, because  $\cot(\pi z)$  blows up every integer !??)



As  $R \rightarrow \infty, I_C \rightarrow 0$ . Hence:

$$I_C = 0 = 2\pi i \left( \sum_{n=-1}^{-\infty} \frac{1}{n^2} + K_0 + \sum_{n=1}^{\infty} \frac{1}{n^2} \right) \quad \Rightarrow \quad K_0 + 2 \sum_{n=1}^{\infty} \frac{1}{n^2} = 0, \quad \sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{-K_0}{2} = \frac{\pi^2}{6}.$$

## Multi-valued Functions

Many functions are **multi-valued** (despite the apparent oxymoron), i.e. for a single point in the domain, the function can have multiple values. An obvious example is a square-root function: given a complex number, there are two complex square roots of it. Thus, the square root function is two-valued. Another example is arc-tangent: given any complex number, there are an infinite number of complex numbers whose tangent is the given complex number.

[picture??]

We refer now to “nice” functions, which are locally (i.e., within any small finite region) analytic, but multi-valued. If you’re not careful, such “multi-valuedness” can violate the assumptions of analyticity, by introducing discontinuities in the function. Without analyticity, all our developments break down: no contour integrals, no sums of series. But, you can avoid such a breakdown, and preserve the tools we’ve developed, by treating multi-valued functions in a slightly special way to insure continuity, and therefore analyticity.

A **regular** function, or region, is analytic and single valued. (You can get a regular function from a multi-valued one by choosing a Riemann sheet. More below.)

A **branch point** is a point in the domain of a function  $f(z)$  with this property: when you traverse a closed path around the branch point, following continuous values of  $f(z)$ ,  $f(z)$  has a different value at the end point of the path than at the beginning point, even though the beginning and end point are the same point in the domain. Example TBS: square root around the origin. Sometimes branch points are also singularities.

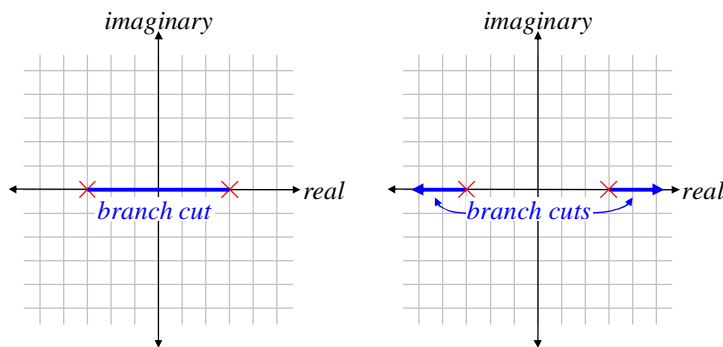
A **branch cut** is an arbitrary (possibly curved) path connecting branch points, or running from a branch point to infinity (“connecting” the branch point to infinity). If you now evaluate integrals of contours that never cross the branch cuts, you insure that the function remains continuous (and thus analytic) over the domain of the integral.

When the contour of integration is entirely in the domain of analyticity of the integrand, “ordinary” contour integration, and the residue theorem, are valid.

This solves the problem of integrating across discontinuities. Branch cuts are like fences in the domain of the function: your contour integral can’t cross them. Note that you’re free to choose your branch cuts wherever you like, so long as the function remains continuous when you don’t cross the branch cuts. Connecting branch points is one way to insure this.

A **Riemann sheet** is the complex plane plus a choice of branch cuts, and a choice of branch. This defines a domain on which a function is regular.

A **Riemann surface** is a continuous joining of Riemann sheets, gluing the edges together. This “looks like” sheets layered on top of each other, and each sheet represents one of the multiple values a multi-valued analytic function may have. TBS: consider  $\sqrt{(z-a)(z-b)}$ .



## 6 Conceptual Linear Algebra

Instead of lots of summation signs, we describe linear algebra concepts, visualizations, and ways to think about linear operations as algebraic operations. This allows fast understanding of linear algebra methods that is extremely helpful in almost all areas of physics. Tensors rely heavily on linear algebra methods, so this section is a good warm-up for tensors. Matrices and linear algebra are also critical for quantum mechanics.

**Caution** In this section, **vector** means a column or row of numbers. In other sections, “vector” has a more general meaning.

In this section, we use bold capitals for matrices (**A**), and bold lower-case for vectors (**a**).

### Matrix Multiplication

It is often helpful to view a matrix as a horizontal concatenation of column-vectors. You can think of it as a row-vector, where each element of the row-vector is itself a column vector.

$$\mathbf{A} = \left[ \begin{array}{c|c|c} & & \\ \hline \mathbf{a} & \mathbf{b} & \mathbf{c} \\ \hline & & \end{array} \right] \quad \text{or} \quad \mathbf{A} = \left[ \begin{array}{c} \mathbf{d} \\ \mathbf{e} \\ \mathbf{f} \end{array} \right].$$

Equally valid, you can think of a matrix as a vertical concatenation of row-vectors, like a column-vector where each element is itself a row-vector.

Matrix multiplication is defined to be the operation of linear transformation, e.g., from one set of coordinates to another. The following properties follow from the standard definition of matrix multiplication:

**Matrix times a vector:** A matrix **B** times a column vector **v**, is a weighted sum of the columns of **B**:

$$\mathbf{B}\mathbf{v} = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} \equiv v^x \begin{bmatrix} B_{11} \\ B_{21} \\ B_{31} \end{bmatrix} + v^y \begin{bmatrix} B_{12} \\ B_{22} \\ B_{32} \end{bmatrix} + v^z \begin{bmatrix} B_{13} \\ B_{23} \\ B_{33} \end{bmatrix}$$

We can visualize this by laying the vector on its side above the columns of the matrix, multiplying each matrix-column by the vector component, and summing the resulting vectors:

$$\mathbf{B}\mathbf{v} = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & B_{23} \\ B_{31} & B_{32} & B_{33} \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} = \begin{bmatrix} v^x \\ \times \\ B_{11} \\ B_{21} \\ B_{31} \end{bmatrix} + \begin{bmatrix} v^y \\ \times \\ B_{12} \\ B_{22} \\ B_{32} \end{bmatrix} + \begin{bmatrix} v^z \\ \times \\ B_{13} \\ B_{23} \\ B_{33} \end{bmatrix} = v^x \begin{bmatrix} B_{11} \\ B_{21} \\ B_{31} \end{bmatrix} + v^y \begin{bmatrix} B_{12} \\ B_{22} \\ B_{32} \end{bmatrix} + v^z \begin{bmatrix} B_{13} \\ B_{23} \\ B_{33} \end{bmatrix}$$

The columns of **B** are the vectors which are weighted by each of the input vector components,  $v^j$ .

Another important way of conceptualizing a matrix times a vector: the resultant vector is a column of dot products. The  $i^{th}$  element of the result is the dot product of the given vector, **v**, with the  $i^{th}$  row of **B**. Writing **B** as a column of row-vectors:

$$\mathbf{B} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{bmatrix} \rightarrow \mathbf{B}\mathbf{v} = \begin{bmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{bmatrix} \begin{bmatrix} \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_1 \cdot \mathbf{v} \\ \mathbf{r}_2 \cdot \mathbf{v} \\ \mathbf{r}_3 \cdot \mathbf{v} \end{bmatrix}.$$

This view derives from the one above, where we lay the vector on its side above the matrix, but now consider the effect on each row separately: it is exactly that of a dot product.

In linear algebra, even if the matrices are complex, we do not conjugate the left vector in these dot products. If they need conjugation, the application must conjugate them separately from the matrix multiplication, i.e. during the construction of the matrix.

We use this dot product concept later when we consider a change of basis.

**Matrix times a matrix:** Multiplying a matrix  $\mathbf{B}$  times another matrix  $\mathbf{C}$  is *defined* as multiplying each column of  $\mathbf{C}$  by the matrix  $\mathbf{B}$ . Therefore, *by definition*, matrix multiplication distributes to the right across the columns:

$$\text{Let } \mathbf{C} = \begin{bmatrix} \mathbf{x} & \mathbf{y} & \mathbf{z} \end{bmatrix}, \text{ then } \mathbf{BC} = \mathbf{B} \begin{bmatrix} \mathbf{x} & \mathbf{y} & \mathbf{z} \end{bmatrix} \equiv \begin{bmatrix} \mathbf{Bx} & \mathbf{By} & \mathbf{Bz} \end{bmatrix}.$$

[Matrix multiplication also distributes to the left across the rows, but we don't use that as much.]

## Determinants

This section assumes you've seen matrices and determinants, but probably didn't understand the reasons why they work.

The determinant operation on a matrix produces a scalar. It is the only operation (up to a constant factor) which is (1) linear in each row and each column of the matrix; and (2) antisymmetric under exchange of any two rows or any two columns.

The above two rules, linearity and antisymmetry, allow determinants to help solve simultaneous linear equations, as we show later under "Cramer's Rule." In more detail:

1. The determinant is linear in each column-vector (and row-vector). This means that multiplying any column (or row) by a scalar multiplies the determinant by that scalar. E.g.,

$$\det \begin{bmatrix} k\mathbf{a} & \mathbf{b} & \mathbf{c} \end{bmatrix} = k \det \begin{bmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \end{bmatrix}; \text{ and } \det \begin{bmatrix} \mathbf{a} + \mathbf{d} & \mathbf{b} & \mathbf{c} \end{bmatrix} = \det \begin{bmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \end{bmatrix} + \det \begin{bmatrix} \mathbf{d} & \mathbf{b} & \mathbf{c} \end{bmatrix}.$$

2. The determinant is anti-symmetric with respect to any two column-vectors (or row-vectors). This means swapping any two columns (or rows) of the matrix negates its determinant.

The above properties of determinants imply some others:

3. Expansion by minors/cofactors (see below), whose derivation proves the determinant operator is unique (up to a constant factor).
4. The determinant of a matrix with any two columns equal (or proportional) is zero. (From anti-symmetry, swap the two equal columns, the determinant must negate, but its negative now equals itself. Hence, the determinant must be zero.)

$$\det \begin{bmatrix} \mathbf{b} & \mathbf{b} & \mathbf{c} \end{bmatrix} = -\det \begin{bmatrix} \mathbf{b} & \mathbf{b} & \mathbf{c} \end{bmatrix} \Rightarrow \det \begin{bmatrix} \mathbf{b} & \mathbf{b} & \mathbf{c} \end{bmatrix} = 0.$$

5.  $\det|\mathbf{A}| \cdot \det|\mathbf{B}| = \det|\mathbf{AB}|$ . This is crucially important. It also fixes the overall constant factor of the determinant, so that the determinant (with this property) is a completely unique operator.
6. Adding a multiple of any column (row) to any other column (row) does not change the determinant:

$$\det \begin{vmatrix} \mathbf{a} + k\mathbf{b} & \mathbf{b} & \mathbf{c} \end{vmatrix} = \det \begin{vmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \end{vmatrix} + \det \begin{vmatrix} k\mathbf{b} & \mathbf{b} & \mathbf{c} \end{vmatrix} = \det \begin{vmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \end{vmatrix} + k \det \begin{vmatrix} \mathbf{b} & \mathbf{b} & \mathbf{c} \end{vmatrix} = \det \begin{vmatrix} \mathbf{a} & \mathbf{b} & \mathbf{c} \end{vmatrix}$$

7.  $\det|\mathbf{A} + \mathbf{B}| \neq \det|\mathbf{A}| + \det|\mathbf{B}|$ . The determinant operator is *not* distributive over matrix addition.
8.  $\det|k\mathbf{A}| = k^n \det|\mathbf{A}|$ .

The *ij*-th **minor**,  $M_{ij}$ , of an  $n \times n$  matrix ( $\mathbf{A} \equiv A_{ab}$ ) is the product  $A_{ij}$  times the determinant of the  $(n-1) \times (n-1)$  matrix formed by crossing out the *i*-th row and *j*-th column:

$$\begin{array}{c}
 \text{\color{red} } j^{\text{th}} \text{ column} \\
 \begin{vmatrix} A_{11} & \cdot & \cdot & \cdot & A_{1n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \text{\color{red} } i^{\text{th}} \text{ row} & \cdot & A_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ A_{n1} & \cdot & \cdot & \cdot & A_{nn} \end{vmatrix}
 \end{array}
 \rightarrow
 M_{ij} \equiv A_{ij} \det \begin{vmatrix} A'_{11} & \cdot & \cdot & A'_{1,n-1} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ A'_{n-1,1} & \cdot & \cdot & A'_{n-1,n-1} \end{vmatrix}$$

A **cofactor** is just a minor with a plus or minus sign affixed:

$$C_{ij} = (-1)^{i+j} M_{ij} = (-1)^{i+j} A_{ij} \det \left[ [\mathbf{A}] \text{ without } i^{\text{th}} \text{ row and } j^{\text{th}} \text{ column} \right].$$

### Cramer's Rule

It's amazing how many textbooks describe Cramer's rule, and how few explain or derive it. I spent years looking for this, and finally found it in [Arf ch 3]. Cramer's rule is a turnkey method for solving simultaneous linear equations. It is horribly inefficient, and virtually worthless above  $3 \times 3$ , however, it does have important theoretical implications. Cramer's rule solves for  $n$  equations in  $n$  unknowns:

Given  $\mathbf{Ax} = \mathbf{b}$ , where  $\mathbf{A}$  is a coefficient matrix,  
 $\mathbf{x}$  is a vector of unknowns,  $x_i$   
 $\mathbf{b}$  is a vector of constants,  $b_i$

To solve for the  $i^{\text{th}}$  unknown  $x_i$ , we replace the  $i^{\text{th}}$  column of  $\mathbf{A}$  with the constant vector  $\mathbf{b}$ , take the determinant, and divide by the determinant of  $\mathbf{A}$ . Mathematically:

Let  $\mathbf{A} = [\mathbf{a}_1 \mid \mathbf{a}_2 \mid \cdots \mid \mathbf{a}_n]$  where  $\mathbf{a}_i$  is the  $i^{\text{th}}$  column of  $\mathbf{A}$ . We can solve for  $x_i$  as

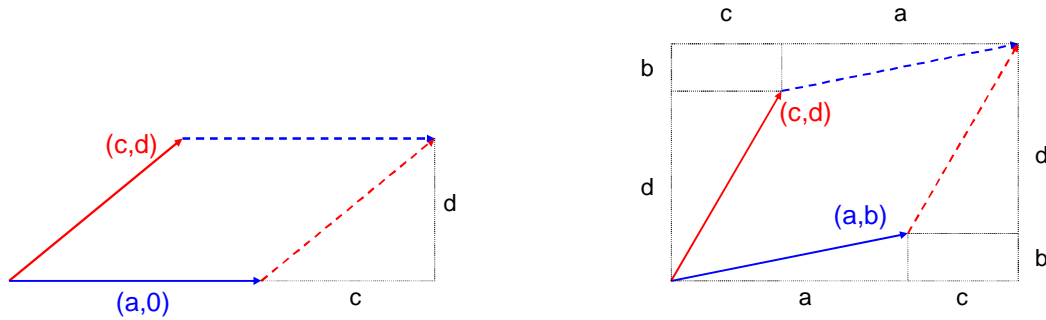
$$x_i = \frac{\det \begin{vmatrix} \mathbf{a}_1 & \cdots & \mathbf{a}_{i-1} & \mathbf{b} & \mathbf{a}_{i+1} & \cdots & \mathbf{a}_n \end{vmatrix}}{\det|\mathbf{A}|}$$

where  $\mathbf{a}_i$  is the  $i^{\text{th}}$  column of  $\mathbf{A}$

This seems pretty bizarre, and one has to ask, why does this work? It's quite simple, if we recall the properties of determinants. Let's solve for  $x_1$ , noting that all other unknowns can be solved analogously. Start by simply multiplying  $x_1$  by  $\det \mathbf{A}$ :

$$\begin{aligned}
 x_1 \det \mathbf{A} &= \det \begin{vmatrix} | & | & | & | \\ x_1 \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & | & | \end{vmatrix} \\
 &= \det \begin{vmatrix} | & | & | & | \\ x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & | & | \end{vmatrix} && \text{adding a multiple of any column to} \\
 & && \text{another doesn't change the determinant} \\
 &= \det \begin{vmatrix} | & | & | & | \\ x_1 \mathbf{a}_1 + x_2 \mathbf{a}_2 + \dots + x_n \mathbf{a}_n & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & | & | \end{vmatrix} && \text{ditto } (n-2) \text{ more times} \\
 &= \det \begin{vmatrix} | & | & | & | \\ \mathbf{Ax} & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & | & | \end{vmatrix} = \det \begin{vmatrix} | & | & | & | \\ \mathbf{b} & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & | & | \end{vmatrix} && \text{rewriting the first column} \\
 \Rightarrow x_1 &= \frac{\det \begin{vmatrix} | & | & | & | \\ \mathbf{b} & \mathbf{a}_2 & \dots & \mathbf{a}_n \\ | & | & | & | \end{vmatrix}}{\det \mathbf{A}}.
 \end{aligned}$$

**Area and Volume as a Determinant**



Determining areas of regions defined by vectors is crucial to geometric physics in many areas. It is the essence of the Jacobian matrix used in variable transformations of multiple integrals. What is the area of the parallelogram defined by two vectors? This is the archetypal area for generalized (oblique, non-normal) coordinates. We will proceed in a series of steps, gradually becoming more general.

First, consider that the first vector is horizontal (above left). The area is simply base  $\times$  height:  $A = ad$ . We can obviously write this as a determinant of the matrix of column vectors, though it is as-yet contrived:



$$A = \det \begin{vmatrix} a & c \\ 0 & d \end{vmatrix} = ad - (0)c = ad.$$

For a general parallelogram (above right), we can take the big rectangle and subtract the smaller rectangles and triangles, by brute force:

$$\begin{aligned} A &= (a+c)(b+d) - 2bc - 2\left(\frac{1}{2}\right)cd - 2\left(\frac{1}{2}\right)ab = \cancel{ab} + ad + cb + \cancel{cd} - 2bc - \cancel{cd} - \cancel{ab} \\ &= ad - bc = \det \begin{vmatrix} a & c \\ b & d \end{vmatrix}. \end{aligned}$$

This is simple enough in 2-D, but is incomprehensibly complicated in higher dimensions. We can achieve the same result more generally, in a way that allows for extension to higher dimensions by induction. Start again with the diagram above left, where the first vector is horizontal. We can rotate that to arrive at any arbitrary pair of vectors, thus removing the horizontal restriction:

Let  $\mathbf{R}$  = the rotation matrix. Then the rotated vectors are  $\mathbf{R} \begin{bmatrix} a \\ 0 \end{bmatrix}$  and  $\mathbf{R} \begin{bmatrix} c \\ d \end{bmatrix}$

$$\det \begin{vmatrix} \mathbf{R} \begin{bmatrix} a \\ 0 \end{bmatrix} & \mathbf{R} \begin{bmatrix} c \\ d \end{bmatrix} \end{vmatrix} = \det \left( \mathbf{R} \begin{bmatrix} a & c \\ 0 & d \end{bmatrix} \right) = (\det \mathbf{R}) \det \begin{vmatrix} a & c \\ 0 & d \end{vmatrix} = \det \begin{vmatrix} a & c \\ 0 & d \end{vmatrix}$$

The final equality is because rotation matrices are orthogonal, with  $\det = 1$ . Thus the determinant of arbitrary vectors defining arbitrary parallelograms equals the determinant of the vectors spanning the parallelogram rotated to have one side horizontal, which equals the area of the parallelogram.

What about the sign? If we reverse the two vectors, the area comes out negative! That's ok, because in differential geometry, 2-D areas are signed: positive if we travel counter-clockwise from the first vector to the 2<sup>nd</sup>, and negative if we travel clockwise. The above areas are positive.

In 3-D, the **signed volume** of the parallelepiped defined by 3 vectors  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $\mathbf{c}$ , is the determinant of the matrix formed by the vectors as columns (positive if  $\mathbf{abc}$  form a right-handed set, negative if  $\mathbf{abc}$  are a left-handed set). We show this with rotation matrices, similar to the 2-D case: First, assume that the parallelogram defined by  $\mathbf{bc}$  lies in the  $x$ - $y$  plane ( $b_z = c_z = 0$ ). Then the volume is simply (area of the base)  $\times$  height:

$$V = (\text{area of base})(\text{height}) = \left( \det \begin{vmatrix} \mathbf{b} & \mathbf{c} \\ \vdots & \vdots \end{vmatrix} \right) (a_z) = \det \begin{vmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & 0 & 0 \end{vmatrix}.$$

where the last equality is from expansion by cofactors along the bottom row. But now, as before, we can rotate such a parallelepiped in 3 dimensions to get any arbitrary parallelepiped. As before, the rotation matrix is orthogonal ( $\det = 1$ ), and does not change the determinant of the matrix of column vectors.

This procedure generalizes to arbitrary dimensions: the signed hyper-volume of a parallelepiped defined by  $n$  vectors in  $n$ -D space is the determinant of the matrix of column vectors. The sign is positive if the 3-D submanifold spanned by each contiguous subset of 3 vectors ( $\mathbf{v}_1\mathbf{v}_2\mathbf{v}_3$ ,  $\mathbf{v}_2\mathbf{v}_3\mathbf{v}_4$ ,  $\mathbf{v}_3\mathbf{v}_4\mathbf{v}_5$ , ...) is right-handed, and negated for each subset of 3 vectors that is left-handed.

### The Jacobian Determinant and Change of Variables

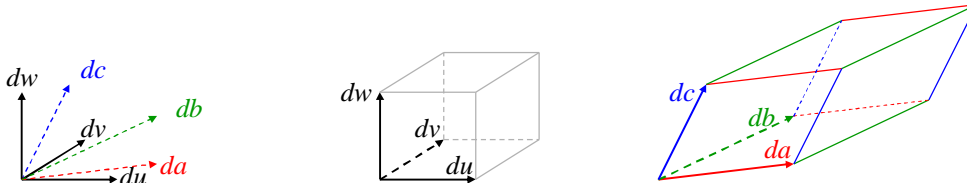
How do we change multiple variables in a multiple integral? Given

$\iiint f(a,b,c) da db dc$  and the change of variables to  $u, v, w$ :

$a = a(u, v, w), \quad b = b(u, v, w), \quad c = c(u, v, w).$  The simplistic

$$\iiint f(a,b,c) da db dc \rightarrow \iiint f[a(u,v,w), b(u,v,w), c(u,v,w)] du dv dw \quad (\text{wrong!})$$

fails, because the “volume”  $du dv dw$  associated with each point of  $f(\cdot)$  is different than the volume  $da db dc$  in the original integral.



Example of new-coordinate volume element ( $du dv dw$ ), and its corresponding old-coordinate volume element ( $da db dc$ ). The new volume element is a rectangular parallelepiped. The old-coordinate parallelepiped has sides straight to first order in the original integration variables.

In the diagram above, we see that the “volume” ( $du dv dw$ ) is smaller than the old-coordinate “volume” ( $da db dc$ ). Note that “volume” is a *relative measure* of volume in coordinate space; it has nothing to do with a “metric” on the space, and “distance” need not even be defined.

There is a concept of relative “volume” in any space, even if there is no definition of “distance.”  
Relative volume is defined as products of coordinate differentials.

The integrand is constant (to first order in the integration variables) over the whole volume element.

Without some correction, the weighting of  $f(\cdot)$  throughout the new-coordinate domain is different than the original integral, and so the integrated sum (i.e., the integral) is different. We correct this by putting in the *original-coordinate* differential volume ( $da db dc$ ) as a function of the *new* differential coordinates,  $du, dv, dw$ . Of course, this function varies throughout the domain, so we can write

$$\iiint f(a,b,c) da db dc \rightarrow \iiint f[a(u,v,w), b(u,v,w), c(u,v,w)] V(u,v,w) du dv dw$$

where  $V(u,v,w)$  takes  $(du dv dw) \rightarrow (da db dc)$

To find  $V(\cdot)$ , consider how the  $a$ - $b$ - $c$  space vector  $da\hat{\mathbf{a}}$  is created from the new  $u$ - $v$ - $w$  space. It has contributions from displacements in all 3 new dimensions,  $u, v,$  and  $w$ :

$$da\hat{\mathbf{a}} = \left( \frac{\partial a}{\partial u} du + \frac{\partial a}{\partial v} dv + \frac{\partial a}{\partial w} dw \right) \hat{\mathbf{a}} \quad \text{Similarly,}$$

$$db\hat{\mathbf{b}} = \left( \frac{\partial b}{\partial u} du + \frac{\partial b}{\partial v} dv + \frac{\partial b}{\partial w} dw \right) \hat{\mathbf{b}}$$

$$dc\hat{\mathbf{c}} = \left( \frac{\partial c}{\partial u} du + \frac{\partial c}{\partial v} dv + \frac{\partial c}{\partial w} dw \right) \hat{\mathbf{c}}$$

The volume defined by the 3 vectors  $du\hat{\mathbf{u}}, dv\hat{\mathbf{v}},$  and  $dw\hat{\mathbf{w}}$  maps to the volume spanned by the corresponding 3 vectors in the original  $a$ - $b$ - $c$  space. The  $a$ - $b$ - $c$  space volume is given by the determinant of the components of the vectors  $d\mathbf{a}, d\mathbf{b},$  and  $d\mathbf{c}$  (written as rows below, to match equations above):

$$volume = \det \begin{vmatrix} \frac{\partial a}{\partial u} du & \frac{\partial a}{\partial v} dv & \frac{\partial a}{\partial w} dw \\ \frac{\partial b}{\partial u} du & \frac{\partial b}{\partial v} dv & \frac{\partial b}{\partial w} dw \\ \frac{\partial c}{\partial u} du & \frac{\partial c}{\partial v} dv & \frac{\partial c}{\partial w} dw \end{vmatrix} = \det \begin{vmatrix} \frac{\partial a}{\partial u} & \frac{\partial a}{\partial v} & \frac{\partial a}{\partial w} \\ \frac{\partial b}{\partial u} & \frac{\partial b}{\partial v} & \frac{\partial b}{\partial w} \\ \frac{\partial c}{\partial u} & \frac{\partial c}{\partial v} & \frac{\partial c}{\partial w} \end{vmatrix} (du dv dw).$$

where the last equality follows from linearity of the determinant. Note that all the partial derivatives are functions of  $u$ ,  $v$ , and  $w$ . Hence,

$$V(u, v, w) = \det \begin{vmatrix} \frac{\partial a}{\partial u} & \frac{\partial a}{\partial v} & \frac{\partial a}{\partial w} \\ \frac{\partial b}{\partial u} & \frac{\partial b}{\partial v} & \frac{\partial b}{\partial w} \\ \frac{\partial c}{\partial u} & \frac{\partial c}{\partial v} & \frac{\partial c}{\partial w} \end{vmatrix} \equiv J(u, v, w) \quad [\text{the Jacobian}], \quad \text{and}$$

$$\iiint f(a, b, c) da db dc \quad \rightarrow \quad \iiint f[a(u, v, w), b(u, v, w), c(u, v, w)] J(u, v, w) du dv dw$$

QED.

### Expansion by Cofactors

Let us construct the determinant operator from its two defining properties: linearity, and antisymmetry. First, we'll define a linear operator, then we'll make it antisymmetric. [This section is optional, though instructive.]

We first construct an operator which is linear in the first column. For the determinant to be linear in the first column, it must be a sum of terms each containing exactly one factor from the first column:

$$\text{Let} \quad \mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix}. \quad \text{Then} \quad \det \mathbf{A} = A_{11}(\dots) + A_{21}(\dots) + \dots + A_{n1}(\dots).$$

To be linear in the first column, the parentheses above must have *no* factors from the first column (else they would be quadratic in some components). Now to also be linear in the 2<sup>nd</sup> column, *all* of the parentheses above must be linear in all the remaining columns. Therefore, to fill in the parentheses we need a linear operator on columns 2... $n$ . But that is the same kind of operator we set out to make: a linear operator on columns 1.. $n$ . Recursion is clearly called for, therefore the parentheses should be filled in with more determinants:

$$\det \mathbf{A} = A_{11}(\det \mathbf{M}_1) + A_{21}(\det \mathbf{M}_2) + \dots + A_{n1}(\det \mathbf{M}_n) \quad (\text{so far}).$$

We now note that the determinant is linear both in the columns, and in the rows. This means that  $\det \mathbf{M}_1$  must not have any factors from the first row *or* the first column of  $\mathbf{A}$ . Hence,  $\mathbf{M}_1$  must be the submatrix of  $\mathbf{A}$  with the first row and first column stricken out.

$$\begin{array}{c}
 \text{1st column} \\
 \text{1st row}
 \end{array}
 \left[ \begin{array}{cccc}
 A_{11} & \cdot & \cdot & A_{1n} \\
 A_{21} & A_{22} & \cdot & A_{2n} \\
 \cdot & \cdot & A_{ij} & \cdot \\
 \cdot & \cdot & \cdot & \cdot \\
 A_{n1} & A_{n2} & \cdot & A_{nn}
 \end{array} \right] \rightarrow \mathbf{M}_1,
 \quad
 \begin{array}{c}
 \text{1st column} \\
 \text{2nd row}
 \end{array}
 \left[ \begin{array}{cccc}
 A_{11} & A_{12} & \cdot & A_{1n} \\
 A_{21} & \cdot & \cdot & A_{2n} \\
 A_{31} & A_{32} & \cdot & A_{3n} \\
 \cdot & \cdot & \cdot & \cdot \\
 A_{n1} & A_{n2} & \cdot & A_{nn}
 \end{array} \right] \rightarrow \mathbf{M}_2, \quad \text{etc.}$$

Similarly,  $\mathbf{M}_2$  must be the submatrix of  $\mathbf{A}$  with the 2<sup>nd</sup> row and first column stricken out. And so on, through  $\mathbf{M}_n$ , which must be the submatrix of  $\mathbf{A}$  with the  $n^{\text{th}}$  row and first column stricken out. We now have an operator that is linear in all the rows and columns of  $\mathbf{A}$ .

So far, this operator is not unique. We could multiply each term in the operator by a constant, and still preserve linearity in all rows and columns:

$$\det \mathbf{A} = k_1 A_{11} (\det \mathbf{M}_1) + k_2 A_{21} (\det \mathbf{M}_2) + \dots + k_n A_{n1} (\det \mathbf{M}_n).$$

We choose these constants to provide the 2<sup>nd</sup> property of determinants: antisymmetry. The determinant is antisymmetric on interchange of any two rows. We start by considering swapping the first two rows: Define  $\mathbf{A}' \equiv (\mathbf{A} \text{ with } A_{1*} \leftrightarrow A_{2*})$ .

$$\begin{array}{c}
 \text{swap}
 \end{array}
 \left[ \begin{array}{cccc}
 A_{11} & A_{12} & \cdot & A_{1n} \\
 A_{21} & \cdot & \cdot & A_{2n} \\
 \cdot & \cdot & A_{ij} & \cdot \\
 \cdot & \cdot & \cdot & \cdot \\
 A_{n1} & \cdot & \cdot & A_{nn}
 \end{array} \right] \rightarrow \mathbf{A}'
 \quad
 \begin{array}{c}
 \text{swapped}
 \end{array}
 \left[ \begin{array}{cccc}
 A_{21} & \cdot & \cdot & A_{2n} \\
 A_{11} & A_{12} & \cdot & A_{1n} \\
 \cdot & \cdot & A_{ij} & \cdot \\
 \cdot & \cdot & \cdot & \cdot \\
 A_{n1} & A_{n2} & \cdot & A_{nn}
 \end{array} \right] \rightarrow \mathbf{M}'_1, \quad \text{etc.}$$

Recall that  $\mathbf{M}_1$  strikes out the first row, and  $\mathbf{M}_2$  strikes out the 2<sup>nd</sup> row, so swapping row 1 with row 2 replaces the first two terms of the determinant:

$$\det \mathbf{A} = k_1 A_{11} (\det \mathbf{M}_1) + k_2 A_{21} (\det \mathbf{M}_2) + \dots \rightarrow \det \mathbf{A}' = k_1 A_{21} (\det \mathbf{M}'_1) + k_2 A_{11} (\det \mathbf{M}'_2) + \dots$$

But  $\mathbf{M}'_1 = \mathbf{M}_2$ , and  $\mathbf{M}'_2 = \mathbf{M}_1$ . So we have:

$$\rightarrow \det \mathbf{A}' = k_1 A_{21} (\det \mathbf{M}_2) + k_2 A_{11} (\det \mathbf{M}_1) + \dots$$

This last form is the same as  $\det \mathbf{A}$ , but with  $k_1$  and  $k_2$  swapped. To make our determinant antisymmetric, we must choose constants  $k_1$  and  $k_2$  such that terms 1 and 2 are antisymmetric on interchange of rows 1 and 2. This simply means that  $k_1 = -k_2$ . So far, the determinant is unique only up to an arbitrary factor, so we choose the simplest such constants:  $k_1 = 1, k_2 = -1$ .

For  $\mathbf{M}_3$  through  $\mathbf{M}_n$ , swapping the first two rows of  $\mathbf{A}$  swaps the first two rows of  $\mathbf{M}'_3$  through  $\mathbf{M}'_n$ :

$$\begin{array}{c}
 \text{swapped}
 \end{array}
 \left[ \begin{array}{cccc}
 A_{21} & A_{22} & \cdot & A_{2n} \\
 A_{11} & A_{12} & \cdot & A_{1n} \\
 A_{31} & \cdot & \cdot & \cdot \\
 A_{41} & A_{42} & \cdot & A_{4n} \\
 A_{n1} & A_{n2} & \cdot & A_{nn}
 \end{array} \right] \rightarrow \mathbf{M}'_3, \quad \text{etc.}$$

Since  $\mathbf{M}_3$  through  $\mathbf{M}_n$  appear inside determinant operators, and such operators are defined to be antisymmetric on interchange of rows, terms 3 through  $n$  also change sign on swapping the first two rows of  $\mathbf{A}$ . Thus, all the terms 1 through  $n$  change sign on swapping rows 1 and 2, and  $\det \mathbf{A} = -\det \mathbf{A}'$ .

We are almost done. We have now a unique determinant operator, with  $k_1 = 1, k_2 = -1$ . We must determine  $k_3$  through  $k_n$ . So consider swapping rows 1 and 3 of  $\mathbf{A}$ , which must also negate our determinant:

$$\begin{matrix} \text{swap} \\ \left[ \begin{array}{cccc} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & \cdot & \cdot & A_{2n} \\ A_{31} & \cdot & \cdot & A_{3n} \\ \cdot & \cdot & \cdot & \cdot \\ A_{n1} & \cdot & \cdot & A_{nn} \end{array} \right] \rightarrow \mathbf{A}'' \end{matrix} \quad \begin{matrix} \text{swapped} \\ \left[ \begin{array}{cccc} A_{31} & \cdot & \cdot & A_{3n} \\ A_{21} & A_{22} & \cdot & A_{2n} \\ A_{11} & A_{12} & \cdot & A_{1n} \\ \cdot & \cdot & \cdot & \cdot \\ A_{n1} & A_{n2} & \cdot & A_{nn} \end{array} \right] \rightarrow \mathbf{M}''_1, \quad \text{etc.} \end{matrix}$$

Again,  $\mathbf{M}''_4$  through  $\mathbf{M}''_n$  have rows 1 & 3 swapped, and thus terms 4 through  $n$  are negated by their determinant operators. Also,  $\mathbf{M}''_2$  (formed by striking out row 2 of  $\mathbf{A}$ ) has its rows 1 & 2 swapped, and is also thus negated.

The terms remaining to be accounted for are  $A_{11}(\det \mathbf{M}_1)$  and  $k_3 A_{31}(\det \mathbf{M}_3)$ . The new  $\mathbf{M}''_1$  is the same as the old  $\mathbf{M}_3$ , but with its first two rows swapped. Similarly, the new  $\mathbf{M}''_3$  is the same as the old  $\mathbf{M}_1$ , but with its first two rows swapped. Hence, both terms 1 and 3 are negated by their determinant operators, so we must choose  $k_3 = 1$  to preserve that negation.

Finally, proceeding in this way, we can consider swapping rows 1 & 4, etc. We find that the odd numbered  $k$ 's are all 1, and the even numbered  $k$ 's are all  $-1$ .

We could also have started from the beginning by linearizing with column 2, and then we find that the  $k$  are opposite to those for column 1: this time for odd numbered rows,  $k_{\text{odd}} = -1$ , and for even numbered rows,  $k_{\text{even}} = +1$ . The  $k$ 's simply alternate sign. This leads to the final form of cofactor expansion about any column  $c$ :

$$\det \mathbf{A} = (-1)^{1+c} A_{1c} (\det \mathbf{M}_1) + (-1)^{2+c} A_{2c} (\det \mathbf{M}_2) + \dots + (-1)^{n+c} A_{nc} (\det \mathbf{M}_n).$$

Note that:

We can perform a cofactor expansion down any column, or across any row, to compute the determinant of a matrix.

We usually choose an expansion order which includes as many zeros as possible, to minimize the computations needed.

### Proof That the Determinant Is Unique

If we compute the determinant of a matrix two ways, from two different cofactor expansions, do we get the same result? Yes. We here prove the determinant is unique by showing that in a cofactor expansion, every possible combination of elements from the rows and columns appears exactly once. This is true no matter what row or column we expand on. Thus all expansions include the same terms, but just written in a different order.

Also, this complete expansion of *all* combinations of elements is a useful property of the cofactor expansion which has many applications beyond determinants. For example, by performing a cofactor expansion without the alternating signs (in other word, an expansion in *minors*), we can fully symmetrize a set of functions (such as boson wave functions).

The proof: let's count the number of terms in a cofactor expansion of a determinant for an  $n \times n$  matrix. We do this by mathematical induction. For the first level of expansion, we choose a row or column, and construct  $n$  terms, where each term includes a cofactor (a sub-determinant of an  $(n-1) \times (n-1)$  matrix). Thus, the number of terms in an  $n \times n$  determinant is  $n$  times the number of terms in an  $(n-1) \times (n-1)$  determinant. Or, turned around,

$$\# \text{terms in } (n+1 \times n+1) = (n+1)(\# \text{terms in } n \times n).$$

There is one term in a 1×1 determinant, 2 terms in a 2×2, 6 terms in a 3×3, and thus  $n!$  terms in an  $n \times n$  determinant. Each term is unique within the expansion: by construction, no term appears twice as we work our way through the cofactor expansion.

Let's compare this to the number of terms *possible* which are linear in every row and column: we have  $n$  choices for the first factor,  $n-1$  choices for the second factor, and so on down to 1 choice for the last factor. That is, there are  $n!$  ways to construct terms linear in all the rows and columns. That is exactly the number of terms in the cofactor expansion, which means every cofactor expansion is a sum of *all possible* terms which are linear in the rows and columns. This proves that the determinant is unique up to a sign.

To prove the sign of the cofactor expansion is also unique, we can consider one specific term in the sum. Consider the term which is the product of the main diagonal elements. This term is always positive, since TBS ??

### Getting Determined

You may have noticed that computing a determinant by cofactor expansion is computationally infeasible for  $n > \sim 15$ . There are  $n!$  terms of  $n$  factors each, requiring  $O(n \cdot n!)$  operations. For  $n = 15$ , this is  $\sim 10^{13}$  operations, which would take about a day on a few GHz computer. For  $n = 20$ , it would take years.

Is there a better way? Fortunately, yes. It can be done in  $O(n^3)$  operations, so one can easily compute the determinant for  $n = 1000$  or more. We do this by using the fact that adding a multiple of any row to another row does not change the determinant (which follows from anti-symmetry and linearity). Performing such row operations, we can convert the matrix to **upper-right-triangular** form, i.e., all the elements of  $\mathbf{A}'$  below the main diagonal are zero.

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix} \quad \rightarrow \quad \mathbf{A}' = \begin{bmatrix} A'_{11} & A'_{12} & \dots & A'_{1,n-1} & A'_{1n} \\ 0 & A'_{22} & \dots & A'_{2,n-1} & A'_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & A'_{n-1,n-1} & A'_{n-1,n} \\ 0 & 0 & \dots & 0 & A'_{nn} \end{bmatrix}$$

By construction,  $\det|\mathbf{A}'| = \det|\mathbf{A}|$ . Using the method of cofactors on  $\mathbf{A}'$ , we expand down the first column of  $\mathbf{A}'$  and first column of every submatrix in the expansion. E.g.,

$$\mathbf{A}' = \begin{bmatrix} A'_{11} & x & x & x \\ 0 & A'_{22} & x & x \\ 0 & 0 & A'_{33} & x \\ 0 & 0 & 0 & A'_{44} \end{bmatrix}$$

Only the first term in each expansion survives, because all the others are zero. Hence,  $\det|\mathbf{A}'|$  is the product of its diagonal elements:

$$\det \mathbf{A} = \det \mathbf{A}' = \prod_{i=1}^n A'_{ii} \quad \text{where } A'_{ii} \text{ are the diagonal elements of } \mathbf{A}'.$$

Let's look at the row operations needed to achieve upper-right-triangular form. We multiply the first row by  $(A_{21} / A_{11})$  and subtract it from the 2<sup>nd</sup> row. This makes the first element of the 2<sup>nd</sup> row zero (below left):

$$\mathbf{A} \rightarrow \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{1n} \\ 0 & B_{22} & B_{23} & B_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ A_{41} & A_{42} & A_{43} & A_{44} \end{bmatrix} \rightarrow \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{1n} \\ 0 & B_{22} & B_{23} & B_{24} \\ 0 & B_{32} & B_{33} & B_{34} \\ 0 & B_{42} & B_{43} & B_{44} \end{bmatrix} \rightarrow \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{1n} \\ 0 & B_{22} & B_{23} & B_{24} \\ 0 & 0 & C_{33} & C_{34} \\ 0 & 0 & C_{43} & C_{44} \end{bmatrix}$$

Perform this operation for rows 3 through  $n$ , and we have made the first column below row 1 all zero (above middle). Similarly, we can zero the 2<sup>nd</sup> column below row 2 by multiplying the (new) 2<sup>nd</sup> row by  $(B_{32} / B_{22})$  and subtracting it from the 3<sup>rd</sup> row. Perform this again on the 4<sup>th</sup> row, and we have the first two columns of the upper-right-triangular form (above right). Iterating for the first  $(n - 1)$  columns, we complete the upper-right-triangular form. The determinant is now the product of the diagonal elements.

About how many operations did that take? There are  $n(n - 1)/2$  row-operations needed, or  $O(n^2)$ . Each row-operation takes from 1 to  $n$  multiplies (average  $n/2$ ), and 1 to  $n$  additions (average  $n/2$ ), summing to  $O(n)$  operations. Total operations is then of order

$$O(n)O(n^2) \sim O(n^3).$$

TBS: Proof that  $\det|\mathbf{AB}| = \det|\mathbf{A}| \det|\mathbf{B}|$

## Advanced Matrices

### Getting to Home Basis

We often wish to change the basis in which we express vectors and matrix operators, e.g. in quantum mechanics. We use a transformation matrix to transform the components of the vectors from the old basis to the new basis. Note that:

We are not transforming the vectors; we are transforming the components of the vector from one basis to another. The vector itself is unchanged.

There are two ways to visualize the transformation. In the first method, we write the decomposition of a vector into components in matrix form. We use the visualization from above that a matrix times a vector is a weighted sum of the columns of the matrix:

$$\mathbf{v} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \mathbf{e}_x & \mathbf{e}_y & \mathbf{e}_z \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} = v^x \mathbf{e}_x + v^y \mathbf{e}_y + v^z \mathbf{e}_z$$

This is a vector equation which is true in any basis. In the  $x$ - $y$ - $z$  basis, it looks like this:

$$\mathbf{v} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} = \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} \quad \text{where} \quad \mathbf{e}_x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{e}_y = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{e}_z = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

If we wish to convert to the  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  basis, we simply write  $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$  in the 1-2-3 basis:

$$\mathbf{v} = \begin{bmatrix} a & d & g \\ b & e & h \\ c & f & i \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} = \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} \quad \text{where (in the 1-2-3 basis):} \quad \mathbf{e}_x = \begin{bmatrix} a \\ b \\ c \end{bmatrix}, \quad \mathbf{e}_y = \begin{bmatrix} d \\ e \\ f \end{bmatrix}, \quad \mathbf{e}_z = \begin{bmatrix} g \\ h \\ i \end{bmatrix}.$$

Thus:

The columns of the transformation matrix are the old basis vectors written in the new basis.  
This is true even for non-ortho-normal bases.

Now let us look at the same transformation matrix, from the viewpoint of its rows. For this, we must restrict ourselves to ortho-normal bases. This is usually not much of a restriction. Recall that the component of a vector  $\mathbf{v}$  in the direction of a basis vector  $\mathbf{e}_i$  is given by:

$$v^i = \mathbf{e}_i \cdot \mathbf{v} \quad \Rightarrow \quad \mathbf{v} = (\mathbf{e}_x \cdot \mathbf{v})\mathbf{e}_x + (\mathbf{e}_y \cdot \mathbf{v})\mathbf{e}_y + (\mathbf{e}_z \cdot \mathbf{v})\mathbf{e}_z.$$

But this is a vector equation, valid in any basis. So  $i$  above could also be 1, 2, or 3 for the new basis:

$$v^1 = \mathbf{e}_1 \cdot \mathbf{v}, \quad v^2 = \mathbf{e}_2 \cdot \mathbf{v}, \quad v^3 = \mathbf{e}_3 \cdot \mathbf{v} \quad \mathbf{v} = (\mathbf{e}_1 \cdot \mathbf{v})\mathbf{e}_1 + (\mathbf{e}_2 \cdot \mathbf{v})\mathbf{e}_2 + (\mathbf{e}_3 \cdot \mathbf{v})\mathbf{e}_3.$$

Recall from the section above on matrix multiplication that multiplying a matrix by a vector is equivalent to making a set of dot products, one from each row, with the vector:

$$\begin{bmatrix} \mathbf{e}_1 \\ \text{-----} \\ \mathbf{e}_2 \\ \text{-----} \\ \mathbf{e}_3 \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{v} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{v} \\ \mathbf{e}_2 \cdot \mathbf{v} \\ \mathbf{e}_3 \cdot \mathbf{v} \end{bmatrix} = \begin{bmatrix} v^1 \\ v^2 \\ v^3 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} (\mathbf{e}_1)_x & (\mathbf{e}_1)_y & (\mathbf{e}_1)_z \\ \text{-----} \\ (\mathbf{e}_2)_x & (\mathbf{e}_2)_y & (\mathbf{e}_2)_z \\ \text{-----} \\ (\mathbf{e}_3)_x & (\mathbf{e}_3)_y & (\mathbf{e}_3)_z \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{v} \\ \mathbf{e}_2 \cdot \mathbf{v} \\ \mathbf{e}_3 \cdot \mathbf{v} \end{bmatrix} = \begin{bmatrix} v^1 \\ v^2 \\ v^3 \end{bmatrix}.$$

Thus:

The rows of the transformation matrix are the new basis vectors written in the old basis.  
This is only true for ortho-normal bases.

There is a beguiling symmetry, and nonsymmetry, in the above two boxed statements about the columns and rows of the transformation matrix.

For complex vectors, we must use the dot product defined with the conjugate of the row basis vector, i.e. the rows of the transformation matrix are the hermitian adjoints of the new basis vectors written in the old basis:

$$\begin{bmatrix} \mathbf{e}_1^\dagger \\ \text{-----} \\ \mathbf{e}_2^\dagger \\ \text{-----} \\ \mathbf{e}_3^\dagger \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \mathbf{v} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{v} \\ \mathbf{e}_2 \cdot \mathbf{v} \\ \mathbf{e}_3 \cdot \mathbf{v} \end{bmatrix} = \begin{bmatrix} v^1 \\ v^2 \\ v^3 \end{bmatrix}.$$

### Diagonalizing a Self-Adjoint Matrix

A special case of basis changing comes up often in quantum mechanics: we wish to change to the basis of eigenvectors of a given operator. In this basis, the basis vectors (which are also eigenvectors) always have the form of a single '1' component, and the rest 0. E.g.,

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad \mathbf{e}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

The matrix operator  $\mathbf{A}$ , in this basis (its own eigenbasis), is diagonal, because:

$$\left. \begin{array}{l} \mathbf{A}\mathbf{e}_1 = \lambda_1\mathbf{e}_1 \\ \mathbf{A}\mathbf{e}_2 = \lambda_2\mathbf{e}_2 \\ \mathbf{A}\mathbf{e}_3 = \lambda_3\mathbf{e}_3 \end{array} \right\} \Rightarrow \quad \mathbf{A} = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \lambda_3 \end{bmatrix}.$$



Finding the unitary (i.e., unit magnitude) transformation from a given basis to the eigenbasis of an operator is called **diagonalizing the matrix**. We saw above that the transformation matrix from one basis to another is just the hermitian adjoint of the new basis vectors written in the old basis. We call this matrix **U**:

$$\begin{bmatrix} \mathbf{e}_1^\dagger \\ \text{-----} \\ \mathbf{e}_2^\dagger \\ \text{-----} \\ \mathbf{e}_3^\dagger \end{bmatrix} \begin{bmatrix} \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{v} \\ \mathbf{e}_2 \cdot \mathbf{v} \\ \mathbf{e}_3 \cdot \mathbf{v} \end{bmatrix} = \begin{bmatrix} v^1 \\ v^2 \\ v^3 \end{bmatrix} \Rightarrow \mathbf{U} = \begin{bmatrix} \mathbf{e}_1^\dagger \\ \text{-----} \\ \mathbf{e}_2^\dagger \\ \text{-----} \\ \mathbf{e}_3^\dagger \end{bmatrix}.$$

**U** transforms vectors, but how do we transform the operator matrix **A** itself? The simplest way to see this is to note that we can perform the operation **A** in any basis by transforming the vector back to the original basis, using **A** in the original basis, and then transforming the result to the new basis:

$$\mathbf{v}_{new} = \mathbf{U}\mathbf{v}_{old} \Rightarrow \mathbf{v}_{old} = \mathbf{U}^{-1}\mathbf{v}_{new}$$

$$\mathbf{A}_{new}\mathbf{v}_{new} = \mathbf{U}(\mathbf{A}_{old}\mathbf{v}_{old}) = \mathbf{U}(\mathbf{A}_{old}\mathbf{U}^{-1}\mathbf{v}_{new}) = (\mathbf{U}\mathbf{A}_{old}\mathbf{U}^{-1})\mathbf{v}_{new} \Rightarrow \mathbf{A}_{new} = (\mathbf{U}\mathbf{A}_{old}\mathbf{U}^{-1})$$

where we used the fact that matrix multiplication is associative. Thus:

The unitary transformation that diagonalizes a (complex) self-adjoint matrix is the matrix of normalized eigen-row-vectors.

We can see this another way by starting with:

$$\mathbf{A}\mathbf{U}^{-1} = \mathbf{A} \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{A}\mathbf{e}_1 & \mathbf{A}\mathbf{e}_2 & \mathbf{A}\mathbf{e}_3 \end{bmatrix} = \begin{bmatrix} \lambda_1\mathbf{e}_1 & \lambda_2\mathbf{e}_2 & \lambda_3\mathbf{e}_3 \end{bmatrix}$$

where  $\mathbf{e}_i$  are the orthonormal eigenvectors  
 $\lambda_i$  are the eigenvalues

Recall the eigenvectors (of self-adjoint matrices) are orthogonal, so we can now pre-multiply by the hermitian conjugate of the eigenvector matrix:

$$\mathbf{U}\mathbf{A}\mathbf{U}^{-1} = \begin{bmatrix} \mathbf{e}_1^\dagger \\ \text{-----} \\ \mathbf{e}_2^\dagger \\ \text{-----} \\ \mathbf{e}_3^\dagger \end{bmatrix} \mathbf{A} \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1^\dagger \\ \text{-----} \\ \mathbf{e}_2^\dagger \\ \text{-----} \\ \mathbf{e}_3^\dagger \end{bmatrix} \begin{bmatrix} \lambda_1\mathbf{e}_1 & \lambda_2\mathbf{e}_2 & \lambda_3\mathbf{e}_3 \end{bmatrix}$$

$$= \begin{bmatrix} \lambda_1(\mathbf{e}_1 \cdot \mathbf{e}_1) & \lambda_2(\cancel{\mathbf{e}_1 \cdot \mathbf{e}_2}) & \cancel{\phantom{\lambda_2(\mathbf{e}_1 \cdot \mathbf{e}_2)}} \\ \lambda_1(\cancel{\mathbf{e}_2 \cdot \mathbf{e}_1}) & \lambda_2(\mathbf{e}_2 \cdot \mathbf{e}_2) & \cancel{\phantom{\lambda_2(\mathbf{e}_2 \cdot \mathbf{e}_2)}} \\ \cancel{\phantom{\lambda_1(\mathbf{e}_2 \cdot \mathbf{e}_1)}} & \cancel{\phantom{\lambda_2(\mathbf{e}_2 \cdot \mathbf{e}_2)}} & \lambda_3(\mathbf{e}_3 \cdot \mathbf{e}_3) \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

where the final equality is because each element of the result is the inner product of two eigenvectors, weighted by an eigenvalue. The only non-zero inner products are between the same eigenvectors (orthogonality), so only diagonal elements are non-zero. Since the eigenvectors are normalized, their inner product is 1, leaving only the weight (i.e., the eigenvalue) as the result.

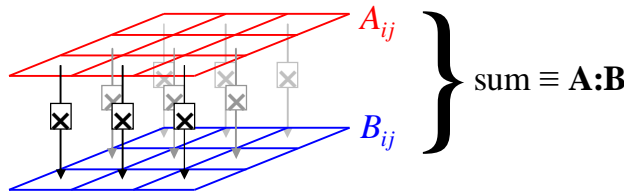
**Warning** Some reference write the diagonalization as  $\mathbf{U}^{-1}\mathbf{A}\mathbf{U}$ , instead of the correct  $\mathbf{U}\mathbf{A}\mathbf{U}^{-1}$ . This is confusing, and inconsistent with vector transformation. Many of these very references

then *change* their notation when they have to transform a vector, because nearly all references agree that vectors transform with  $U$ , and not  $U^{-1}$ .

### Contraction of Matrices

You don't see a dot product of matrices defined very often, but the concept comes up in physics, even if they don't call it a "dot product." We see such products in QM density matrices, and in tensor operations on vectors. We use it below in the "Trace" section for traces of products.

For two matrices of the same size, we define the **contraction of two matrices** as the sum of the products of the corresponding elements (much like the dot product of two vectors). The contraction is a scalar. Picture the contraction as overlaying one matrix on top of the other, multiplying the stacked numbers (elements), and adding all the products:



We use a colon to convey that the summation is over 2 dimensions (rows and columns) of  $A$  and  $B$  (whereas the single-dot dot product of vectors sums over the 1 dimensional list of vector components):

$$A : B \equiv \sum_{i,j=1}^n a_{ij} b_{ij} \quad \text{For example, for } 3 \times 3 \text{ matrices:}$$

$$A : B = \begin{matrix} a_{11}b_{11} & +a_{12}b_{12} & +a_{13}b_{13} \\ +a_{21}b_{21} & +a_{22}b_{22} & +a_{23}b_{23} \\ +a_{31}b_{31} & +a_{32}b_{32} & +a_{33}b_{33} \end{matrix} = a_{11}b_{11} + a_{12}b_{12} + a_{13}b_{13} + a_{21}b_{21} + a_{22}b_{22} + a_{23}b_{23} + a_{31}b_{31} + a_{32}b_{32} + a_{33}b_{33}$$

which is a single number.

If the matrices are complex, we do not conjugate the left matrix (such conjugation is often done in defining the dot product of complex vectors).

### Trace of a Product of Matrices

The trace of a matrix is defined as the sum of the diagonal elements:

$$\text{Tr}(\mathbf{A}) \equiv \sum_{j=1}^n a_{jj} \quad \text{E.g.: } \mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}, \quad \text{Tr}(\mathbf{A}) = a_{11} + a_{22} + a_{33}.$$

The trace of a product of matrices comes up often, e.g. in quantum field theory. We first show that  $\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA})$ :

Let  $\mathbf{C} = \mathbf{AB}$ .  $\Rightarrow \text{Tr}(\mathbf{AB}) = c_{11} + c_{22} + \dots + c_{mm}$

Define  $a_{r*}$  as the  $r^{\text{th}}$  row of  $\mathbf{A}$ , and  $b_{*c}$  as the  $c^{\text{th}}$  column of  $\mathbf{B}$

$$c_{11} = a_{1*} \cdot b_{*1}, \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} (\mathbf{B}^T)_{11} & (\mathbf{B}^T)_{12} & (\mathbf{B}^T)_{13} \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{pmatrix}$$

$$c_{22} = a_{2*} \cdot b_{*2}, \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} \begin{pmatrix} \cdot & \cdot & \cdot \\ (\mathbf{B}^T)_{21} & (\mathbf{B}^T)_{22} & (\mathbf{B}^T)_{23} \\ \cdot & \cdot & \cdot \end{pmatrix}$$

and so on.

The diagonal elements of the product  $\mathbf{C}$  are the sums of the overlays of the rows of  $\mathbf{A}$  on the columns of  $\mathbf{B}$ . But this is the same as the overlays of the rows of  $\mathbf{A}$  on the rows of  $\mathbf{B}^T$ . Then we sum the overlays, i.e., we overlay  $\mathbf{A}$  onto  $\mathbf{B}^T$ , and sum all the products of all the overlaid elements:

$$\text{Tr}(\mathbf{AB}) = \mathbf{A} : \mathbf{B}^T .$$

Now consider  $\text{Tr}(\mathbf{BA}) = \mathbf{B} : \mathbf{A}^T$ . But visually,  $\mathbf{B} : \mathbf{A}^T$  overlays the same pairs of elements as  $\mathbf{A} : \mathbf{B}^T$ , but in the transposed order. When we sum over all the products of the pairs, we get the same sum either way:

$$\text{Tr}(\mathbf{AB}) = \text{Tr}(\mathbf{BA}) \quad \text{because} \quad \mathbf{A} : \mathbf{B}^T = \mathbf{B} : \mathbf{A}^T .$$

This leads to the important cyclic property for the trace of the product of several matrices:

$$\text{Tr}(\mathbf{AB...C}) = \text{Tr}(\mathbf{CAB...}) \quad \text{because} \quad \text{Tr}((\mathbf{AB...})\mathbf{C}) = \text{Tr}(\mathbf{C}(\mathbf{AB...})) .$$

and matrix multiplication is associative. By simple induction, any cyclic rotation of the matrices leaves the trace unchanged.

## Linear Algebra Briefs

**The determinant equals the product of the eigenvalues:**

$$\det \mathbf{A} = \prod_{i=1}^n \lambda_i \quad \text{where} \quad \lambda_i \text{ are the eigenvalues of } \mathbf{A} .$$

This is because the eigenvalues are unchanged through a similarity transformation. If we diagonalize the matrix, the main diagonal consists of the eigenvalues, and the determinant of a diagonal matrix is the product of the diagonal elements (by cofactor expansion).

## 7 Probability, Statistics, and Data Analysis

I think probability and statistics are among the most conceptually difficult topics in mathematical physics. We start with a brief overview of the basics, but overall, we assume you are familiar with simple probabilities, and gaussian distributions.

### Probability and Random Variables

We assume you have a basic idea of probability, and since we seek here understanding over mathematical purity, we give here intuitive definitions. A **random variable**, say  $X$ , is a quantity that you can observe (or measure), multiple times (at least in principle), and is not completely predictable. Each observation (**instance**) of a random variable may give a different value. Random variables may be **discrete** (the roll of a die), or **continuous** (the angle of a game spinner after you spin it). A **uniform** random variable has all its values equally likely. Thus the roll of a (fair) die is a uniform discrete random variable. The angle of a game spinner is a uniform continuous random variable. But in general, the values of a random variable are not necessarily equally likely. For example, a **gaussian** (aka “normal”) random variable is more likely to be near the mean.

Given a large sample of observations of any physical quantity  $X$ , there will be some structure to the values  $X$  assumes. For discrete random variables, each possible value will appear (close to) some fixed fraction of the time in any large sample. The fraction of a large sample that a given value appears is that value’s **probability**. For a 6-sided die, the probability of rolling 1 is  $1/6$ , i.e.  $\Pr(1) = 1/6$ . Because probability is a fraction of a total, it is always between 0 and 1 inclusive:

$$0 \leq \Pr(\text{anything}) \leq 1.$$

[Note that one can imagine systems of chance specifically constructed to *not* provide consistency between samples, at least not on realistic time scales. By definition, then, observations of such a system do *not* constitute a random variable in the sense of our definition.]

Strictly speaking, a **statistic** is a number that summarizes in some way a set of random values. Many people use the word informally, though, to mean the raw data from which we compute true statistics.

### Conditional Probability

Probability, in general, is a combination of physics and knowledge: the physics of the system in question, and what you know about its state.

**Conditional probability** specifically addresses probability when the state of the system is partly known. A **a priori probability** generally implies less knowledge of state (“a priori” means “in the beginning” or “beforehand”). But there is no true, fundamental distinction, because *all* probabilities are in some way dependent on both physics *and* knowledge.

Suppose you have one bag with 2 white and 2 black balls. You draw 2 balls without replacement. What is the chance the 2<sup>nd</sup> ball will be white? A priori, it’s obviously  $1/2$ . However, suppose the first ball is known white. Now  $\Pr(2^{\text{nd}} \text{ ball is white}) = 1/3$ . So we say the conditional probability that the 2<sup>nd</sup> ball will be white, given that the first ball is white, is  $1/3$ . In symbols:

$$\Pr(2^{\text{nd}} \text{ ball white} \mid \text{first ball white}) = 1/3.$$

Another example of how conditional probability of an event can be different than the a priori probability of that event: I have a bag of white and a bag of black balls. I give you a bag at random. What is the chance the 2<sup>nd</sup> ball will be white? A priori, it’s  $1/2$ . After seeing the 1<sup>st</sup> ball is white, now  $\Pr(2^{\text{nd}} \text{ ball is white}) = 1$ . In this case,

$$\Pr(2^{\text{nd}} \text{ ball white} \mid \text{first ball white}) = 1.$$

**Precise Statement of the Question Is Critical**

Many arguments arise about probability because the questions are imprecise, each combatant has a different interpretation of the question, but neither realizes the other is arguing a different issue. Consider this:

You deal 4 cards from a shuffled standard deck of 52 cards. I tell you 3 of them are aces. What is the probability that the 4<sup>th</sup> card is also an ace?

The question is ambiguous, and could reasonably be interpreted two ways, but the two interpretations have quite different answers. It is very important to know exactly how I have discovered 3 of them are aces.

**Case 1:** I look at the 4 cards and say “At least 3 of these cards are aces.” There are 193 ways that 4 cards can hold at least 3 aces, and only 1 of those ways has 4 aces. Therefore, the chance of the 4<sup>th</sup> card being an ace is 1/193.

**Case 2:** I look at only 3 of the 4 cards and say, “These 3 cards are aces.” There are 49 unseen cards, all equally likely to be the 4<sup>th</sup> card. Only one of them is an ace. Therefore, the chance of the 4<sup>th</sup> card being an ace is 1/49.

It may help to show that we can calculate the 1/49 chance from the 193 hands that have at least 3 aces: Of the 192 that have exactly 3 aces, we expect that 1/4 of them = 48 will show aces as their first 3 cards (because the non-ace has probability 1/4 of being last) . Additionally, the one hand of 4 aces will always show aces as its first 3 cards. Hence, of the 193 hands with at least 3 aces, 49 show aces as their first 3 cards, of which exactly 1 will be the 4-ace hand. Hence, its conditional probability, given that the first 3 cards are aces, is 1/49.

**Let’s Make a Deal**

This is an example of a problem that confuses many people (including me), and how to properly analyze it. We hope this example illustrates some general methods of analysis that you can use to navigate more general confusing questions. In particular, the methods used here apply to renormalizing entangled quantum states when a measurement of one value is made.

You’re in the Big Deal on the game show *Let’s Make a Deal*. There are 3 doors. Hidden behind two of them are goats; behind the other is the Big Prize. You choose door #1. Monty Hall, the MC, knows what’s behind each door. He opens door #2, and shows you a goat. Now he asks, do you want to stick with your door choice, or switch to door #3? Should you switch?

Without loss of generality (WLOG), we assume you choose door #1 (and of course, it doesn’t matter which door you choose). We make a chart of mutually exclusive events, and their probabilities:

Bgg	shows door #2	1/6
	shows door #3	1/6
gBg	shows door #3	1/3
ggB	shows door #2	1/3

After you choose, Monty shows you that door #2 is a goat. So from the population of possibilities, we strike out those that are no longer possible (i.e., where he shows door #3, and those where the big prize is #2), and renormalize the remaining probabilities:

Bgg	shows door #2	<del>1/6</del> 1/3
	<del>shows door #3</del>	<del>1/6</del>
gBg	<del>shows door #3</del>	<del>1/3</del>
ggB	shows door #2	<del>1/3</del> 2/3

Another way to think of this: Monty showing you door #2 is equivalent to saying, “The big prize is either the door you picked, or it’s door #3.” Since your chance of having picked right (1/3) is unaffected by

Monty telling you this,  $\text{Pr}(\text{big prize is \#3}) = 2/3$ . Monty uses his knowledge to always pick a door with a goat. That gives you information, which improves your ability to guess right on your second guess.

You can also see it this way: There's a  $1/3$  chance you picked right the first time. Then you'll switch, and lose. But there's a  $2/3$  chance you picked wrong the first time. Then you'll switch, and win. So you win twice as often as you lose, much better odds than  $1/3$  of winning.

Let's take a more extreme example: suppose there are 100 doors, and you pick #1. Now Monty tells you, "The big prize is either the door you picked, or it's door #57." Should you switch? Of course. The chance you guessed right is tiny, but Monty knows for sure.

## How to Lie With Statistics

In 2007, on the front page of newspapers, was a story about a big study of sexual behavior in America. The *headline point* was that on average, heterosexual men have 7 partners in their lives, and women have only 4.

*Innumeracy*, a book about math and statistics, uses this exact claim from a previous study of sexual behavior, and noted that one can easily prove that the *average* number of heterosexual partners of men and women must be *exactly* the same (if there are equal numbers of men and women in the population. The US has equal numbers of men and women to better than 1%).

The only explanation for the survey results is that *many people are lying*. Typically, men lie on the high side, women lie on the low side. The article goes on to quote all kinds of statistics and "facts," oblivious to the fact that these claims are based on lies. So how much can you believe anything the subjects said?

Even more amazing to me is that the "scientists" doing the study seem equally oblivious to the mathematical impossibility of their results. Perhaps some graduate student got a PhD out of this study, too.

**The proof:** every heterosexual encounter involves a man and a woman. If the partners are new to each other, then it counts as a new partner for both the man and the woman. The average number of partners for men is the total number of new partners for all men divided by the number of men in the US. But this is equal to the total number of new partners for all women divided by the number of women in the US. QED.

[An insightful friend noted, "Maybe to some women, some guys aren't worth counting."]

## Choosing Wisely: An Informative Puzzle

Here's a puzzle which illuminates the physical meaning of the  $\binom{n}{k}$  binomial forms. Try it yourself before reading the answer. Really. First, recall that:

$$\binom{n}{k} \equiv n \text{ choose } k \equiv \frac{n!}{k!(n-k)!}.$$

is the number of combinations of  $k$  items taken from  $n$  distinct items; more precisely,  $\binom{n}{k}$  is the number of ways of choosing  $k$  items from  $n$  distinct items, without replacement, where the order of choosing doesn't matter.

**The puzzle:** Show in words, without algebra, that  $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$ .

Some purists may complain that the demonstration below lacks rigor (not true), or that the algebraic demonstration is "shorter." However, though the algebraic proof is straightforward, it is dull and uninformative. Some may like the demonstration here because it uses the physical meaning of the mathematics to reach an iron-clad conclusion.

**The solution:** The LHS is the number of ways of choosing  $k$  items from  $n + 1$  items. Now there are two *distinct* subsets of those ways: those ways that include the  $(n + 1)^{th}$  item, and those that don't. In the first subset, after choosing the  $(n + 1)^{th}$  item, we must choose  $k - 1$  more items from the remaining  $n$ , and there are  $\binom{n}{k-1}$  ways to do this. In the second subset, we must choose all  $k$  items from the first  $n$ , and there are  $\binom{n}{k}$  ways to do this. Since this covers all the possible ways to choose  $k$  items from  $n + 1$  items, it must be that  $\binom{n+1}{k} = \binom{n}{k-1} + \binom{n}{k}$ . QED.

### Multiple Events

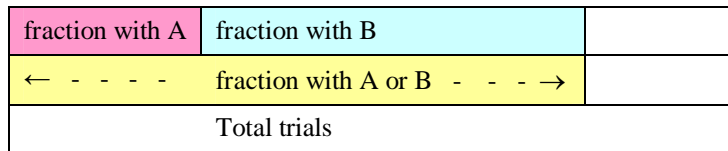
First we summarize the rules for computing the probability of combinations of independent events from their individual probabilities, then we justify them:

- $\Pr(A \text{ and } B) = \Pr(A) \cdot \Pr(B)$ ,                      A and B independent
- $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ ,                      A and B mutually exclusive
- $\Pr(\text{not } A) = 1 - \Pr(A)$
- $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A)\Pr(B)$ ,    always .

For independent events A and B,  $\Pr(A \text{ and } B) = \Pr(A) \cdot \Pr(B)$ . This follows from the definition of probability as a fraction. If A and B are **independent** (have nothing to do with each other), then  $\Pr(A)$  is the fraction of trials with event A. Then of the fraction of those with event A, the fraction that also has B is  $\Pr(B)$ . Therefore, the fraction of the total trials with both A and B is:

$$\Pr(A \text{ and } B) = \Pr(A) \cdot \Pr(B).$$

For mutually exclusive events,  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ . This also follows from the definition of probability as a fraction. The fraction of trials with event A  $\equiv \Pr(A)$ ; fraction with event B  $\equiv \Pr(B)$ . If no trial can contain both A and B, then the fraction with either is simply the sum (figure below).

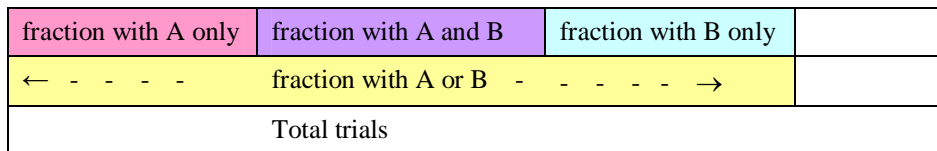


$\Pr(\text{not } A) = 1 - \Pr(A)$ . Since  $\Pr(A)$  is the fraction of trials with event A, and all trials must either have event A or not:

$$\Pr(A) + \Pr(\text{not } A) = 1.$$

Notice that A and (not A) are mutually exclusive events (a trial can't both have A and not have A), so their probabilities add.

By  $\Pr(A \text{ or } B)$  we mean  $\Pr(A \text{ or } B \text{ or both})$ . For independent events, you might think that  $\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$ , but this is not so. A simple example shows that it can't be: suppose  $\Pr(A) = \Pr(B) = 0.7$ . Then  $\Pr(A) + \Pr(B) = 1.4$ , which can't be the probability of anything. The reason for the failure of simple addition of probabilities is that doing so counts the probability of (A and B) twice (figure below):



Note that  $\Pr(A \text{ or } B)$  is equivalent to  $\Pr(A \text{ and maybe } B)$  or  $\Pr(B \text{ and maybe } A)$ . But  $\Pr(A \text{ and maybe } B)$  includes the probability of both  $A$  and  $B$ , as does  $\Pr(B \text{ and maybe } A)$ , hence it is counted twice. So subtracting the probability of  $(A \text{ and } B)$  makes it counted only once:

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A)\Pr(B), \quad A \text{ and } B \text{ independent.}$$

A more complete statement, which breaks down  $(A \text{ or } B)$  into mutually exclusive events is:

$$\Pr(A \text{ or } B) = \Pr(A \text{ and not } B) + \Pr(\text{not } A \text{ and } B) + \Pr(A \text{ and } B)$$

Since the right hand side is now mutually exclusive events, their probabilities add:

$$\begin{aligned} \Pr(A \text{ or } B) &= \Pr(A)[1 - \Pr(B)] + \Pr(B)[1 - \Pr(A)] + \Pr(A)\Pr(B) \\ &= \Pr(A) + \Pr(B) - 2\Pr(A)\Pr(B) + \Pr(A)\Pr(B) \\ &= \Pr(A) + \Pr(B) - \Pr(A)\Pr(B). \end{aligned}$$

TBS: Example of rolling 2 dice.

### Combining Probabilities

Here is a more in-depth view of multiple events, with several examples. This section should be called “Probability Calculus,” but most people associate “calculus” with something hard, and I didn’t want to scare them off. In fact, **calculus** simply means “a method of calculation.”

Probabilities describe binary events: an event either happens, or it doesn’t.  
Therefore, we can use some of the methods of Boolean algebra in probability.

**Boolean algebra** is the mathematics of expressions and variables that can have one of only two values: usually taken to be “true” and “false.” We will use only a few simple, intuitive aspects of Boolean algebra here.

An **event** is something that can either happen, or not (it’s binary!). We define the **probability** of an event as the fraction of time, out of many (possibly hypothetical) trials, that the given event happens. For example, the probability of getting a “heads” from a toss of a fair coin is 0.5, which we might write as  $\Pr(\text{heads}) = 0.5 = 1/2$ . Probability is a fraction of a whole, and so lies in  $[0, 1]$ .

We now consider *two* random events. Two events have one of 3 relationships: independent, mutually exclusive, or conditional (aka conditionally dependent). We will soon see that the first two are special cases of the “conditional” relationship. We now consider each relationship, in turn.

**Independent:** For now, we define independent events as events that have nothing to do with each other, and no effect on each other. For example, consider two events: tossing a heads, and rolling a 1 on a 6-sided die. Then  $\Pr(\text{heads}) = 1/2$ , and  $\Pr(\text{rolling } 1) = 1/6$ . The events are independent, since the coin cannot influence the die, and the die cannot influence the coin. We define one “trial” as two actions: a toss and a roll. Since probabilities are fractions, of all trials,  $1/2$  will have “heads”, and  $1/6$  of those will roll a 1. Therefore,  $1/12$  of all trials will contain both a “heads” and a 1. We see that probabilities of independent events multiply. We write:

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B). \quad (\text{independent events}).$$

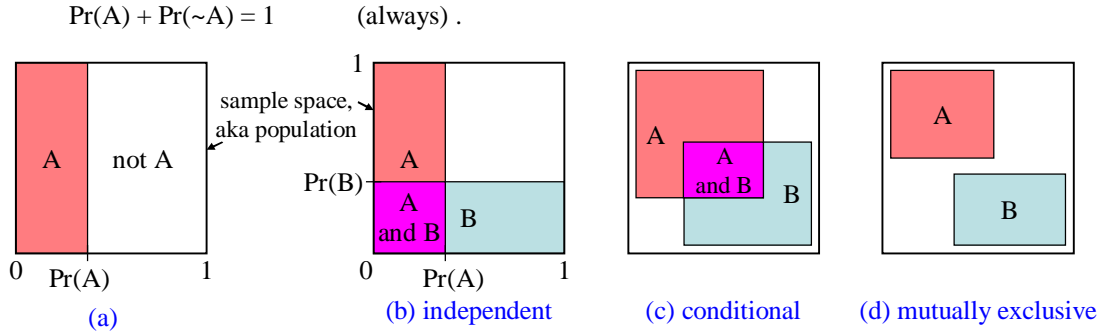
In fact, this is the precise *definition* of independence: if the probability of two events both occurring is the product of the individual probabilities, then the events are **independent**.

[Aside: This definition extends to PDFs: if the joint PDF of two random variables is the product of their individual PDFs, then the random variables are independent.]

Geometric diagrams are very helpful in understanding the probability calculus. We can picture the probabilities of  $A$ ,  $B$ , and  $(A \text{ and } B)$  as areas. The **sample space** or **population** is the set of all possible outcomes of trials. We draw that as a rectangle. Each point in the rectangle represents one possible outcome. Therefore, the probability of an outcome being within a region of the population is proportional to the area of the region.

Figure 7.1 (a): An event  $A$  either happens, or it doesn’t. Therefore:





**Figure 7.1** The (continuous) sample space is the square. Areas are proportional to probabilities. (a) An event either happens, or it doesn't. (b) Events A and B are independent. (c) A and B are dependent. (d) A and B are mutually exclusive.

Figure 7.1 (b):  $\Pr(A)$  is the same whether B occurs or not, shown by the fraction of B covered by A is the same as the fraction of the sample space covered by A. Therefore, *by definition*, A and B are independent.

Figure 7.1 (c): The probability of (A or B (or both)) is the red, blue, and magenta areas. Geometrically then, we see:

$$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B) \quad (\text{always}).$$

This is always true, regardless of any dependence between A and B.

**Conditionally dependent:** From the diagram, when A and B are **conditionally dependent**, we see that the  $\Pr(B)$  depends on whether A happens or not.  $\Pr(B \text{ given that } A \text{ occurred})$  is written as  $\Pr(B | A)$ , and read as “probability of B given A.” From the ratio of the magenta area to the red, we see

$$\Pr(B | A) = \Pr(B \text{ and } A) / \Pr(A) . \quad (\text{always}).$$

**Mutually exclusive:** Two events are **mutually exclusive** when they cannot both happen (diagram above, (d)). Thus,

$$\Pr(A \text{ and } B) = 0, \quad \text{and} \quad \Pr(A \text{ or } B) = \Pr(A) + \Pr(B) \quad (\text{mutually exclusive}).$$

Note that  $\Pr(A \text{ or } B)$  follows the rule from above, which always applies.

We see that independent events are an extreme case of conditional events: independent events satisfy:

$$\Pr(B | A) = \Pr(B) \quad (\text{independent})$$

since the occurrence of A has no effect on B. Also, mutually exclusive events satisfy:

$$\Pr(B | A) = 0 \quad (\text{mutually exclusive})$$

**Summary of Probability Calculus**

Always	
$\Pr(\sim A) = 1 - \Pr(A)$	$\Pr(\text{entire sample space}) = 1$ (diagram above, (a))
$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$	Subtract off any double-count of “A and B” (diagram above, (c))
A & B independent	All from diagram above, (b)
$\Pr(A \text{ and } B) = \Pr(A)\Pr(B)$	Precise def'n of “independent”
$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A)\Pr(B)$	Using the “and” and “or” rules above

$\Pr(B   A) = \Pr(B)$	special case of conditional probability
A & B mutually exclusive	All from diagram above, (d)
$\Pr(A \text{ and } B) = 0$	Def'n of "mutually exclusive"
$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B)$	Nothing to double-count; special case of $\Pr(A \text{ or } B)$ from above
$\Pr(B   A) = \Pr(A   B) = 0$	Can't both happen
Conditional probabilities	All from diagram above, (c)
$\Pr(B   A) = \Pr(B \text{ and } A) / \Pr(A)$	fraction of A that is also B.
$\Pr(B \text{ and } A) = \Pr(B   A)\Pr(A) = \Pr(A   B)\Pr(B)$	Bayes' Rule: Shows relationship between $\Pr(B   A)$ and $\Pr(A   B)$
$\Pr(A \text{ or } B) = \Pr(A) + \Pr(B) - \Pr(A \text{ and } B)$	Same as "Always" rule above

Note that the "and" rules are often simpler than the "or" rules.

### To B, or To Not B?

Sometimes its easier to compute  $\Pr(\sim A)$  than  $\Pr(A)$ . Then we can find  $\Pr(A)$  from  $\Pr(A) = 1 - \Pr(\sim A)$ .

**Example:** What is the probability of rolling 4 or more with two dice?

The population has 36 possibilities. To compute this directly, we use:

$$\underbrace{3}_{\text{ways to roll 4}} + \underbrace{4}_{\text{ways to roll 5}} + \underbrace{5}_{\dots} + 6 + 5 + 4 + \underbrace{3}_{\dots} + \underbrace{2}_{\text{ways to roll 11}} + \underbrace{1}_{\text{ways to roll 12}} = 33 \quad \Rightarrow \quad \Pr(\geq 4) = \frac{33}{36}$$

That's a lot of addition. It's much easier to note that:

$$\Pr(< 4) = \underbrace{1}_{\text{ways to roll 2}} + \underbrace{2}_{\text{ways to roll 3}} = 3 \quad \Rightarrow \quad \Pr(< 4) = \frac{3}{36}, \quad \text{and} \quad \Pr(\geq 4) = 1 - \Pr(< 4) = \frac{33}{36}$$

In particular, the "and" rules are often simpler than the "or" rule. Therefore, when asked for the probability of "this or that", it is sometimes simpler to convert to its complementary "and" statement, compute the "and" probability, and subtract it from 1 to find the "or" probability.

**Example:** From a standard 52-card deck, draw a single card. What is the chance it is a spade or a face-card (or both)? Note that these events are independent.

To compute directly, we use the "or" rule:

$$\Pr(\text{spade}) = 1/4, \quad \Pr(\text{facecard}) = 3/13,$$

$$\Pr(\text{spade or facecard}) = \frac{1}{4} + \frac{3}{13} - \frac{1}{4} \cdot \frac{3}{13} = \frac{13 + 12 - 3}{52} = \frac{22}{52}$$

It may be simpler to compute the probability of drawing neither a spade nor a face-card, and subtracting from 1:

$$\Pr(\sim \text{spade}) = 3/4, \quad \Pr(\sim \text{facecard}) = 10/13,$$

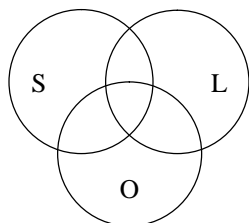
$$\Pr(\text{spade or facecard}) = 1 - \Pr(\sim \text{spade and } \sim \text{facecard}) = 1 - \frac{3}{4} \cdot \frac{10}{13} = 1 - \frac{30}{52} = \frac{22}{52}$$

The benefit of converting to the simpler "and" rule increases with more "or" terms, as shown in the next example.

**Example:** Remove the 12 face cards from a standard 52-card deck, leaving 40 number cards (aces are 1). Draw a single card. What is the chance it is a spade (S), low (L) (4 or less), or odd (O)? Note that these 3 events are independent.

To compute directly, we can count up the number of ways the conditions can be met, and divide by the population of 40 cards. There are 10 spades, 16 low cards, and 20 odd numbers. But we can't just sum those numbers, because we would double (and triple) count many of the cards.

To compute directly, we must extend the “or” the rules to 3 conditions, shown below.



Venn diagram for Spade, Low, and Odd.

Without proof, we state that the direct computation from a 3-term “or” rule is this:

$$\begin{aligned} \Pr(S) &= 1/4, & \Pr(L) &= 4/10, & \Pr(O) &= 1/2 \\ \Pr(S \text{ or } L \text{ or } O) &= \Pr(S) + \Pr(L) + \Pr(O) \\ &\quad - \Pr(S) \Pr(L) - \Pr(S) \Pr(O) - \Pr(L) \Pr(O) + \Pr(S) \Pr(L) \Pr(O) \\ &= \frac{1}{4} + \frac{4}{10} + \frac{1}{2} - \left(\frac{1}{4} \cdot \frac{4}{10}\right) - \left(\frac{1}{4} \cdot \frac{1}{2}\right) - \left(\frac{4}{10} \cdot \frac{1}{2}\right) + \left(\frac{1}{4} \cdot \frac{4}{10} \cdot \frac{1}{2}\right) \\ &= \frac{10 + 16 + 20 - 4 - 5 - 8 + 2}{40} = \frac{31}{40} \end{aligned}$$

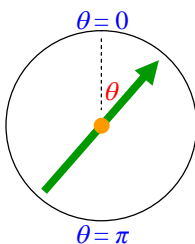
It is far easier to compute the chance that it is neither spade, nor low, nor odd:

$$\begin{aligned} \Pr(\sim S) &= 3/4, & \Pr(\sim L) &= 6/10, & \Pr(\sim O) &= 1/2 \\ \Pr(S \text{ or } L \text{ or } O) &= 1 - \Pr(\sim S \text{ and } \sim L \text{ and } \sim O) = 1 - \Pr(\sim S) \Pr(\sim L) \Pr(\sim O) \\ &= 1 - \frac{3}{4} \cdot \frac{6}{10} \cdot \frac{1}{2} = 1 - \frac{9}{40} = \frac{31}{40}. \end{aligned}$$

You may have noticed that converting “S or L or O” into “ $\sim(\sim S \text{ and } \sim L \text{ and } \sim O)$ ” is an example of De Morgan’s theorem from Boolean algebra.

## Continuous Random Variables and Distributions

Probability is a little more complicated for continuous random variables. A continuous population is a set of random values that can take on values in a continuous interval of real numbers; for example, if I spin a board-game spinner, the little arrow can point in any direction:  $0 \leq \theta < 2\pi$ .



Board game spinner

Furthermore, all angles are equally likely. By inspection, we see that the probability of being in the first quadrant is 1/4, i.e.  $\Pr(0 \leq \theta < \pi/2) = 1/4$ . Similarly, the probability of being in any interval  $d\theta$  is:

$$\Pr(\theta \text{ in any interval } d\theta) = \frac{1}{2\pi} d\theta .$$

If I ask, “what is the chance that it will land at exactly  $\theta = \pi$ ?” the probability goes to zero, because the interval  $d\theta$  goes to zero. In this simple example, the probability of being in any interval  $d\theta$  is the same as being in any other interval of the same size. In general, however, some systems have a probability per unit interval that varies with the value of the random variable (call it  $X$ ) (I wish I had a simple, everyday example of this??). So:

$$\Pr(X \text{ in an infinitesimal interval } dx \text{ around } x) = \text{pdf}(x) dx , \quad \text{where}$$

$$\text{pdf}(x) \equiv \text{the probability distribution function} .$$

pdf(x) has units of 1/x.

By summing mutually exclusive probabilities, the probability of  $X$  in any *finite* interval  $[a, b]$  is:

$$\Pr(a \leq X \leq b) = \int_a^b dx \text{pdf}(x) .$$

Since any random variable  $X$  must have *some* real value, the total probability of  $X$  being between  $-\infty$  and  $+\infty$  must be 1:

$$\Pr(-\infty < X < \infty) = \int_{-\infty}^{\infty} dx \text{pdf}(x) = 1 .$$

The probability distribution function of a random variable tells you everything there is to know about that random variable.

**Population and Samples**

A **population** is a (often infinite) set of all possible values that a random variable may take on, along with their probabilities. A **sample** is a finite set of values of a random variable, where those values come from the population of all possible values. The same value may be repeated in a sample. We often use samples to estimate the characteristics of a much larger population.

A **trial** or **instance** is *one* value of a random variable.

There is enormous confusion over the binomial (and similar) distributions, because each instance of a binomial random variable comes from many **attempts** at an event, where each attempt is labeled either “success” or “failure.” Superficially, an “attempt” looks like a “trial,” and many sources confuse the terms. In the binomial distribution,  $n$  attempts go into making a *single* trial (or instance) of a binomial random variable.

**Population Variance**

The **variance** of a population is a measure of the “spread” of any distribution, i.e. it is some measure of how widely spread out values of a random variable are likely to be [there are other measures of spread,

too]. The variance of a population or sample is among the most important parameters in statistics. Variance is always  $\geq 0$ , and is defined as the average squared-difference between the random values and their average value:

$$\text{var}(X) \equiv \left\langle (X - \bar{X})^2 \right\rangle \quad \text{where } \langle \rangle \text{ is an operator which takes the average } \bar{X} \equiv \langle X \rangle.$$

Note that:

Whenever we write an operator such as  $\text{var}(X)$ , we can think of it as *functional* of the PDF of  $X$  (recall that a functional acts on a function to produce a number).

$$\text{var}(X) \equiv \text{var}[\text{pdf}_X(x)] \equiv \int_{-\infty}^{\infty} (x - \bar{X})^2 \text{pdf}_X(x) dx \equiv \left\langle (X - \bar{X})^2 \right\rangle.$$

The units of variance are the square of the units of  $X$ . From the definition, we see that if I multiply a set of random numbers by a constant  $k$ , then I multiply their variance by  $k^2$ :

$$\text{var}(kX) = k^2 \text{var}(X) \quad \text{where } X \text{ is any set of random numbers.}$$

Any function, including variance, with the above property is **homogeneous-of-order-2** (2<sup>nd</sup> order homogeneous??). We will return later to methods of estimating the variance of a population.

### Population Standard Deviation

The **standard deviation** of a population is another measure of the “spread” of a distribution, defined simply as the square root of the variance. Standard deviation is always  $\geq 0$ , and equals the root-mean-square (RMS) of the deviations from the average:

$$\text{dev}(X) \equiv \sqrt{\text{var}(X)} = \sqrt{\left\langle (X - \bar{X})^2 \right\rangle} \quad \text{where } \langle \rangle \text{ is an operator which takes the average.}$$

As with variance, we can think of  $\text{dev}(X)$  as a functional acting on  $\text{pdf}_X(x)$ :  $\text{dev}[\text{pdf}_X(x)]$ . The units of standard deviation are the units of  $X$ . From the definition, we see that if I multiply a set of random numbers by a constant  $k$ , then I multiply their standard deviation by  $k$ :

$$\text{dev}(kX) = k \text{dev}(X) \quad \text{where } X \text{ is any set of random numbers.}$$

Standard deviation and variance are useful measures, even for non-normal populations.

They have many universal properties, some of which we discuss as we go. There exist bounds on the percentage of *any* population contained with  $\pm c\sigma$ , for any number  $c$ . Even stronger bounds apply for all unimodal populations.

### Normal (aka Gaussian) Distribution

From [mathworld.wolfram.com/NormalDistribution.html](http://mathworld.wolfram.com/NormalDistribution.html) : “While statisticians and mathematicians uniformly use the term ‘normal distribution’ for this distribution, physicists sometimes call it a gaussian distribution and, because of its curved flaring shape, social scientists refer to it as the ‘bell curve.’ ”

A gaussian distribution is one of a 2-parameter family of distributions defined as a population with:

$$\text{pdf}(x) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2} \quad \text{where } \begin{array}{l} \mu \equiv \text{population average} \\ \sigma \equiv \text{population standard deviation} \end{array} \quad [\text{picture??}].$$

$\mu$  and  $\sigma$  are parameters:  $\mu$  can be any real value, and  $\sigma > 0$  and real. This illustrates a common feature of named distributions: they are usually a family of distributions, parameterized by one or more parameters. A gaussian distribution is a 2-parameter distribution:  $\mu$  and  $\sigma$ . As noted below:

Any linear combination of gaussian random variables is another gaussian random variable.

Gaussian distributions are the *only* such distributions [ref??].

### New Random Variables From Old Ones

Given two random variables  $X$  and  $Y$ , we can construct new random variables as functions of  $x$  and  $y$  (trial values of  $X$  and  $Y$ ). One common such new random variable is simply the sum:

$$\text{Define } Z \equiv X + Y \quad \text{which means } \forall \text{ trials } i, \quad z_i \equiv x_i + y_i.$$

We then ask, given  $\text{pdf}_X(x)$  and  $\text{pdf}_Y(y)$  (which is all we can know about  $X$  and  $Y$ ), what is  $\text{pdf}_Z(z)$ ? To answer this, consider a particular value  $x$  of  $X$ ; we see that:

$$\text{Given } x: \quad \text{Pr}(Z \text{ within } dz \text{ of } z) = \text{Pr}(Y \text{ within } dz \text{ of } (z - x)).$$

But  $x$  is a value of a random variable, so the total  $\text{Pr}(Z \text{ within } dz \text{ of } z)$  is the sum (integral) over all  $x$ :

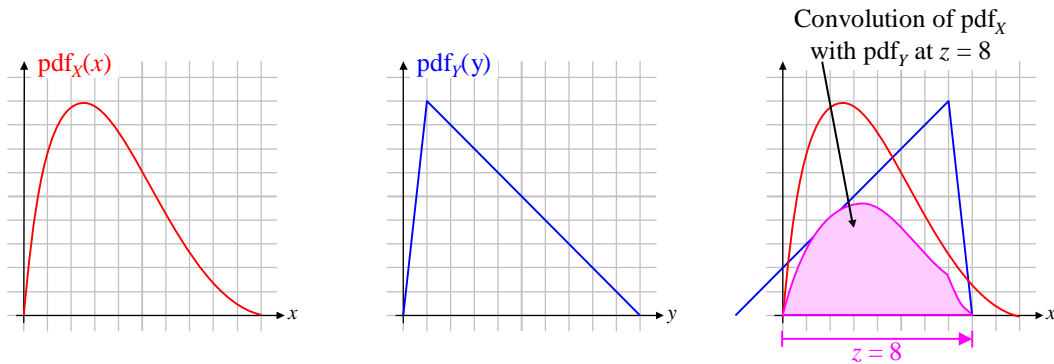
$$\text{Pr}(Z \text{ within } dz \text{ of } z) = \int_{-\infty}^{\infty} dx \text{pdf}_X(x) \text{Pr}(Y \text{ within } dz \text{ of } (z - x)), \quad \text{but}$$

$$\text{Pr}(Y \text{ within } dz \text{ of } (z - x)) = \text{pdf}_Y(z - x) dz, \quad \text{so}$$

$$\text{Pr}(Z \text{ within } dz \text{ of } z) = dz \int_{-\infty}^{\infty} dx \text{pdf}_X(x) \text{pdf}_Y(z - x)$$

$$\Rightarrow \quad \text{pdf}_Z(z) = \int_{-\infty}^{\infty} dx \text{pdf}_X(x) \text{pdf}_Y(z - x).$$

This integral way of combining two functions,  $\text{pdf}_X(x)$  and  $\text{pdf}_Y(y)$  with a parameter  $z$  is called the **convolution** of  $\text{pdf}_X$  and  $\text{pdf}_Y$ , which is a function of a number,  $z$ .



The convolution evaluated at  $z$  is the area under the product  $\text{pdf}_X(x)\text{pdf}_Y(z - x)$ .

From the above, we can easily deduce the  $\text{pdf}_Z(z)$  if  $Z \equiv X - Y = X + (-Y)$ . First, we find  $\text{pdf}_{(-Y)}(y)$ , and then use the convolution rule. Note that:

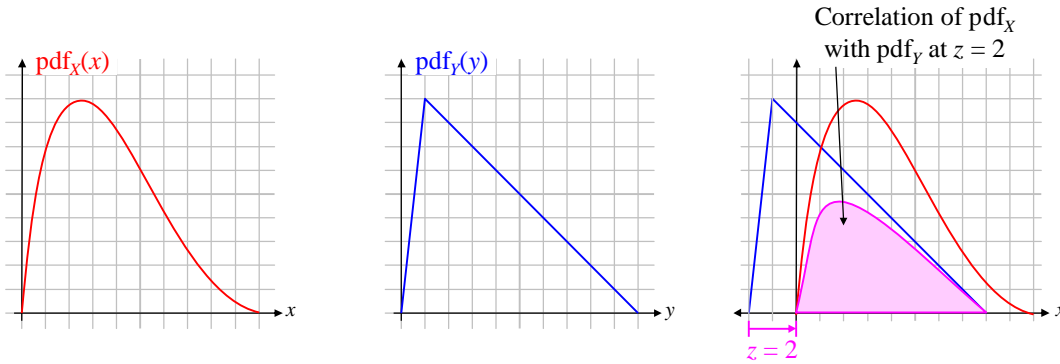
$$\text{pdf}_{(-Y)}(y) = \text{pdf}_Y(-y)$$

$$\Rightarrow \quad \text{pdf}_Z(z) = \int_{-\infty}^{\infty} dx \text{pdf}_X(x) \text{pdf}_{(-Y)}(z - x) = \int_{-\infty}^{\infty} dx \text{pdf}_X(x) \text{pdf}_Y(x - z)$$

Since we are integrating from  $-\infty$  to  $+\infty$ , we can shift  $x$  with no effect:

$$x \rightarrow x + z \quad \Rightarrow \quad \text{pdf}_Z(z) = \int_{-\infty}^{\infty} dx \text{pdf}_X(x + z) \text{pdf}_Y(x),$$

which is the standard form for the **correlation function** of two functions,  $\text{pdf}_X(x)$  and  $\text{pdf}_Y(y)$ .



The correlation function evaluated at  $z$  is the area under the product  $\text{pdf}_x(x+z)\text{pdf}_y(x)$ .

The PDF of the sum of two random variables is the convolution of their PDFs.  
 The PDF of the difference of two random variables is the correlation function of their PDFs.

Note that the convolution of a gaussian distribution with a different gaussian is another gaussian. Therefore, the sum of a gaussian random variable with *any* other gaussian random variable is gaussian.

### Some Distributions Have Infinite Variance, or Infinite Average

In principle, the only requirement on a PDF is that it be normalized:

$$\int_{-\infty}^{\infty} \text{pdf}(x) dx = 1.$$

Such a distribution has well-defined probabilities for all  $x$ . However, even given that, it is possible that the variance is infinite (or properly, undefined). For example, consider:

$$\left. \begin{matrix} \text{pdf}(x) = 2x^{-3} & x \geq 1 \\ = 0 & x < 1 \end{matrix} \right\} \Rightarrow \bar{x} = \int_1^{\infty} x \text{pdf}(x) dx = 2, \quad \text{but} \quad \sigma^2 = \int_1^{\infty} x^2 \text{pdf}(x) dx \rightarrow \infty.$$

The above distribution is normalized, and has finite average, but infinite deviation. The following example is even worse:

$$\left. \begin{matrix} \text{pdf}(x) = x^{-2} & x \geq 1 \\ = 0 & x < 1 \end{matrix} \right\} \Rightarrow \bar{x} = \int_0^{\infty} x \text{pdf}(x) dx \rightarrow \infty, \quad \text{and} \quad \sigma^2 = \int_0^{\infty} x^2 \text{pdf}(x) dx \rightarrow \infty.$$

This distribution is normalized, but has both infinite average and infinite deviation.

Are such distributions physically meaningful? Sometimes. The Lorentzian (aka Breit-Wigner) distribution is common in physics, or at least, a good approximation to physical phenomena. It has infinite average and deviation. It's standard and parameterized forms are:

$$L(x) = \frac{1}{\pi(1+x^2)} \quad L(x; x_0, \gamma) = \frac{1}{\pi\gamma} \cdot \frac{1}{1 + ((x-x_0)/\gamma)^2}$$

where  $x_0 \equiv$  location of peak,  $\gamma \equiv$  half-width at half-maximum

This is approximately the energy distribution of particles created in high-energy collisions. It's CDF is:

$$\text{cdf}_{\text{Lorentzian}}(x) = \frac{1}{\pi} \arctan\left(\frac{x-x_0}{\gamma}\right) + \frac{1}{2}.$$

## Samples and Parameter Estimation

### Why Do We Use Least Squares, and Least Chi-Squared ( $\chi^2$ )?

We frequently use “least sum-squared-residuals” (aka **least squares**) as our definition of “best.” Why sum-squared-residuals? Certainly, one could use other definitions (see least-sum-magnitudes below). However, least squares residuals are most common because they have many useful properties:

- Squared residuals are reasonable: they’re always positive.
- Squared residuals are continuous and differentiable functions of things like fit parameters (magnitude residual is not differentiable). Differentiable means we can analytically minimize it, and for linear fits, the equations are linear.
- We can compute *many* analytic results from least squares, which is not generally true with other residual measures.
- Variance is defined as average of squared deviation (aka “residual”), and variances of uncorrelated random values simply add.
- The central limit theorem causes gaussian distributions to appear frequently in the natural world, and one of its two natural parameters is variance (an average squared-residual).
- For gaussian residuals, least squares parameter estimates are also maximum likelihood.

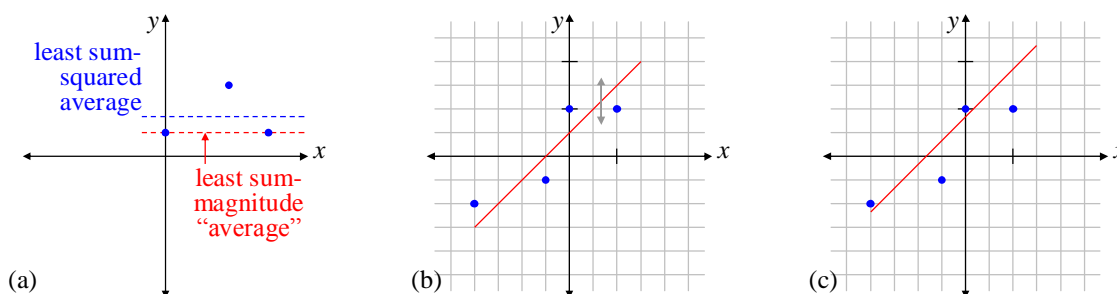
Most other measures of residuals have fewer nice properties.

### Why Not Least-Sum-Magnitudes?

A common question is “Why not magnitude of residuals, instead of squared residuals?” Least-sum-magnitude residuals have at least two serious problems. First, they often yield clearly bad results; and second, least-sum-magnitude-residuals can be highly degenerate: there are often an infinite number of solutions that are “equally” good, and that’s bad.

To illustrate, Figure 7.2a shows the least sum magnitude “average” for 3 points. Sliding the average line up or down increases the magnitude difference for points 1 and 2, and decreases the magnitude difference by the same amount for point 2. Points 1 and 2 totally dominate the result, regardless of how large point 2 is. This is intuitively undesirable for most purposes.

Figure 7.2b and c shows the degeneracy: both lines have equal sum magnitudes, but intuitively fit (b) is vastly better for most purposes.



**Figure 7.2** (a) least-sum-magnitude “average”. (b) Example fit to least-sum-magnitude-residuals. The sum-magnitude is unchanged by moving the “fit line” straight up or down. (c) Alternative “fit” has same sum-magnitude-residuals, but is a much less-likely fit for most residual distributions.

### Other Residual Measures

There *are* some cases where least squares residuals does *not* work well, in particular, if you have outliers in your data. When you square the residual to an outlier, you get a really big number. This squared-residual swamps out all your (real) residuals, thus wreaking havoc with your results. The usual practice is to identify the outliers, remove them, and analyze the remaining data with least-squares.



However, on rare occasions, one might work with a residual measure other than least squared residuals [Myers ??].

When working with data where each measurement has its own uncertainty, we usually replace the least squared residuals criterion with least-chi-squared. We discuss this later when considering data with individual uncertainties.

### Average, Variance, and Standard Deviation

In statistics, an **efficient estimator**  $\equiv$  the *most* efficient estimator [ref??]. There is none better (i.e., none with smaller variance). You can prove mathematically that the average and variance of a sample are the most efficient estimators (least variance) of the population average and variance. It is impossible to do any better, so it's not worth looking for better ways. The most efficient estimators are **least squares** estimators, which means that over many samples, they minimize the sum-squared error from the true value. We discuss least-squares vs. maximum-likelihood estimators later.

Note, however, that given a set of measurements, some of them may not actually measure the population of interest (i.e., they may be noise). If you can identify those bad measurements from a sample, you should remove them before estimating any parameter. Usually, in real experiments, there is always some unremovable corruption of the desired signal, and this contributes to the uncertainty in the measurement.

The **sample average** is defined as:

$$\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i,$$

and is the least variance estimate of the average  $\langle X \rangle$  of any population. It is **unbiased**, which means the average of many sample estimates approaches the true population average:

$$\langle \bar{x} \rangle_{\text{many samples}} = \langle X \rangle \quad \text{where } \langle \bullet \rangle_{\text{over what}} \equiv \text{average, over the given parameter if not obvious.}$$

Note that the definition of unbiased is *not* that the estimator approaches the true value for large samples; it is that the *average* of the estimator approaches the true value over many samples, even small samples.

The sample variance and standard deviation are defined as:

$$s^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{where } \bar{x} \text{ is the sample average, as above: } \bar{x} \equiv \langle x_i \rangle$$

$$s \equiv \sqrt{s^2}$$

The sample variance is an efficient and unbiased estimate of  $\text{var}(X)$ , which means no other estimate of  $\text{var}(X)$  is better. Note that  $s^2$  is unbiased, but  $s$  is *biased*, because the square root of the average is not equal to the average of the square root:

$$\langle s \rangle_{\text{many samples}} \neq \text{dev}(X) \quad \text{because} \quad \langle \sqrt{s^2} \rangle \neq \sqrt{\langle s^2 \rangle}.$$

This exemplifies the importance of properly defining "bias":

$$\langle s \rangle_{\text{many samples}} \neq \text{dev}(X) \quad \text{even though} \quad \lim_{n \rightarrow \infty} s = \text{dev}(X).$$

Sometimes you see variance defined with  $1/n$ , and sometimes with  $1/(n-1)$ . Why? The **population variance** is *defined* as the mean-squared deviation from the population average. For a finite population (such as test scores in a given class), we find the population variance using  $1/N$ , where  $N$  is the number of values in the whole population:

$$\text{var}(X) \equiv \frac{1}{N} \sum_{i=1}^N (X_i - \mu)^2$$

$N$  is the # of values in the entire population  
*where*  $X_i$  is the  $i^{\text{th}}$  value of the population  
 $\mu \equiv$  exact population average .

In contrast, the **sample variance** is the variance of a *sample* taken from a population. The population average  $\mu$  is usually unknown. We can only estimate  $\mu \approx \langle x \rangle$ . Then to make  $s^2$  unbiased (as we show that later), we must use  $1/(n - 1)$ , where  $n$  is the sample size (not population size).

The sample variance is actually a special case of curve fitting, where we fit a constant,  $\langle x \rangle$ , to the population. This is a single parameter, and so removes 1 degree of freedom from our fit errors. Hence, the mean-squared fit error (i.e.,  $s^2$ ) has 1 degree of freedom less than the sample size. (Much more on curve fitting later).

For a *sample* from a population when the average  $\mu$  is exactly known, we use  $n$  as the weighting for an unbiased estimator  $s^2$ :

$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2, \quad \text{which is just the above equation with } X_i \rightarrow x_i, N \rightarrow n.$$

Notice that infinite populations with unknown  $\mu$  can only have samples, and thus always use  $n-1$ . But as  $n \rightarrow \infty$ , it doesn't matter, so we can compute the population variance either way:

$$\text{var}(X) \equiv \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n x_i^2 = \lim_{n \rightarrow \infty} \frac{1}{n-1} \sum_{i=1}^n x_i^2, \text{ because } n-1 \rightarrow n, \text{ when } n \rightarrow \infty.$$

**Central Limit Theorem For Continuous And Discrete Populations**

The central limit theorem is important because it allows us to estimate some properties of a population given only sample of the population, with no a priori information. Given a population, we can take a sample of it, and compute its average. If we take many samples, each will (likely) produce a different average. Hence, the average of a sample is a *new* random variable, created from the original.

The central limit theorem says that for *any* population, as the sample size grows, the sample average approaches a gaussian random variable, with average equal to the population average, and variance equal to the population variance divided by  $n$ .

Mathematically, given a random variable  $X$ , with mean  $\mu$  and variance  $\sigma_X^2$ :

$$\lim_{n \rightarrow \infty} \langle x \rangle \in \text{gaussian} \left( \mu, \frac{\sigma_X^2}{n} \right) \quad \text{where } \langle x \rangle \equiv \text{sample average} .$$

Note that the central limit theorem applies only to multiple samples from a *single* population (though there are some variations that can be applied to multiple populations). [It is possible to construct large sums of *multiple* populations whose averages are *not* gaussian, e.g. in communication theory, inter-symbol interference (ISI). But we will not go further into that.]

**How does the Central Limit Theorem apply to a discrete population?** If a population is discrete, then any sample average is also discrete. But the gaussian distribution is continuous. So how can the sample average approach a gaussian for large sample size  $N$ ? Though the sample average is discrete, the density of allowed values increases with  $N$ . If you simply plot the discrete values as points, those points approach the gaussian curve. For very large  $N$ , the points are so close, they “look” continuous.

TBS: Why binomial (discreet), Poisson (discreet), and chi-squared (continuous) distributions approach gaussian for large  $n$  (or  $v$ ).

### Uncertainty of Average

The sample average  $\langle x \rangle$  gives us an estimate of the population average  $\mu$ . The sample average, when taken as a set of values of many samples, is itself a random variable. The Central Limit Theorem (CLT) says that if we know the population standard deviation  $\sigma$ , the sample average  $\langle x \rangle$  will have standard deviation:

$$u_{\langle x \rangle} = \frac{\sigma}{\sqrt{n}} \quad (\text{proof below}).$$

In statistics,  $u_{\langle x \rangle}$  is called the **standard error of the mean**. In experiments,  $u_{\langle x \rangle}$  is the 1-sigma **uncertainty** in our estimate of the population average  $\mu$ . However, most often, we know neither  $\mu$  nor  $\sigma$ , and must estimate *both* from our sample, using  $\langle x \rangle$  and  $s$ . For “large” samples, we use simply  $\sigma \approx s$ , and then:

$$u_{\langle x \rangle} \approx \frac{s}{\sqrt{n}} \quad \text{for "large" samples, i.e. } n \text{ is "large" .}$$

For small samples, we must still use  $s$  as our estimate of the population deviation, since we have nothing else. But instead of assuming that  $u_{\langle x \rangle}$  is gaussian, we use the *exact* distribution, which is a little wider, called a **T-distribution** [W&M ??], which is complicated to write explicitly. It take an argument  $t$ , similar to the gaussian  $z \equiv (x - \mu)/\sigma$ , which measures its dimensionless distance from the mean:

$$t \equiv \frac{x - \langle x \rangle}{s} \quad \text{where } \langle x \rangle = \text{sample average, } s = \text{sample standard deviation .}$$

We then use  $t$ , and  $t$ -tables, to establish confidence intervals [ref??].

### Uncertainty of Uncertainty: How Big Is Infinity?

Sometimes, we need to know the uncertainty in our estimate of the population variance (or standard deviation). So let’s look more closely at the uncertainty in our estimate  $s^2$  of the population variance  $\sigma^2$ .

The random variable  $\frac{(n-1)s^2}{\sigma^2}$  has chi-squared distribution with  $n - 1$  degrees of freedom [W&M Thm 6.16 p201]. So:

$$s^2 \in \frac{\sigma^2}{n-1} \chi_{n-1}^2 \quad \Rightarrow \quad \text{var}(s^2) = \left( \frac{\sigma^2}{n-1} \right)^2 2(n-1) = \frac{2\sigma^4}{n-1},$$

$$\text{dev}(s^2) = \frac{\sigma^2}{n-1} \sqrt{2(n-1)} = \sqrt{\frac{2}{n-1}} \sigma^2 .$$

However, usually we’re more interested in the *uncertainty* of the standard deviation estimate, rather than its variance. For that, we use the fact that  $s$  is function of  $s^2$ :  $s \equiv (s^2)^{1/2}$ . For moderate or bigger sample sizes, and confidence ranges up to 95% or so, we can use the approximate formula for the deviation of a function of a random variable (see “Functions of Random Variables,” elsewhere):

$$Y = f(X) \quad \Rightarrow \quad \text{dev}(Y) \approx f'(\langle X \rangle) \text{dev}(X) \quad \text{for small dev}(X) .$$

$$s \equiv (s^2)^{1/2} \quad \Rightarrow \quad \text{dev}(s) \approx \frac{1}{2} (\sigma^2)^{-1/2} \text{dev}(s^2) = \frac{1}{2\sigma} \sqrt{\frac{2}{n-1}} \sigma^2 = \frac{1}{\sqrt{2(n-1)}} \sigma \approx \frac{1}{\sqrt{2(n-1)}} s .$$

This allows us to address the rule of thumb: “ $n > 30$ ” is statistical infinity.

This rule is most often used in estimating the standard error of the mean  $u_{\langle x \rangle}$  (see above), given by  $u_{\langle x \rangle} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}}$ . We don’t know the population deviation,  $\sigma$ , so we approximate it with  $s \approx \sigma$ . For small samples, this isn’t so good. Then, as noted above, the uncertainty  $u_{\langle x \rangle}$  needs to include both the true

sampling uncertainty in  $\langle x \rangle$  and the uncertainty in  $s$ . To be confident that our  $\langle x \rangle$  is within our claim, we need to expand our confidence limits, to allow for the chance that  $s$  happens to be low. The Student T-distribution exactly handles this correction to our confidence limits on  $\langle x \rangle$  for all sample sizes.

However, when can we ignore this correction? In other words, how big should  $n$  be for the gaussian (as opposed to  $T$ ) distribution to be a good approximation. The uncertainty in  $s$  is:

$$u_s \equiv \text{dev}(s) = \frac{1}{\sqrt{2(n-1)}} \sigma .$$

This might seem circular, because we still have  $\sigma$  (which we don't know) on the right hand side. However, it's effect is now reduced by the fraction multiplying it. So the uncertainty in  $\sigma$  is also reduced by this factor, and we can neglect it. Thus to first order, we have:

$$u_s \equiv \text{dev}(s) = \sigma \frac{1}{\sqrt{2(n-1)}} \approx \frac{1}{\sqrt{2(n-1)}} s .$$

So long as  $u_s \ll s$ , we can ignore it. In other words:

$$u_s \ll s \Rightarrow \frac{1}{\sqrt{2(n-1)}} \ll 1, \text{ for } u_{\langle x \rangle} \text{ to be approximately gaussian, and } s \approx \sigma .$$

(You may notice that  $u_s$  is correlated with  $s$ : bigger  $s$  implies bigger (estimated)  $u_s$ , so the contribution to  $u_{\langle x \rangle}$  from  $u_s$  does *not* add in quadrature to  $s/\sqrt{n}$ .) When  $n = 30$ :

$$\frac{1}{\sqrt{2(30-1)}} = 0.13 \ll 1 .$$

13% is pretty reasonable for the uncertainty of the uncertainty  $u_{\langle x \rangle}$ , and  $n = 30$  is the generally agreed upon bound for good confidence that  $s \approx \sigma$ .

### Functions of Random Variables

It follows from the definition of probability that the average value of any function of a random variable is:

$$\langle f(X) \rangle = \int_{-\infty}^{\infty} dx f(x) \text{pdf}_X(x) .$$

We can apply this to our definitions of population average and population variance:

$$\bar{X} \equiv \langle X \rangle = \int_{-\infty}^{\infty} dx x \text{pdf}_X(x), \quad \text{and} \quad \text{var}(X) = \int_{-\infty}^{\infty} dx (x - \bar{X})^2 \text{pdf}_X(x) .$$

---

### Statistically Speaking: What Is The Significance of This?

Before we compute any uncertainties, we should understand what they mean. Statistical significance interprets uncertainties. It is one of the most misunderstood, and yet most important, concepts in science. It underlies virtually all experimental and simulation results. Beliefs (correct and incorrect) about statistical significance drive experiment, research, funding, and policy.

Understanding statistical significance is a prerequisite to understanding science.

This cannot be overstated, and yet many (if not most) scientists and engineers receive no formal training in statistics. The following few pages describe statistical significance, surprisingly using almost no math.

## Overview of Statistical Significance

The term “statistically significant” has a precise meaning which is, unfortunately, different than the common meaning of the word “significant.”

Many experiments compare quantitative measures of two populations, e.g. the IQs of ferrets vs. gophers. In any real experiment, the two measures will almost certainly differ. How should we interpret this difference?

We can use statistics to tell us the meaning of the difference. A difference which is not “statistically significant” in some particular experiment may, in fact, be quite important. But we can only determine its importance if we do another experiment with finer resolution, enough to satisfy our subjective judgment of “importance.” For this section, I use the word **importance** to mean a *subjective* assessment of a measured result.

The statement “We could not measure a difference” is very different from “There is no important difference.” Statistical significance is a quantitative comparison of the magnitude of an effect and the resolution of the statistics used to measure it.

This section requires an understanding of probability and uncertainty.

Statistical significance can be tricky, so we start with several high level statements about what statistical significance is and is not. We then give more specific statements and examples.

**Statistical significance** is many things:

Statistical significance is a measure of an experiment’s ability to resolve its own measured result. It is not a measure of the importance of a result.

Statistical significance is closely related to uncertainty.

Statistical significance is a quantitative statement of the probability that a result is real, instead of a measurement error or the random result of sampling that just happened to turn out that way (by chance).

“Statistically significant” means “measurable by this experiment.” “Not statistically significant” means that we cannot fully trust the result *from this experiment alone*; the experiment was too crude to have confidence in its own result.

Statistical significance is a one-way street: if a result is statistically significant, it is (probably) real. However, it may or may not be important. In contrast, if a result is *not* statistically significant, then we don’t know if it’s real or not. However, we will see that even a not significant result can sometimes provide meaningful and useful information.

If the difference between two results in an experiment is not statistically significant, that difference may still be very real and important.

## Details of Statistical Significance

A meaningful measurement must contain two parts: the magnitude of the result, and the confidence limits on it, both of which are quantitative statements. When we say, “the average IQ of ferrets in our experiment is  $102 \pm 5$  points,” we mean that there is a 95% chance that the *actual* average IQ is between 97 and 107. We could also say that our 95% **confidence limits** are 97 to 107. Or, we could say that our 95% **uncertainty** is 5 points. The confidence limits are sometimes called **error bars**, because on a graph, confidence limits are conventionally drawn as little bars above and below the measured values.

Suppose we test gophers and find that their average IQ is  $107 \pm 4$  points. Can we say “on average, gophers have higher IQs than ferrets?” In other words, is the difference we measured significant, or did it happen just by chance? To assess this, we compute the difference, and its uncertainty (recall that uncorrelated uncertainties add in quadrature):

$$\Delta IQ = (107 - 102) \pm \sqrt{4^2 + 5^2} = 5 \pm 6 \quad (\text{gophers} - \text{ferrets})$$

This says that the difference lies within our uncertainty, so we are not 95% confident that gophers have higher IQs. Therefore, we still don't know if either population has higher IQs than the other. Our experiment was not precise enough to measure a difference. This does *not* mean that there is no difference. However, we can say that there is a 95% chance that the difference is between  $-1$  and  $11$  ( $5 \pm 6$ ). A given experiment measuring a difference can produce one of two results of statistical significance: (1) the difference is statistically significant; or (2) it is not. In this case, the difference is not (statistically) significant at the 95% level.

In addition, confidence limits yield one of three results of “importance:” (1) confirm that a difference is important; or (2) not important, or (3) be inconclusive. But the *judgment* of how much is “important” is outside the scope of the experiment. For example, we may know from prior research that a 10 point average IQ difference makes a population a better source for training pilots, enough better to be “important.” Note that this is a subjective statement, and its precise meaning is outside our scope here.

Five of the six combinations of significance and importance are possible, as shown by the following examples.

**Example 1, not significant, and inconclusive importance:** With the given numbers,  $\Delta IQ = 5 \pm 6$ , the “importance” of our result is inconclusive, because we don't know if the average IQ difference is more or less than 10 points.

**Example 2, not significant, but definitely not important:** Suppose that prior research showed (somehow) that a difference needed to be 20 points to be “important.” Then our experiment shows that the difference is not important, because the difference is very unlikely to be as large as 20 points. In this case, *even though the results are not statistically significant, they are very valuable*; they tell us something meaningful and worthwhile, namely, the difference between the average IQs of ferrets and gophers is not important for using them as a source for pilots. The experimental result is valuable, even though not significant, because it establishes an *upper bound* on the difference.

**Example 3, significant, but inconclusive importance:** Suppose again that a difference of 10 points is important, but our measurements are: ferrets average  $100 \pm 3$  points, and gophers average  $107 \pm 2$  points. Then the difference is:

$$\Delta IQ = (107 - 100) \pm \sqrt{2^2 + 3^2} = 7 \pm 4 \quad (\text{gophers} - \text{ferrets})$$

These results are statistically significant: there is better than a 95% chance that the average IQs of ferrets and gophers are different. However, the importance of the result is still inconclusive, because we don't know if the difference is more or less than 10 points.

**Example 4, significant and important:** Suppose again that a difference of 10 points is important, but we measure that ferrets average  $102 \pm 3$  points, and gophers average  $117 \pm 2$  points. Then the difference is:

$$\Delta IQ = (117 - 102) \pm \sqrt{2^2 + 3^2} = 15 \pm 4 \quad (\text{gophers} - \text{ferrets})$$

Now the difference is both statistically significant, and important, because there is a 95% chance that the difference is  $> 10$  points. We are better off choosing gophers to go to pilot school.

**Example 5, significant, but not important:** Suppose our measurements resulted in

$$\Delta IQ = 5 \pm 4$$

Then the difference is significant, but not important, because we are confident that the difference  $< 10$ . This result established an upper bound on the difference. In other words, our experiment was precise enough that if the difference were important (i.e., big enough to matter), then we'd have measured it.

Finally, note that we cannot have a result that is not significant, but important. Suppose our result was:

$$\Delta IQ = 11 \pm 12$$

The difference is unmeasurably small, and possibly zero, so we certainly cannot say the difference is important. In particular, we can't say the difference is greater than anything.

Thus we see that stating “there is a statistically significant difference” is (by itself) not saying much, because the difference could be tiny, and physically unimportant.

We have used here the common confidence limit fraction of 95%, often taken to be  $\sim 2\sigma$ . The next most common fraction is 68%, or  $\sim 1\sigma$ . Another common fraction is 99%, taken to be  $\sim 3\sigma$ . More precise gaussian fractions are 95.45% and 99.73%, but the digits after the decimal point are usually meaningless (i.e., not statistically significant!) Note that we cannot round 99.73% to the nearest integer, because that would be 100%, which is meaningless in this context. Because of the different confidence fractions in use, you should always state your fractions explicitly. You can state your confidence fraction once, at the beginning, or along with your uncertainty, e.g.  $10 \pm 2 (1\sigma)$ .

**Caveat:** We are assuming **random errors**, which are defined as those that average out with larger sample sizes. **Systematic errors** do not average out, and result from biases in our measurements. For example, suppose the IQ test was prepared mostly by gophers, using gopher cultural symbols and metaphors unfamiliar to most ferrets. Then gophers of equal intelligence will score higher IQs because the test is not fair. This bias changes the meaning of all our results, possibly drastically.

Ideally, when stating a difference, one should put a lower bound on it that is physically important, and give the probability (confidence) that the difference is important. E.g. “We are 95% confident the difference is at least 10 points” (assuming that 10 points on this scale matters).

**Examples**

Here are some examples of meaningful and not-so-meaningful statements:

Meaningless Statements (appearing frequently in print)	Meaningful Statements, possibly subjective (not appearing enough)
The difference in IQ between groups A and B is not statistically significant. (Because your experiment was bad, or because the difference is small?)	Our data show there is a 99% likelihood that the IQ difference between groups A and B is less than 1 point.
We measured an average IQ difference of 5 points. (With what confidence?)	Our experiment had insufficient resolution to tell if there was an important difference in IQ.
Group A has a statistically significantly higher IQ than group B. (How much higher? Is it important?)	Our data show there is a 95% likelihood that the IQ difference between groups A and B is greater than 10 points.

**Statistical significance summary:** “Statistical significance” is a quantitative statement about an experiment’s ability to resolve its own result. We use “importance” as a subjective assessment of a measurement that may be guided by other experiments, and/or gut feel. Statistical significance says nothing about whether the measured result is important or not.

**Predictive Power: Another Way to Be Significant, but Not Important**

Suppose that we have measured IQs of millions of ferrets and gophers over decades. Suppose their population IQs are gaussian, and given by (note the use of  $1\sigma$  uncertainties):

$$\text{ferrets: } 101 \pm 20 \qquad \text{gophers: } 103 \pm 20 \qquad (1\sigma).$$

The average difference is small, but because we have millions of measurements, the uncertainty in the average is even smaller, and we have a statistically significant difference between the two groups.

Suppose we have only one slot open in pilot school, but two applicants: a ferret and a gopher. Who should get the slot? We haven’t measured these two individuals, but we might say, “Gophers have ‘significantly’ higher IQs than ferrets, so we’ll accept the gopher.” Is this valid?

To quantitatively assess the validity of this reasoning, let us suppose (simplistically) that pilot students with an IQ of 95 or better are 20% more likely (1.2x) to succeed than those with  $IQ < 95$ . From the given statistics, 61.8% of ferrets have  $IQs > 95$ , vs. 65.5% of gophers. That is, 61.8% of ferrets get the 1.2x boost in likelihood of success, and similarly for the gophers. Then the *relative* probabilities of success are:

ferrets:  $0.382 + 0.618(1.2) = 1.12$

gophers:  $0.345 + 0.655(1.2) = 1.13$ .

Thus a random gopher is 113/112 times (less than 0.7% more) likely to succeed than a random ferret. This is pretty unimportant. In other words, species (between ferrets and gophers) is not a good predictor of success. Species is *so* bad that many, many other facts will be better predictors of success. Height, eyesight, years of schooling, and sports ability are probably all better predictors. The key point is this:

Differences in average between two populations, that are much smaller than the deviations *within* the populations, are poor predictors of individual outcomes.

## Unbiased vs. Maximum-Likelihood Estimators

In experiments, we frequently have to estimate parameters from data. There is a very important difference between “unbiased” and “maximum likelihood” estimates, even though *sometimes* they are the same. Sadly, two of the most popular experimental statistics books confuse these concepts, and their distinction.

[A common error is to try to “derive” unbiased estimates using the principle of “maximum likelihood,” which is impossible since the two concepts are very different. The incorrect argument goes through the exercise of “deriving” the formula for sample variance from the principle of maximum likelihood, and (of course) gets the wrong answer! Hand waving is then applied to wiggle out of the mistake.]

Everything in this section applies to *arbitrary* distributions, not just gaussian. We follow these steps:

1. Terse definitions, which won't be entirely clear at first.
2. Example of estimating the variance of a population (things still fuzzy).
3. Silly example of the need for maximum-likelihood in repeated trials.
4. Real-world physics examples of different situations leading to different choices between unbiased and maximum-likelihood.
5. Closing comments.

**Terse definitions:** In short:

An unbiased statistic is one whose average is exactly right: in the limit of an infinite number of estimates, the average of an unbiased statistic is exactly the population parameter.

Therefore, the average of many samples of an unbiased statistic is likely closer to the right answer than one sample is.

A **maximum likelihood** statistic is one which is most likely to have produced the given the data. Note that if it is biased, then the average of many maximum likelihood estimates does *not* get you closer to right answer. In other words, given a fixed set of data, maximum-likelihood estimates have some merit, but biased ones can't be combined well with other sets of data (perhaps future data, not yet taken). This concept should become more clear below.

Which is better, an unbiased estimate or a maximum-likelihood estimate? It depends on what you goals are.

**Example of population variance:** Given a sample of values from a population, an *unbiased* estimate of the population variance is

$$\sigma^2 \approx \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1} \quad (\text{unbiased estimate}).$$

If we take several samples of the population, compute an unbiased estimate of the variance for each sample, and average those estimates, we'll get a better estimate of the population variance. Usually, unbiased estimators are those that minimize the sum-squared-error from the true value (**principle of least-squares**).



However, suppose we only get *one* shot at estimating the population variance? Suppose Monty Hall says “I’ll give you a zillion dollars if you can estimate the variance (to within some tolerance)”? What estimate should we give him? Since we only get one chance, we don’t care about the average of many estimates being accurate. We want to give Mr. Hall the variance estimate that is *most likely* to be right. One can show that the most likely estimate is given by using  $n$  in the denominator, instead of  $(n - 1)$ :

$$\sigma^2 \approx \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} \quad (\text{maximum-likelihood estimate}).$$

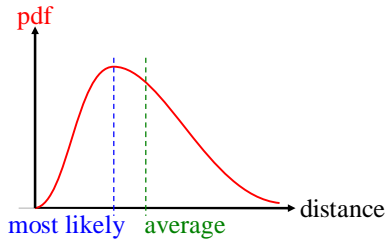
This is the estimate most likely to win the prize. Perhaps more realistically, if you need to choose how long to fire a retro-rocket to land a spacecraft on the moon, do you choose (a) the burn time that, averaged over many spacecraft, reaches the moon, or (b) the burn time that is most likely to land your one-and-only craft on the moon?

In the case of variance, the maximum-likelihood estimate is smaller than the unbiased estimate by a factor of  $(n - 1)/n$ . If we were to make many maximum-likelihood estimates, each one would be small by the same factor. The average would then also be small by that factor. No amount of averaging would ever fix this error. Our average estimate of the population variance would *not* get better with more estimates.

You might conclude that maximum-likelihood estimates are only good for situations where you get a single trial. However, we now show that maximum-likelihood estimates can be useful even when there are many trials of a statistical process.

**Example: Maximum likelihood vs. unbiased:** You are a medieval peasant barely keeping your family fed. Every morning, the benevolent king goes to the castle tower overlooking the public square, and tosses out a gold coin to the crowd. Whoever catches it, keeps it.

Being better educated than most medieval peasants, each day you record how far the coin goes, and generate a PDF (probability distribution function) for the distance from the tower. It looks like Figure 7.3.



**Figure 7.3** Gold coin toss distance PDF.

The most-likely distance is notably different than the average distance. Given this information, where do you stand each day? Answer: At the most-likely distance, because that maximizes your payoff not only for one trial, but across many trials over a long time. The “best” estimator is in the eye of the beholder: as a peasant, you don’t care much for least squares, but you do care about most money.

Note that the previous example of landing a spacecraft is the same as the gold coin question: even if you launch many spacecraft, for each one you would give the burn *most-likely* to land the craft. The average of many failed landings has no value.

**Real physics examples: Example 1:** Suppose you need to generate a beam of ions, all moving at very close to the same speed. You generate your ions in a plasma, with a Maxwellian thermal speed distribution (roughly the same shape as the gold coin toss PDF). Then you send the ions through a velocity selector to pick out only those very close to a single speed. You can tune your velocity selector to pick any speed. Now ions are not cheap, so you want your velocity selector to get the most ions from the speed distribution that it can. That speed is the most-likely speed, not the average speed. So here again, we see that most-likely has a valid use even in repeated trials of random processes.

**Example 2:** Suppose you are tracing out the orbit of the moon around the earth by measuring the distance between the two. Any given day’s measurement has limited ability to trace out an entire orbit, so

you must make many measurements over several years. You have to fit a model of the moon’s orbit to this large set of measurements. You’d like your fit to get better as you collect more data. Therefore, each day you choose to make *unbiased* estimates of the distance, so that on-average, over time, your estimate of the orbit gets better and better. If instead you chose each day’s maximum-likelihood estimator, you’d be off of the average (in the same direction) every day, and no amount of averaging would ever fix that.

**Wrap up:** When you have a symmetric, unimodal distribution for a parameter estimate (symmetric around a single maximum), then the unbiased and maximum-likelihood estimates are identical. This is true, for example, for the average of a gaussian distribution. For asymmetric or multi-modal distributions, the unbiased and maximum-likelihood estimates are different, and have different properties. In general, unbiased estimates are the **most efficient estimators**, which means they have the smallest variance of all possible estimators. Unbiased estimators are also least-squares estimators, which means they minimize the sum-squared error from the true value. This property follows from being unbiased, since the average of a population is the least-squares estimate of all its values.

---

## Correlation and Dependence

To take a sample of a random variable  $X$ , we get a value of  $X_i$  for each sample point  $i, i = 1 \dots n$ . Sometimes when we take a sample, for each sample point we get not one, but two, random variables,  $X_i$  and  $Y_i$ . The two random variables  $X_i$  and  $Y_i$  may or may not be related to each other. We define the joint probability distribution function of  $X$  and  $Y$  such that:

$$\Pr(x < X < x + dx \text{ and } y < Y < y + dy) \equiv \text{pdf}_{XY}(x, y).$$

This is just a 2-dimensional version of a typical pdf. Since  $X$  and  $Y$  are random variables, we could look at either of them and find its individual pdf:  $\text{pdf}_X(x)$ , and  $\text{pdf}_Y(y)$ . If  $X$  and  $Y$  have nothing to do with each other (i.e.,  $X$  and  $Y$  are **independent**), then a fundamental axiom of probability says that the probability density of finding  $x < X < x + dx$  and  $y < Y < y + dy$  is the product of the two pdfs:

$$X \text{ and } Y \text{ are independent} \Rightarrow \text{pdf}_{XY}(x, y) = \text{pdf}_X(x)\text{pdf}_Y(y)$$

The above equation is the *definition* of statistical independence:

Two random variables are independent if and only if their joint distribution function is the product of the individual distribution functions.

A very different concept is “correlation.” Correlation is a measure of how linearly related two random variables are. We discuss correlation in more detail later, but it turns out that we can define correlation mathematically by the correlation coefficient:

$$\rho \propto \langle (X - \bar{X})(Y - \bar{Y}) \rangle \equiv \text{cov}(X, Y).$$

If  $\rho = 0$ , then  $X$  and  $Y$  are uncorrelated. If  $\rho \neq 0$ , then  $X$  and  $Y$  are **correlated**. For a discrete random variable,

$$\rho \propto \sum_{i=1}^{\text{population}} (x_i - \bar{x})(y_i - \bar{y}).$$

Note that :

$$\rho = 0 \Leftrightarrow \text{cov}(X, Y) = 0.$$

Two random variables are uncorrelated if and only if their covariance, defined above, is zero.

Being independent is a stronger statement than uncorrelated. Random variables which are independent are necessarily uncorrelated (proof below). But variables which are uncorrelated can be highly dependent. For example, suppose we have a random variable  $X$ , which is uniformly distributed over  $[-1, 1]$ . Now define a new random variable  $Y$  such that  $Y = X^2$ . Clearly,  $Y$  is dependent on  $X$ , but  $Y$  is uncorrelated with

$X$ .  $Y$  and  $X$  are dependent because given either, we know a lot about the other. They are uncorrelated because for every  $Y$  value, there is one positive and one negative value of  $X$ . So for every value of  $(X - \bar{X})(Y - \bar{Y})$ , there is its negative, as well. The average is therefore 0; hence,  $\text{cov}(X, Y) = 0$ .

A crucial point is:

Variances add for uncorrelated variables, even if they are dependent.

This is easy to show. Given that  $X$  and  $Y$  are uncorrelated,

$$\begin{aligned} \text{var}(X + Y) &= \left\langle [X + Y - (\bar{X} + \bar{Y})]^2 \right\rangle = \left\langle [(X - \bar{X}) + (Y - \bar{Y})]^2 \right\rangle \\ &= \left\langle (X - \bar{X})^2 + 2(X - \bar{X})(Y - \bar{Y}) + (Y - \bar{Y})^2 \right\rangle \\ &= \left\langle (X - \bar{X})^2 \right\rangle + 2\left\langle (X - \bar{X})(Y - \bar{Y}) \right\rangle + \left\langle (Y - \bar{Y})^2 \right\rangle \\ &= \text{var}(X) + \text{var}(Y). \end{aligned}$$

All we needed to prove that variances add is that  $\text{cov}(X, Y) = 0$ .

### Independent Random Variables are Uncorrelated

It is extremely useful to know that independent random variables are necessarily uncorrelated. We prove this now, in part to introduce some methods of statistical analysis, and to emphasize the distinction between “uncorrelated” and “independent.” Understanding analysis methods enables you to analyze a new system reliably, so learning these methods is important for research.

Two random variables are **independent** if they have no relationship at all. Mathematically, the definition of **statistical independence** of two random variables is that the joint density is simply the product of the individual densities:

$$\text{pdf}_{x,y}(x, y) = \text{pdf}_x(x) \text{pdf}_y(y) \quad \text{statistical independence .}$$

The definition of **uncorrelated** is that the covariance, or equivalently the correlation coefficient, is zero:

$$\text{cov}(x, y) = \left\langle (x - \mu_x)(y - \mu_y) \right\rangle = 0 \quad \text{uncorrelated random variables .} \tag{7.1}$$

These definitions are all we need to prove that independent random variables are uncorrelated. First, we prove a slightly simpler claim: independent *zero-mean* random variables are uncorrelated:

$$\text{Given: } \mu_x \equiv \int_{-\infty}^{\infty} dx \text{pdf}_x(x) = 0, \quad \mu_y \equiv \int_{-\infty}^{\infty} dy \text{pdf}_y(y) = 0,$$

then the integral factors into  $x$  and  $y$  integrals, because the joint density of independent random variables factors:

$$\text{cov}(x, y) = \langle xy \rangle = \iint_{-\infty}^{\infty} dx dy \text{pdf}_{x,y}(x, y) xy = \left( \int_{-\infty}^{\infty} dx \text{pdf}_x(x) \right) \left( \int_{-\infty}^{\infty} dy \text{pdf}_y(y) \right) = 0.$$

For non-zero-mean random variables,  $(x - \mu_x)$  is a zero-mean random value, as is  $(y - \mu_y)$ . But these are the quantities that appear in the definition of covariance (7.1). Therefore, the covariance of *any* two independent random variables is zero.

Note well:

Independent random variables are necessarily uncorrelated, but the converse is *not* true:  
uncorrelated random variables may still be dependent.

For example, if  $X \in \text{uniform}(-1,1)$ , and  $Y \equiv X^2$ , then  $X$  and  $Y$  are uncorrelated, but highly dependent.

## Statistical Analysis Algebra

Statistical analysis relies on a number of basic properties of combining random variables (RVs), which define an algebra of statistics. This algebra of RV interaction relates to distributions, averages, variances, and other properties. Within this algebra, there is much confusion about which results apply universally, and which apply only conditionally: e.g., gaussian distributions, independent RVs, uncorrelated RVs, etc. We explicitly address all conditions here. We will use all of these methods later, especially when we derive the lesser-known results for uncertainty weighted data.

### The Average of a Sum: Easy?

We all know that  $\langle x + y \rangle = \langle x \rangle + \langle y \rangle$ . But is this true even if  $x$  and  $y$  are *dependent* random variables (RVs)? Let's see. We can find  $\langle x + y \rangle$  for dependent variables by integrating over the joint density:

$$\begin{aligned} \langle x + y \rangle &\equiv \iint_{-\infty}^{\infty} dx dy \text{pdf}_{x,y}(x,y) (x + y) = \iint_{-\infty}^{\infty} dx dy \text{pdf}_{x,y}(x,y) x + \iint_{-\infty}^{\infty} dx dy \text{pdf}_{x,y}(x,y) y \\ &= \langle x \rangle + \langle y \rangle. \end{aligned}$$

Therefore, the result is easy, and essential for all further analyses:

The average of a sum equals the sum of averages, even for RVs of *arbitrary* dependence.

### The Average of a Product

Life sure would be great if the average of a product were the product of the averages ... but it's not, in general. Although, sometimes it is. As scientists, we need to know the difference. Given  $x$  and  $y$  are random variables (RVs), what is  $\langle xy \rangle$ ?

In statistical analysis, it is often surprisingly useful to break up a random variable into its "varying" part plus its average; therefore, we define:

$$x \equiv \delta x + \mu_x, \quad y \equiv \delta y + \mu_y \quad \Rightarrow \quad \langle \delta x \rangle = \langle \delta y \rangle = 0.$$

Note that  $\mu_x$  and  $\mu_y$  are constants. Then we can evaluate:

$$\begin{aligned} \langle xy \rangle &= \langle (\delta x + \mu_x)(\delta y + \mu_y) \rangle = \langle \delta x \delta y \rangle + \mu_y \langle \delta x \rangle + \mu_x \langle \delta y \rangle + \mu_x \mu_y \\ &= \mu_x \mu_y + \langle (x - \mu_x)(y - \mu_y) \rangle \equiv \mu_x \mu_y + \text{cov}(x, y). \end{aligned}$$

The average of the product is the product of the averages *plus* the covariance.

Only if  $x$  and  $y$  are uncorrelated, which is implied if they are independent (see earlier), then the average of the product is the product of the averages.

This rule provides a simple corollary: the average of an RV squared:

$$\langle x^2 \rangle = \mu_x^2 + \text{cov}(x, x) = \mu_x^2 + \sigma_x^2. \tag{7.2}$$

### Variance of a Sum

We frequently need the variance of a sum of possible *dependent* RVs. We derive it here for RVs  $x, y$ :

$$\begin{aligned} \text{var}(x + y) &= \langle (x + y - \mu_x - \mu_y)^2 \rangle = \langle [(x - \mu_x) + (y - \mu_y)]^2 \rangle \\ &= \langle (x - \mu_x)^2 \rangle + \langle (y - \mu_y)^2 \rangle + 2 \langle (x - \mu_x)(y - \mu_y) \rangle = \text{var}(x) + \text{var}(y) + 2 \text{cov}(x, y). \end{aligned}$$

### Covariance Revisited

The covariance comes up so frequently in statistical analysis that it merits an understanding of its properties as part of the statistical algebra. Covariance appears directly in the formulas for the variance of a sum, and the average of a product, of RVs. (You might remember this by considering the units. For a sum  $x + y$ :  $[x] = [y]$  and  $[\text{var}(x + y)] = [x^2] = [y^2] = [\text{cov}(x, y)]$ . For a product  $xy$ :  $[xy] = [\text{cov}(x, y)]$ .) Conceptually, the covariance of two RVs,  $a$  and  $b$ , measures how much  $a$  and  $b$  vary together linearly from their respective averages. If positive, it means  $a$  and  $b$  tend to go up together; if negative, it means  $a$  tends to go up when  $b$  goes down, and vice-versa. Covariance is defined as a population average:

$$\text{cov}(a, b) \equiv \langle (a - \mu_a)(b - \mu_b) \rangle .$$

From the definition, we see that  $\text{cov}()$  is a bilinear, commutative operator:

Given:  $a, b, c, d$  are random variables;  $k \equiv \text{constant}$ :

$$\text{cov}(a, b) = \text{cov}(b, a)$$

$$\text{cov}(ka, b) = \text{cov}(a, kb) = k \text{cov}(a, b)$$

$$\text{cov}(a + c, b) = \text{cov}(a, b) + \text{cov}(c, b), \quad \text{cov}(a, b + d) = \text{cov}(a, b) + \text{cov}(a, d) .$$

Occasionally, when expanding a covariance, there may be constants in the arguments. We can consider a constant as a random variable which always equals its average, so:

$$\text{cov}(a, k) = 0$$

$$\text{cov}(a + k, b) = \text{cov}(a, b + k) = \text{cov}(a, b) .$$

From the definition, we find that the covariance of an RV with itself is the RV's variance:

$$\text{cov}(a, a) = \text{var}(a) .$$

### Capabilities and Limits of the Sample Variance

The following developments yield important results, and illustrate some methods of statistical algebra that are worth understanding. We wish to determine an unbiased estimator for the population variance,  $\sigma^2$ , from a sample (set) of  $n$  independent values  $\{y_i\}$ , in two cases: (1) we already know the population average  $\mu$ ; and (2) we don't know the population average. The first case is easier. We proceed in detail, because we need this foundation of process to be rock solid, since so much is built upon it.

**$\sigma^2$  from sample and known  $\mu$ :** We must start with the *definition* of population variance as an average over the population:

$$\sigma^2 \equiv \langle (y - \mu)^2 \rangle \quad \text{where} \quad \mu \equiv \langle y \rangle \equiv \text{average over population of } y \equiv \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{i=1}^N y_i . \quad (7.3)$$

A simple guess for the estimator of  $\sigma^2$ , motivated by the definition, might be:

$$g^2 \equiv \frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2 \quad (\text{a guess}) .$$

We now analyze our guess over many samples of size  $n$ , to see how it performs. By definition, to be unbiased, the average of  $g^2$  over an ensemble of samples of size  $n$  must equal  $\sigma^2$ :

$$\text{unbiased:} \quad \langle g^2 \rangle_{\text{ensemble}} = \sigma^2 \equiv \langle (y - \mu)^2 \rangle_{\text{population}} .$$

Mathematically, we find an ensemble average by letting the number of ensembles go to infinity, and the definition of population average is given by letting the number of individual values go to infinity. Let  $M$  be the number of ensembles. Then:

$$\langle g^2 \rangle_{ensemble} \equiv \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M g_m^2 = \lim_{M \rightarrow \infty} \frac{1}{M} \sum_{m=1}^M \frac{1}{n} \sum_{i=1}^n (y_i - \mu)^2.$$

Since all the  $y_i$  above are distinct, we can combine the summations. Effectively, we have converted the ensemble average on the RHS to a population average, whose properties we know:

$$\langle g^2 \rangle_{ensemble} = \lim_{M \rightarrow \infty} \frac{1}{Mn} \sum_{i=1}^{Mn} (y_i - \mu)^2 \equiv \langle (y - \mu)^2 \rangle_{population} \equiv \sigma^2.$$

We have proved that our guess is an unbiased estimator of the population variance,  $\sigma^2$ .

(In fact, since we already know that the sample average is an unbiased estimate of the population average, and the variance  $\sigma^2$  is defined as a population average, then we can conclude immediately that the *sample* average of  $\langle (y_i - \mu)^2 \rangle$  in an unbiased estimate of the *population* average  $\langle (y_i - \mu)^2 \rangle \equiv \sigma^2$ . Again, we took the long route above to illustrate important methods that we will use again.)

Note that the denominator is  $n$ , and not  $n - 1$ ,  
because we started with separate knowledge of the population average  $\mu$ .

For example, when figuring the standard deviation of grades in a class, one uses  $n$  in the denominator, since the class average is known exactly.

**$\sigma^2$  from sample alone:** A harder case is estimating  $\sigma^2$  when  $\mu$  is not known. As before, we must start with a guess at an estimator, and then analyze our guess to see how it performs. A simple guess, motivated by the definition, might be:

$$s^2 \propto \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{a guess}) \quad \text{where} \quad \bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i.$$

By definition, to be unbiased, the average of  $s^2$  over an ensemble of samples of size  $n$  must equal  $\sigma^2$ . We now consider the sum in  $s^2$ . We first show a failed attempt, and then how to avoid it. If we try to analyze the sum directly, we get :

$$\left\langle \sum_{i=1}^n (y_i - \bar{y})^2 \right\rangle = \sum_{i=1}^n \langle y_i^2 - 2\bar{y}y_i + \bar{y}^2 \rangle = \sum_{i=1}^n \langle y_i^2 \rangle - 2 \sum_{i=1}^n \langle \bar{y}y_i \rangle + n \langle \bar{y}^2 \rangle.$$

In the equation above, angle brackets mean ensemble average. By tradition, we don't explicitly label our angle brackets to say what we are averaging over, and we make you figure it out. Even better, as we saw earlier, sometimes the angle brackets mean ensemble average, and sometimes they mean population average. (This is a crucial difference in definition, and a common source of confusion in statistical analysis: just what are we averaging over, anyway?) However, on the RHS, the first ensemble average is the same as the population average. However, further analysis of the ensemble averages at this point is messy (more on this later).

To avoid the mess, we note that definition (7.4) requires us to somehow introduce the population average into the analysis, even though it is unknown. By trial and error, we find it is easier to start with the population average, and write it in terms of  $\bar{y}$  :

$$\sum_{i=1}^n (y_i - \mu)^2 = \sum_{i=1}^n ((y_i - \bar{y}) + (\bar{y} - \mu))^2 = \sum_{i=1}^n (y_i - \bar{y})^2 + 2(\bar{y} - \mu) \underbrace{\sum_{i=1}^n (y_i - \bar{y})}_0 + \sum_{i=1}^n (\bar{y} - \mu)^2.$$

$(\bar{y} - \mu)$  does not depend on  $i$ , so it comes out of the summation. The second term is identically zero, because:

$$\sum_{i=1}^n (y_i - \bar{y}) = \sum_{i=1}^n y_i - n\bar{y} = n\bar{y} - n\bar{y} = 0.$$

Now we can take the ensemble average of the remains of the sum-of-squares equation:

$$\begin{aligned} \left\langle \sum_{i=1}^n (y_i - \mu)^2 \right\rangle &= \left\langle \underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{nh^2} \right\rangle + \left\langle \sum_{i=1}^n (\bar{y} - \mu)^2 \right\rangle \\ \sum_{i=1}^n \left\langle (y_i - \mu)^2 \right\rangle &= \left\langle \underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{nh^2} \right\rangle + n \left\langle (\bar{y} - \mu)^2 \right\rangle. \end{aligned}$$

All the ensemble averages in the sum on the LHS are the same, and equal the population average, which is the definition of  $\sigma^2$ . On the RHS, we use the known properties of  $\bar{y}$ :

$$\langle \bar{y} \rangle = \mu, \quad \text{var}(\bar{y}) = \langle (\bar{y} - \mu)^2 \rangle = \sigma^2 / n.$$

Then we have:

$$\begin{aligned} n\sigma^2 &= \left\langle \underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{nh^2} \right\rangle + \sigma^2 \\ \left\langle \sum_{i=1}^n (y_i - \bar{y})^2 \right\rangle &= (n-1)\sigma^2. \end{aligned}$$

Thus we see our guess for  $s^2$  is correct. The last equation implies that the unbiased sample estimator is:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

We made no assumptions at all about the distribution of  $y$ , therefore:

$s^2$  is an unbiased estimator of population variance  $\sigma^2$  for any distribution.

### How to Do Statistical Analysis Wrong, and How to Fix It

The following example development contains one error that illustrates a common mistake in statistical analysis: failure to account for dependence between random values. We then show how to correct the error using our statistical algebra. This example re-analyzes an earlier goal: to determine an unbiased estimator for the population variance,  $\sigma^2$ , from a sample of  $n$  values  $\{y_i\}$ .

As before, we start with a guess that our unbiased estimator of  $\sigma^2$  is proportional to the sum squared deviation from the average (similar to the messy attempt we gave up on earlier). Since we know we must introduce  $\mu$  into the computation, we choose to expand the sum by adding and subtracting  $\mu$ :

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n [(y_i - \mu) + (\mu - \bar{y})]^2 = \sum_{i=1}^n [(y_i - \mu)^2 + 2(y_i - \mu)(\mu - \bar{y}) + (\mu - \bar{y})^2].$$

Now we take ensemble averages, and bring them inside the summations:

$$\left\langle \sum_{i=1}^n (y_i - \bar{y})^2 \right\rangle = \sum_{i=1}^n \left\langle (y_i - \mu)^2 \right\rangle + 2 \sum_{i=1}^n \left\langle (y_i - \mu)(\mu - \bar{y}) \right\rangle + n \left\langle (\mu - \bar{y})^2 \right\rangle. \quad (7.5)$$

All the ensemble averages on the RHS now equal their population averages. We consider each of the three terms in turn:

- $\left\langle (y_i - \mu)^2 \right\rangle_{ensemble} = \left\langle (y_i - \mu)^2 \right\rangle_{population} \equiv \sigma^2$ , and the summation in the first term on the right is  $n$  times this.
- In the 2<sup>nd</sup> term on the RHS, the averages of both factors,  $(y_i - \mu)$  and  $(\mu - \bar{y})$ , are zero, so we drop that term.
- $\left\langle (\mu - \bar{y})^2 \right\rangle = \left\langle (\bar{y} - \mu)^2 \right\rangle = \text{var}(\bar{y}) = \sigma^2 / n$ .

Then:

$$\left\langle \sum_{i=1}^n (y_i - \bar{y})^2 \right\rangle = n\sigma^2 + \sigma^2 = (n+1)\sigma^2 \quad \Rightarrow \quad s^2 = \frac{1}{n+1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{wrong!}). \quad (7.6)$$

Clearly, this is wrong: the denominator should be  $(n - 1)$ . What happened? See if you can figure it out before reading further.

Really, stop reading now, and figure out what went wrong. Apply our statistical algebra.

The error is in the second bullet above: just because two RVs both average to zero doesn't mean their product averages to zero (see the average of a product, earlier). In fact, the average of the product must include their covariance. In this case, any given  $y_i$  correlates (positively) with  $\bar{y}$  because  $\bar{y}$  includes each  $y_i$ . Since the  $\bar{y}$  is negated in the 2<sup>nd</sup> factor, the final correlation is negative. Then for a given  $k$ , using the bilinearity of covariance ( $\mu$  is constant):

$$\text{cov}((y_k - \mu), (\mu - \bar{y})) = -\text{cov}(y_k, \bar{y}) = -\text{cov}\left(y_k, \frac{1}{n} \sum_{j=1}^n y_j\right).$$

By assumption, the  $y_i$  are *independent* samples of  $y$ , and therefore have zero covariance between them:

$$\text{cov}(y_k, y_j) = 0, \quad k \neq j, \quad \text{and} \quad \text{cov}(y_k, y_k) = \sigma^2.$$

The only term in the summation over  $j$  that survives the covariance operation is when  $j = k$ :

$$\text{cov}((y_k - \mu), (\mu - \bar{y})) = -\text{cov}\left(y_k, \frac{1}{n} y_k\right) = -\frac{\sigma^2}{n}.$$

Therefore, equation (7.6) should include the summation term from (7.5) that we incorrectly dropped. The ensemble average of each term in that summation is the same, which we just computed, so the result is  $n$  times  $(-\sigma^2/n)$ :

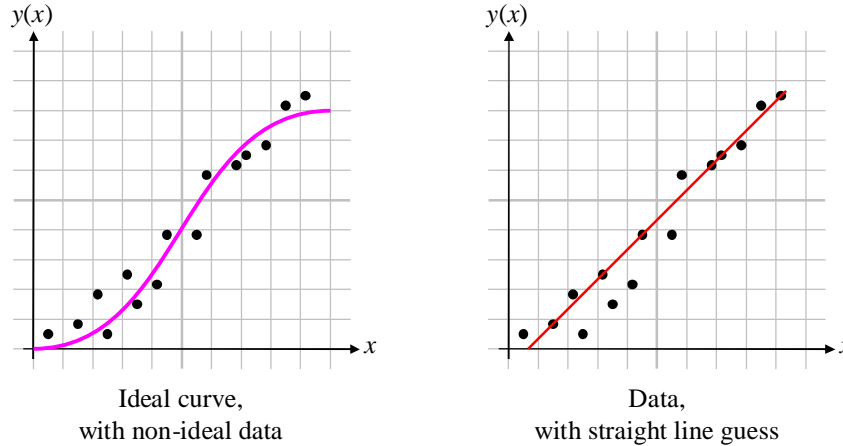
$$\left\langle \sum_{i=1}^n (y_i - \bar{y})^2 \right\rangle = n\sigma^2 - 2n \frac{\sigma^2}{n} + \sigma^2 = (n-1)\sigma^2 \quad \Rightarrow \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{right!}).$$

Order is restored to the universe.



## Introduction to Data Fitting (Curve Fitting)

Suppose we have an ideal process, with an ideal curve mapping an independent variable  $x$  to a dependent variable  $y$ . Now we take a set of measurements of this process, that is, we measure a set of data pairs  $(x_i, y_i)$ , Figure 7.4 left.



**Figure 7.4** (Left) Ideal curve with non-ideal data. (Right) The same data with a straight line fit.

Suppose further we don't *know* the ideal curve, but we have to guess it. Typically, we make a guess of the general form of the curve from theoretical or empirical information, but we leave the exact parameters of the curve "free." For example, we may guess that the form of the curve is a straight line (Figure 7.4 right):

$$y = mx + b,$$

but we leave the slope and intercept ( $m$  and  $b$ ) of the curve as-yet unknown. (We might guess another form, with other, possibly more parameters.) Then we **fit** our curve to the data, which means we compute the values of  $m$  and  $b$  which "best" fit the data. "Best" means that the values of  $m$  and  $b$  minimize some measure of "error," called the **figure of merit**, compared to *all* other values of  $m$  and  $b$ . For data with constant uncertainty, the most common figure of merit is the sum-squared residual:

$$\begin{aligned} \text{sum-squared-residual} &\equiv SSE \equiv \sum_{i=1}^n \text{residual}_i^2 \\ &= \sum_{i=1}^n (\text{measurement}_i - \text{curve}_i)^2 = \sum_{i=1}^n (\text{measurement}_i - f(x_i))^2 \end{aligned}$$

where  $f(x)$  is our fitting function .

The  $(\text{measurement} - \text{curve})$  is often written as  $(O - C)$  for  $(\text{observed} - \text{computed})$ . In our example of fitting to a straight line, for given values of  $m$  and  $b$ , we have:

$$SSE \equiv \sum_{i=1}^n \text{residual}_i^2 = \sum_{i=1}^n (y_i - (mx_i + b))^2 .$$

Curve fitting is the process of finding the values of all our unknown parameters such that (for constant uncertainty) they minimize the sum-squared residual from our data.

The purpose of fitting, in general, is to estimate parameters, some of which may not have simple, closed-form estimators.

We discuss data with varying uncertainty later; in that more general case, we adjust parameters to minimize the  $\chi^2$  parameter.

## Goodness of Fit

### Chi-Squared Distribution

You don't really need to understand the  $\chi^2$  distribution to understand the  $\chi^2$  parameter, but we start there because it's helpful background.

**Notation:**  $X \in D(x)$  means  $X$  is a random variable with probability distribution function (PDF) =  $D(x)$ .  $X \in kD(x)$  means  $X$  is an RV which is  $k$  (a constant) times an RV which is  $\in D$ .

Chi-squared ( $\chi^2$ ) distributions are a family of distributions characterized by one parameter, called  $\nu$  (Greek nu). (Contrast with the gaussian distribution, which has two *real* parameters, the mean,  $\mu$ , and standard deviation,  $\sigma$ .) So we say "chi-squared is a 1-parameter distribution."  $\nu$  is almost always an integer. The simplest case is  $\nu = 1$ : if we define a new random variable  $X$  from a gaussian random variable  $\chi$ , as:

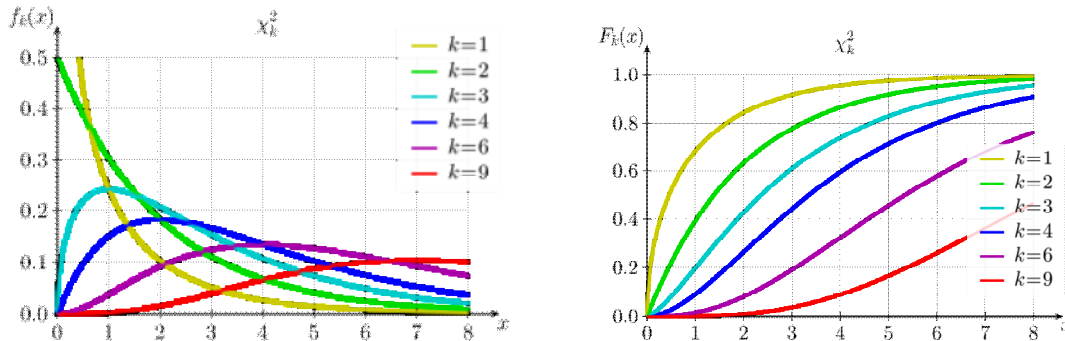
$$X = \chi^2, \quad \text{where } \chi \in \text{gaussian}(\mu = 0, \sigma^2 = 1), \text{ i.e. avg} = 0, \text{ variance} = 1,$$

then  $X$  has a  $\chi^2_1$  distribution. I.e.,  $\chi^2_{\nu=1}(x)$  is the probability distribution function of the square of a zero-mean unit-variance gaussian.

For general  $\nu$ ,  $\chi^2_{\nu}(x)$  is the PDF of the *sum* of the squares of  $\nu$  *independent* gaussian random variables:

$$Y = \sum_{i=1}^{\nu} \chi_i^2, \quad \text{where } \chi_i \in \text{gaussian}(\mu = 0, \sigma^2 = 1), \text{ i.e. avg} = 0, \text{ std deviation} = 1.$$

Thus, the random variable  $Y$  above has a  $\chi^2_{\nu}$  distribution. [picture??] Chi-squared random variables are always  $\geq 0$ , since they are the sums of squares of gaussian random variables. Since the gaussian distribution is continuous, the chi-squared distributions are also continuous.



**Figure 7.5** PDF (left) and CDF (right) of some  $\chi^2$  distributions.  $\chi^2_1(0) \rightarrow \infty$ .  $\chi^2_2(0) = 1/2$ . [[http://en.wikipedia.org/wiki/Chi-squared\\_distribution](http://en.wikipedia.org/wiki/Chi-squared_distribution)]

From the definition, we can also see that the sum of two chi-squared random variables is another chi-squared random variable:

$$\text{Let } A \in \chi^2_n, B \in \chi^2_m, \quad \text{then } A + B \in \chi^2_{n+m}.$$

By the central limit theorem, this means that for large  $\nu$ , chi-squared itself approaches gaussian. However, a  $\chi^2$  random variable (RV) is always positive, whereas any gaussian PDF extends to negative infinity.

We can show that:

$$\begin{aligned} \langle \chi^2_1 \rangle &= 1, & \text{var}(\chi^2_1) &= 2 \\ \Rightarrow \langle \chi^2_{\nu} \rangle &= \nu, & \text{var}(\chi^2_{\nu}) &= 2\nu \Leftrightarrow \text{dev}(\chi^2_{\nu}) = \sqrt{2\nu} \end{aligned}$$

We don't usually need the analytic form, but for completeness:

$$\text{PDF: } \chi^2_{\nu}(x) = \frac{x^{\nu/2-1} e^{-x/2}}{\Gamma(\nu/2) 2^{\nu/2}}. \quad \text{For } \nu \geq 3, \text{ there is a maximum at } \nu - 2.$$

For  $\nu = 1$  or  $2$ , there is no maximum, and the PDF is monotonically decreasing.

### Chi-Squared Parameter

As seen above,  $\chi^2$  is a continuous probability distribution. However, there is a goodness-of-fit test which computes a *parameter* also called “chi-squared.” This parameter is from a distribution that is often close to a  $\chi^2$  distribution, but be careful to distinguish between the *parameter*  $\chi^2$  and the *distribution*  $\chi^2$ .

The chi-squared parameter is not *required* to be from a chi-squared distribution, though it often is. All the chi-squared parameter really requires is that the variances of our residuals add, which is to say that our residuals are uncorrelated (not necessarily independent, though independence implies uncorrelated).

The  $\chi^2$  parameter is valid for any distribution of uncorrelated residuals.  
 The  $\chi^2$  parameter has a  $\chi^2$  distribution only if the residuals are gaussian.

However, for large  $\nu$ , the  $\chi^2$  distribution approaches gaussian, as does the sum of many values of any distribution. Therefore:

The  $\chi^2$  distribution is a reasonable approximation to the distribution of any  $\chi^2$  parameter with  $\nu > \sim 20$ , even if the residuals are not gaussian [ref??].

To illustrate, consider a set of measurements, each with uncertainty  $u$ . Then if the set of  $\{(measurement - model)/u\}$  has zero mean, it has standard-deviation = 1, even for non-gaussian residuals:

Define:  $\text{dev}(X) \equiv$  standard deviation of random variable  $X$ , also written  $\sigma_X$ ,

$$\text{var}(X) \equiv (\text{dev}(X))^2 = \text{variance of random variable } X, \text{ also written } \sigma_X^2.$$

$$\text{dev}\left(\frac{\text{residual}}{u}\right) = 1 \quad \Rightarrow \quad \text{var}\left(\frac{\text{residual}}{u}\right) = 1.$$

As a special case, but *not* required for a  $\chi^2$  parameter, if our residuals are gaussian:

$$\frac{\text{residual}}{u} \in \text{gaussian}(0,1) \quad \Rightarrow \quad \left(\frac{\text{residual}}{u}\right)^2 \in \chi^2_1.$$

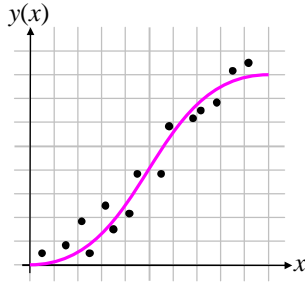
Often, the uncertainties vary from measurement to measurement. In that case, we are fitting a curve to data triples:  $(x_i, y_i, u_i)$ . Still, the error divided by uncertainty for any single measurement is unit deviation:

$$\text{dev}\left(\frac{\text{residual}_i}{u_i}\right) = 1, \quad \text{and} \quad \text{var}\left(\frac{\text{residual}_i}{u_i}\right) = 1, \quad \text{for all } i.$$

If we have  $n$  measurements, with uncorrelated residuals, then because variances add:

$$\text{var}\left(\sum_{i=1}^n \frac{\text{residual}_i}{u_i}\right) = n. \quad \text{For gaussian errors:} \quad \sum_{i=1}^n \left(\frac{\text{residual}_i}{u_i}\right)^2 \in \chi^2_n.$$

Returning to our ideal process from Figure 7.4, with a curve mapping an independent variable  $x$  to a dependent variable  $y$ , we now take a set of measurements with known uncertainties  $u_i$ .



Then our dimensionless parameter  $\chi^2$  is defined as:

$$\chi^2 \equiv \sum_{i=1}^n \left( \frac{\text{residual}_i}{u_i} \right)^2 = \sum_{i=1}^n \left( \frac{\text{measurement}_i - \text{curve}_i}{u_i} \right)^2 \quad \left[ \text{If gaussian residuals, } \chi^2 \in \chi^2_n \right].$$

If  $n$  is large, this sum will be close to the average, and (for zero-mean errors):

$$\langle \chi^2 \rangle = \left\langle \sum_{i=1}^n \left( \frac{\text{residual}_i}{u_i} \right)^2 \right\rangle = n.$$

Now suppose we have **fit** a curve to our data, i.e. we guessed a functional form, and found the parameters which minimize the  $\chi^2$  parameter for that form with our data. If our fit is good, then our curve is very close to the “real” dependence curve for  $y$  as a function of  $x$ , and our errors will be essentially random (no systematic error). We now compute the  $\chi^2$  parameter for our fit:

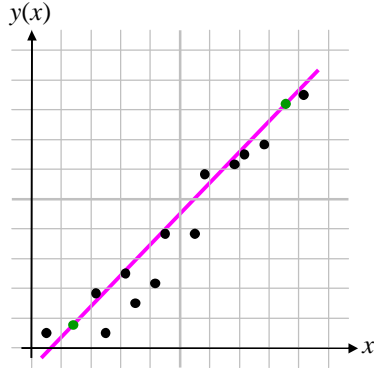
$$\chi^2 \equiv \sum_{i=1}^n \left( \frac{\text{residual}_i}{u_i} \right)^2 = \sum_{i=1}^n \left( \frac{\text{measurement}_i - \text{fit}_i}{u_i} \right)^2.$$

If our fit is good, the number  $\chi^2$  will likely be close to  $n$ . (We will soon modify the distribution of the  $\chi^2$  parameter, but for now, it illustrates our principle.)

If our fit is bad, there will be significant systematic fit error in addition to our random error, and our  $\chi^2$  parameter will be much larger than  $n$ . Summarizing:

If  $\chi^2$  is close to  $n$ , then our fit residuals are no worse than our measurement uncertainties, and the fit is “good.” If  $\chi^2$  is much larger than  $n$ , then our fit residuals are worse than our measurement uncertainties, so our fit must be “bad.”

**Degrees of freedom:** So far we have ignored the “degrees of freedom” of the fit, which we now motivate. (We prove this in detail later.) Consider again a hypothetical fit to a straight line. We are free to choose our parameters  $m$  and  $b$  to define our “fit-line.” But in a set of  $n$  data points, we *could* (if we wanted) choose our  $m$  and  $b$  to *exactly* go through two of the data points:



This guarantees that two of our fit residuals are zero. If  $n$  is large, it won't significantly affect the other residuals, and instead of  $\chi^2$  being the sum of  $n$  squared-residuals, it is approximately the sum of  $(n - 2)$  squared-residuals. In this case,  $\langle \chi^2 \rangle \approx n - 2$ . A rigorous analysis (given later) shows that for the best fit line (which probably doesn't go through *any* of the data points), and gaussian residuals, then  $\langle \chi^2 \rangle = n - 2$ , *exactly*. This concept generalizes quite far:

- Even if we don't fit 2 points exactly to the line;
- Even if our fit-curve is not a line;
- Even if we have more than 2 fit parameters;

the effect is to reduce the  $\chi^2$  parameter to be a sum of less than  $n$  squared-residuals. The effective number of squared-residuals in the sum is called the **degrees of freedom (dof)**, and is given by:

$$dof \equiv n - (\# \text{ fit parameters}).$$

Thus for gaussian residuals, and  $p$  linear fit parameters, the statistics of our  $\chi^2$  parameter are really:

$$\langle \chi^2 \rangle = dof = n - p, \quad \text{dev}(\chi^2) = \sqrt{2(dof)} = \sqrt{2(n - p)}. \tag{7.7}$$

For nonlinear fits, we use the same formula as an *approximation*.

### Reduced Chi-Squared Parameter

Since it is awkward for everyone to know  $n$ , the number of points in our fit, it is convenient to define a "goodness-of-fit" parameter that is independent of  $n$ . We simply divide our chi-squared parameter by  $dof$ , to get the **reduced chi-squared parameter**. Then it has these statistics:

$$\begin{aligned} \text{reduced } \chi^2 &\equiv \frac{\chi^2}{dof} = \frac{1}{dof} \sum_{i=1}^n \left( \frac{\text{measurement}_i - \text{fit}_i}{u_i} \right)^2 &\Rightarrow \\ \langle \text{reduced } \chi^2 \rangle &= \frac{\langle \chi^2 \rangle}{dof} = \frac{dof}{dof} = 1, \\ \text{dev}(\text{reduced } \chi^2) &= \frac{\text{dev}(\chi^2)}{dof} = \frac{\sqrt{2(dof)}}{dof} = \sqrt{\frac{2}{dof}}. \end{aligned}$$

If reduced  $\chi^2$  is close to 1, the fit is "good." If reduced  $\chi^2$  is much larger than 1, the fit is "bad." By "much larger" we mean several deviations away from 1, and the deviation gets smaller with larger  $dof$  (larger  $n$ ).

Of course, our confidence in  $\chi^2$  or reduced- $\chi^2$  depends on how many data points went into computing it, and our confidence in our measurement uncertainties,  $u_i$ . Remarkably, one reference on  $\chi^2$  [which I

don't remember] says that our estimates of measurement uncertainties,  $u_i$ , should come from a sample of at least *five*! That seems to me to be quite small to have much confidence in  $u$ .

---

## Linear Regression

### Review of Multiple Linear Regression

Most intermediate statistics texts cover multiple linear regression, e.g. [W&M p353], but we remind you of some basic concepts here:

A simple example of multiple linear regression is this: you measure some observable  $y$  vs. an independent variable  $x$ , i.e. you measure  $y(x)$  for some set of  $x = \{x_i\}$ . You have a model for  $y(x)$  which is a linear combination of basis functions:

$$y(x) = b_0 + b_1 f_1(x) + b_2 f_2(x) + \dots + b_k f_k(x) = \sum_{m=1}^k b_m f_m(x).$$

You use multiple linear regression to find the coefficients  $b_i$  of the basis functions  $f_i$  which compose the measured function,  $y(x)$ . The basis functions need not be orthonormal. Note that:

Linear regression is *not* limited to fitting data to a straight line.

Fitting data to a line is often called “fitting data to a line” (seriously). We now show that there is no mathematical difference between fitting to a line and *linear* fitting to an arbitrary function (so long as the uncertainties in the  $x$ 's are negligible).

The quirky part is understanding what are the “predictors” (which may be random variables) to which we perform the regression. As above, the predictors can be arbitrary functions of a single independent variable, but they may also be arbitrary functions of *multiple* independent variables. For example, the speed of light in air varies with 3 independent variables: temperature, pressure, and humidity:

$$c = c(T, P, H)$$

Suppose we take  $n$  measurements of  $c$  at various combinations of  $T$ ,  $P$ , and  $H$ . Then our data consists of quintuples:  $(T_i, P_i, H_i, c_i, u_i)$ , where  $u_i$  is the uncertainty in  $c_i$ . We might propose a linear model:

$$c(T, P, H) = b_0 + b_1 T + b_2 P + b_3 H + b_4 TP.$$

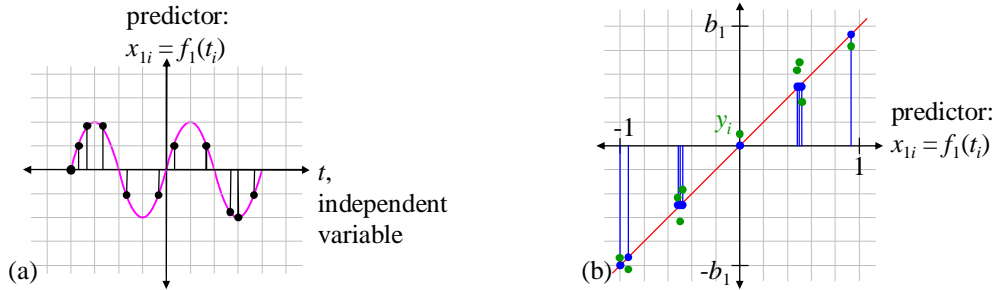
The model is linear because it is a linear combination of arbitrary functions of  $T$ ,  $P$ , and  $H$ . The last term above handles an interaction between temperature and pressure. In terms of linear regression, we have 4 predictors:  $T$ ,  $P$ ,  $H$ , and  $TP$  (the product of T and P).

### We Fit to the Predictors, *Not* the Independent Variable

Figure 7.6 shows an example fit to a model:

$$y_{\text{mod}}(t) = b_1 x_1 = b_1 f_1(t) = b_1 \sin(\omega t) \Rightarrow x_{1i} = \sin(\omega t_i) .$$

There is only 1 fit-function in this example; the predictors are the  $x_{1i}$ . The fit is to the predictors, not to the independent variables  $t_i$ . In some cases, there is no independent variable; there are only predictors (Analysis of Variance includes such cases).



**Figure 7.6** (a) Example predictor: an arbitrary function of independent variable  $t$ . (b) Linear fit to the predictor is a straight line. The fit is *not* to  $t$  itself. Even if the  $t_i$  are evenly spaced, the predictors are not. Note that the predictor values of  $-0.5$  and  $+0.5$  each occur 3 times. This shows a good fit: the measured values (green) are close to the model values.

#### Summarizing:

- Multiple linear regression predicts the values of some random variable  $y_i$  from  $k$  (possibly correlated) **predictors**,  $x_{mi}$ ,  $m = 1, 2, \dots k$ . The predictors may or may not be random variables. In some cases, the predictors are arbitrary functions of a single independent variable,  $t_i$ :  $x_{mi} = f_m(t_i)$ . We assume that all the  $t_i$ ,  $y_i$ , and all the  $f_m$  are *given*., which means all the  $x_{mi} = f_m(t_i)$  are given. In other cases, there are multiple independent variables, and multiple functions of those variables.
- It's *linear* prediction, so our prediction model is that  $y$  is a *linear combination* of the predictors,  $\{x_m\}$ :

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_k x_k = \sum_{m=1}^k b_m x_m .$$

Note that we have included  $b_0$  as a fitted constant, so there are  $k + 1$  fit parameters:  $b_0 \dots b_k$ . This is quite common, in practice, but not always necessary. Note that the *prediction model* has no subscripts of  $i$ , because the model applies to *all*  $x_m$  and  $y$  values.

- Our *measurement* model includes the prediction model, plus measurement noise,  $\varepsilon_i$ :

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \dots + b_k x_{ki} + \varepsilon_i = \left( \sum_{m=1}^k b_m x_{mi} \right) + \varepsilon_i, \quad i = 1, 2, \dots n .$$

For a given set of measurements, the  $\varepsilon_i$  are fixed, but unknown. Over an ensemble of many sets of measurements, the  $\varepsilon_i$  are random variables. The **measurement uncertainty** is defined as the 1-sigma deviation of the noise:

$$u_i \equiv \text{dev}(\varepsilon_i) .$$

Note that the measurement model assumes *additive* noise (as opposed to, say, multiplicative noise).

- Multiple linear regression determines the unknown regression coefficients  $b_0, b_1, \dots b_k$  from  $n$  samples of the  $y$  and each of the  $x_m$ . For least-squares fitting, we simultaneously solve the following  $k + 1$  linear equations in  $k + 1$  unknowns for the  $b_m$  [W&M p355]:

$$b_0 n + b_1 \sum_{i=1}^n x_{1i} + b_2 \sum_{i=1}^n x_{2i} + \dots + b_k \sum_{i=1}^n x_{ki} = \sum_{i=1}^n y_i .$$

And for each  $m = 1, 2, \dots k$ :

$$b_0 \sum_{i=1}^n x_{mi} + b_1 \sum_{i=1}^n x_{mi} x_{1i} + b_2 \sum_{i=1}^n x_{mi} x_{2i} + \dots + b_k \sum_{i=1}^n x_{mi} x_{ki} = \sum_{i=1}^n x_{mi} y_i .$$

Again, all the  $y_i$  and  $x_{mi}$  are given. Therefore, all the sums above are constants, on both the left and right sides. In matrix form, we solve for  $\mathbf{b} \equiv (b_0, b_1, \dots b_k)^T$  from:

$$\mathbf{Xb} = \mathbf{y} \quad \text{or} \quad \begin{bmatrix} n & \sum x_{1i} & \dots & \sum x_{ki} \\ \sum x_{1i} & \sum (x_{1i})^2 & \dots & \sum x_{1i} x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{ki} x_{1i} & \dots & \sum (x_{ki})^2 \end{bmatrix} \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \vdots \\ \sum x_{ki} y_i \end{pmatrix} .$$

**Examples:** For fitting to a line, in our notation, our model is:

$$y(x) = b_0 + b_1 x .$$

$k + 1 = 2$ : our 2 parameters are  $b_0$  and  $b_1$ . Written in terms of functions, we have  $f_1(x) = x$ .

For a sinusoidal periodogram analysis, we typically have a set of measurements  $y_i$  at a set of times  $t_i$ . Given a trial frequency  $\omega$ , we wish to find the least-squares cosine and sine amplitudes that best fit our data. Thus:

$$k = 2: \quad f_1 = \cos, \quad f_2 = \sin, \quad x_{1i} = \cos(\omega t_i), \quad x_{2i} = \sin(\omega t_i), \quad i = 1, 2, \dots n ,$$

and our fit model is:

$$y(t) = b_0 + b_1 \cos(\omega t) + b_2 \sin(\omega t) .$$

(In practice, the (now deprecated) L-S algorithm employs a trick to simplify solving the equations, but we need not consider that here.)

### Fitting to a Polynomial is Multiple Linear Regression

Fitting a polynomial to data is actually a simple example of **multiple linear regression** (see also the Numerical Analysis section for exact polynomial “fits”). Polynomial fit-functions are just a special case of multiple linear regression [W&M p357], where we are predicting  $y_i$  from powers of  $x_i$ . As such, we let

$x_{mi} = (t_i)^m$ , and proceed with standard multiple linear regression:

$$b_0(n) + b_1 \sum_{i=1}^n t_i + b_2 \sum_{i=1}^n t_i^2 + \dots + b_k \sum_{i=1}^n t_i^k = \sum_{i=1}^n y_i .$$

And for each  $m = 1, 2, \dots k$ :

$$b_0 \sum_{i=1}^n t_i^m + b_1 \sum_{i=1}^n t_i^{m+1} + b_2 \sum_{i=1}^n t_i^{m+2} + \dots + b_k \sum_{i=1}^n t_i^{m+k} = \sum_{i=1}^n t_i^m y_i .$$

## The Sum-of-Squares Identity

The sum of squares identity is a crucial tool of linear fitting (aka linear regression). It underlies many of the basic statistics of multiple linear regression and Analysis of Variance (or AOV). The sum of squares identity can be used to define the “coefficient of determination” (and the associated “correlation



coefficient”), and also provides the basis for the F-test and t-test of fit parameter significance. Since ANOVA is actually a special case of multiple linear regression, we describe here the regression view. The ANOVA results then follow directly.

We first consider the case where all the measurements have the same uncertainty,  $\sigma$  (the homoskedastic case). This is a common situation in practice, and also serves as a starting point for the more-involved case where each measurement has its own uncertainty (the heteroskedastic case). Furthermore, there is a transformation from heteroskedastic measurements into an equivalent set of homoskedastic measurements, which are then subject to all of the following homoskedastic results.

We proceed along these steps:

- The raw sum of squares identity.
- The geometric view of a least-squares fit.
- The ANOVA sum of squares identity.
- The failure of the ANOVA sum of squares identity.
- Later, we provide the equivalent formulas for data with individual uncertainties.

Nowhere in this section do we make any assumptions at all about the residuals;  
we do not assume they are gaussian, nor independent, nor even random.

This section assumes you understand the concepts of linear fitting. We provide a brief overview here, and introduce our notation.

A linear fit uses a set of  $p$  coefficients,  $b_1, \dots, b_p$ , as fit parameters in a model with *arbitrary* fit functions. The “model” fit is defined as:

$$y_{\text{mod}}(x) \equiv b_1 f_1(x) + b_2 f_2(x) + \dots + b_p f_p(x) = \sum_{m=1}^p b_m f_m(x).$$

Note that a linear fit does *not* require that  $y$  is a straight-line function of  $x$ .

There is a common special case where we include a constant offset  $b_0$  in the model. In this case, there are  $p-1$  fit functions, since  $p$  is always the *total* number of fit parameters:

$$y_{\text{mod}}(x) \equiv b_0 + b_1 f_1(x) + b_2 f_2(x) + \dots + b_{p-1} f_{p-1}(x) = b_0 + \sum_{m=1}^{p-1} b_m f_m(x).$$

Note that this is equivalent to including a fit function  $f_0(x) = 1$ , so it is really no different than the first model given above. Therefore, the first form is completely general, and includes the second. Anything true of the first form is also true of the second, but the reverse is not true. We use both forms, depending on whether our model includes  $b_0$  or not.

For a set of  $n$  pairs  $(x_i, y_i)$ , the “fit” means finding the values of  $b_m$  that together minimize the sum-squared residual:

define:  $y_{\text{mod},i} \equiv y_{\text{mod}}(x_i) = \sum_{m=1}^p b_m f_m(x_i), \quad \varepsilon_i \equiv y_i - y_{\text{mod},i}.$

minimize:  $SSE \equiv \sum_{i=1}^n (y_i - y_{\text{mod},i})^2 \equiv \sum_{i=1}^n \varepsilon_i^2.$

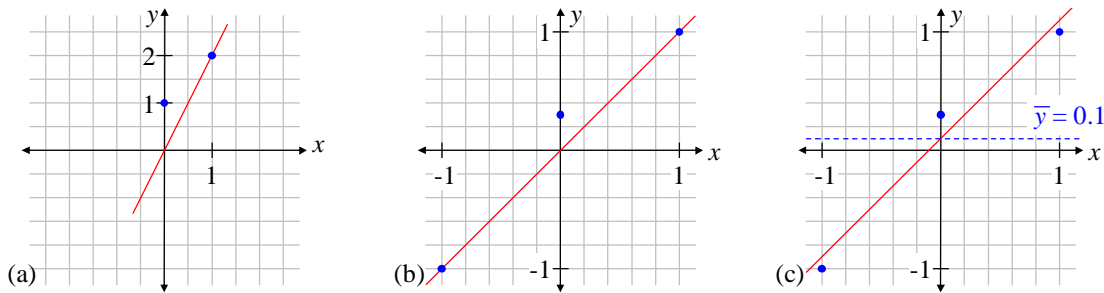
Note that the fit residuals  $\varepsilon_i$  may include both unmodeled behavior, as well as noise (which, by definition, cannot be modeled).

### The Raw Sum-of-Squares Identity

Most references do not consider the raw sum of squares (SSQ) identity. We present it first because it provides a basis for the more-common ANOVA SSQ identity, and it is sometimes useful in its own right. Consider a set of data  $(x_i, y_i), i = 1, \dots, n$ . Conceptually, the SSQ identity says the sum of the squares of the  $y_i$  can be partitioned into a sum of squares of model values plus a sum of squares of “residuals” (often called “errors”):

$$(raw) \quad SST = SSA + SSE: \quad \sum_{i=1}^n y_i^2 = \sum_{i=1}^n y_{mod,i}^2 + \sum_{i=1}^n (y_i - y_{mod,i})^2. \quad (7.8)$$

(The term “errors” can be misleading, so in words we always use “residuals.” However, we write the term as SSE, because that is so common in the literature.) The SSQ identity is *only* true for a least-squares linear fit to a parametrized model, and has some important non-obvious properties. We start with some examples of the identity, and provide simple proofs later.



**Figure 7.7** (a) Two data points,  $n = 2$ , and best-fit 1-parameter model. (b) Three data points,  $n = 3$ , and best-fit 1-parameter model. (c) Three data points,  $n = 3$ , and best-fit 2-parameter model.

**Example:**  $n = 2, p = 1$ : Given a data set of two measurements  $(0, 1)$ , and  $(1, 2)$  (Figure 7.7a). We choose a 1-parameter model:

$$y(x) = b_1 x.$$

The best fit line is  $b_1 = 2$ , and therefore  $y(x) = 2x$ . (We see this because the model is forced through the origin, so the residual at  $x = 0$  is fixed. Then the least squares residuals are those that minimize the error at  $x = 1$ , which we can make zero.) Our raw sum-of-squares identity (7.8) is:

$$\underbrace{1^2 + 2^2}_{SST} = \underbrace{(0^2 + 2^2)}_{SSA} + \underbrace{(1^2 + 0^2)}_{SSE} \quad \rightarrow \quad 5 = 4 + 1.$$

**Example:**  $n = 3, p = 1$ : Given a data set of three measurements  $(-1, -1)$ ,  $(0, 0.3)$ , and  $(1, 1)$  (Figure 7.7b). We choose a 1-parameter model:

$$y(x) = b_1 x.$$

The best fit line is  $b_1 = 1$ , and therefore  $y(x) = x$ . (We see this because the model is forced through the origin, so the residual at  $x = 0$  is fixed. Then the least squares residuals are those that minimize the errors at  $x = -1$  and  $x = 1$ , which we can make zero.) Our raw sum-of-squares identity (7.8) is:

$$\underbrace{1^2 + 0.3^2 + 1^2}_{SST} = \underbrace{((-1)^2 + 0^2 + 1^2)}_{SSA} + \underbrace{(0^2 + 0.3^2 + 0^2)}_{SSE} \quad \rightarrow \quad 2.09 = 2 + 0.09.$$

**Example:**  $n = 3, p = 2$ : We consider the same data:  $(-1, -1)$ ,  $(0, 0.3)$ , and  $(1, 1)$ , but we now include a  $b_0$  DC-offset parameter in the model:

$$y(x) = b_0 + b_1 x.$$

The best fit line is  $b_0 = 0.1, b_1 = 1$ , and therefore  $y(x) = 0.1 + x$ , shown in Figure 7.7c. (We see this because the fit functions are orthogonal over the given  $\{x_i\}$ , and therefore the fit parameters  $\{b_m\}$  can be found by correlating the data with the fit functions, normalized over the  $\{x_i\}$ . Trust me on this.)

$$\frac{1^2 + 0.3^2 + 1^2}{SST} = \underbrace{\left( \frac{(-1)^2 + 0^2 + 1^2}{SSA} \right)}_{SSA} + \underbrace{\left( \frac{0^2 + 0.3^2 + 0^2}{SSE} \right)}_{SSE} \rightarrow 2.09 = 2 + 0.09.$$

The *raw* sum-of-squares identity holds for *any* linear least-squares fit, even with non-gaussian (or non-random) residuals.

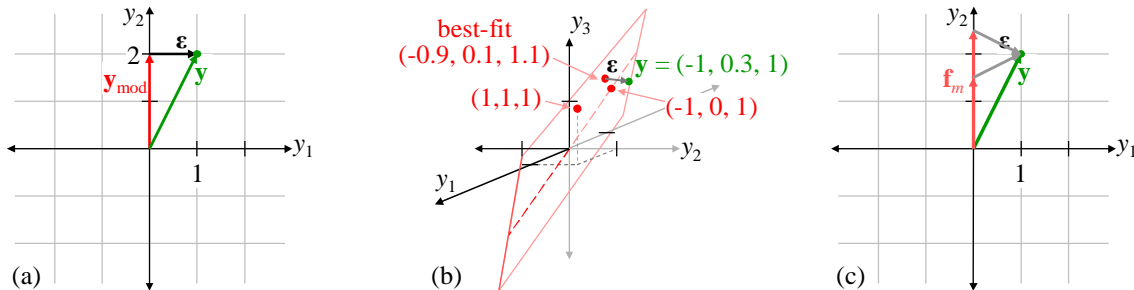
In general, the SSQ identity does not hold for nonlinear fits, as is evident from the following sections. This means that none of the *linear* regression statistics are valid for a *nonlinear* fit.

### The Geometric View of a Least-Squares Fit

The geometric view of least-squares fitting requires defining a new kind of vector space: measurement space (aka “observation space”). This is an  $n$ -dimensional space, where  $n \equiv$  the number of measurements in the data set. Our sets of measurements  $\{y_i\}$ , residuals  $\{\varepsilon_i\}$ , etc. can be viewed as vectors:

$$\mathbf{y} \equiv (y_1, y_2, \dots, y_n), \quad \boldsymbol{\varepsilon} \equiv (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n), \quad \text{etc.}$$

Thus, the entire set of measurements is a *single point* in measurement space (Figure 7.8). We write that point as the displacement vector  $\mathbf{y}$ . If we have 1000 measurements, then measurement space is 1000-dimensional. Measurement space is the space of all possible data sets  $\{y_i\}$ , with the  $\{x_i\}$  fixed.



**Figure 7.8** (a) Measurement space,  $n = 2$ , and best-fit 1-parameter model. (b) Measurement space,  $n = 3$ , and the 2-parameter model surface within it. (c) The shortest  $\boldsymbol{\varepsilon}$  is perpendicular to every  $\mathbf{f}_m$ .

Given a set of parameters  $\{b_m\}$  and the sample points  $\{x_i\}$ , the model (with no residuals) defines a set of measurements,  $y_{mod,i}$ , which can also be plotted as a single point in measurement space. For example, Figure 7.8a shows our  $n = 2$  model  $y = b_1x$ , taken at the two abscissa value  $x_1 = 0$ , and  $x_2 = 1$ , which gives  $y_{mod,1} = 0, y_{mod,2} = b_1$ . The least squares fit is  $b_1 = 2$ . Then the coordinates  $(y_{mod,1}, y_{mod,2}) = (0, 2)$  give the model vector  $\mathbf{y}_{mod}$  in Figure 7.8a.

Note that by varying the  $b_m$ , the model points in measurement space define a  $p$ -dimensional subspace of it. In Figure 7.8a, different values of  $b_1$  trace out a vertical line through the origin. In this case,  $p = 1$ , so the subspace is 1D: a line.

The  $n = 3$  case is shown in Figure 7.8b. Here,  $p = 2$ , so the model subspace is 2D: a plane in the 3D measurement space. Different values of  $b_0$  and  $b_1$  define different model points in measurement space. For a linear fit, the origin is always on the model surface: when all the  $b_m = 0$ , all the model  $y_i = 0$ . Therefore, the plane goes through the origin. Two more points define the plane:

$$\begin{aligned} b_0 = 1, b_1 = 0 &\Rightarrow \mathbf{y} = (1, 1, 1) \\ b_0 = 0, b_1 = 1 &\Rightarrow \mathbf{y} = (-1, 0, 1) \end{aligned}$$

As shown, the model plane passes through these points. Again using linearity, note that any model vector (point) lies on a ray from the origin, and the entire ray is within the model surface. In other words, you can

scale any model vector by any value to get another model vector. To further visualize the plane, note that whenever  $b_1 = -b_0$ ,  $y_3 = 0$ . Then  $y_1 = -b_1 + b_0 = 2b_0$ , and  $y_2 = b_0$ ; therefore, the line  $y_2 = 0.5 y_1$  lies in the model surface, and is shown with a dashed line in Figure 7.8b.

The green dot in Figure 7.8b is the measurement vector  $\mathbf{y}$  (in front of the model plane). The best-fit model point is  $(-0.9, 0.1, 1.1)$ . The residual vector  $\boldsymbol{\varepsilon}$  goes from the model to  $\mathbf{y}$ , and is perpendicular to the model plane.

The model surface is entirely determined by the model (the  $f_m(x)$ ), and the sample points  $\{x_i\}$ .

The measured values  $\{y_i\}$  will then determine the best-fit model, which is a point on the model surface.

In Figure 7.8a and b, we see that the residual vector is perpendicular to the best-fit linear model vector. Is this always the case? Yes. If the model vector were shorter (Figure 7.8c),  $\boldsymbol{\varepsilon}$  would have to reach farther to go from there to the measurement vector  $\mathbf{y}$ . Similarly, if the model vector were longer,  $\boldsymbol{\varepsilon}$  would also be longer. Therefore the shortest residual vector (least sum squared residual) must be perpendicular to the best-fit model vector. This is true in any number of dimensions. From this geometry, we can use the  $n$ -dimensional Pythagorean Theorem to prove the sum of squares identity immediately (in vector notation):

$$\boldsymbol{\varepsilon} \cdot \mathbf{y}_{\text{mod}} = 0 \quad \Rightarrow \quad \underbrace{\mathbf{y}^2}_{SST} = \underbrace{\mathbf{y}_{\text{mod}}^2}_{SSA} + \underbrace{\boldsymbol{\varepsilon}^2}_{SSE} \quad \text{where} \quad \mathbf{y}^2 \equiv \mathbf{y} \cdot \mathbf{y}, \text{ etc.}$$

**Fit parameters as coordinates of the model surface:** We've seen that each point on the model surface corresponds to a unique set of  $\{b_m\}$ . Therefore, the  $b_m$  compose a new coordinate system for the model surface, different from the  $y_i$  coordinates. For example, in Figure 7.8b, the  $b_0$  axis is defined by setting  $b_1 = 0$ . This is the line through the origin and the model point  $\mathbf{y} = (1, 1, 1)$ . The  $b_1$  axis is defined by setting  $b_0 = 0$ . This is the line through the origin and  $\mathbf{y} = (-1, 0, 1)$ . In general, the  $b_m$  axes need not be perpendicular, though in the case, they are.

In Figure 7.8b,  $\boldsymbol{\varepsilon}$  is perpendicular to every vector in the model plane. In general,  $\boldsymbol{\varepsilon}$  is perpendicular to every  $\mathbf{f}_m$  vector (i.e. each of the  $m$  components of the best-fit model vector):

$$\boldsymbol{\varepsilon} \cdot \mathbf{f}_m = 0 \quad \forall m = 1, \dots, p \quad \text{where} \quad \mathbf{f}_m \equiv b_m (f_m(x_1), f_m(x_2), \dots, f_m(x_n)).$$

Again, this must be so to minimize the length of  $\boldsymbol{\varepsilon}$ , because if  $\boldsymbol{\varepsilon}$  had any component parallel to any  $\mathbf{f}_m$ , then we could make that  $\mathbf{f}_m$  longer or shorter, as needed, to shrink  $\boldsymbol{\varepsilon}$  (Figure 7.8c). We'll use this perpendicularity in the section on the algebra of the sum of squares.

### Algebra and Geometry of the Sum-of-Squares Identity

We now prove the sum of squares (SSQ) identity algebraically, and highlight its corresponding geometric features. We start by simply subtracting and adding the model values  $y_{\text{mod},i}$  in the sum of squares:

$$\begin{aligned} \sum_{i=1}^n y_i^2 &= \sum_{i=1}^n ((y_i - y_{\text{mod},i}) + y_{\text{mod},i})^2 = \sum_{i=1}^n (\varepsilon_i + y_{\text{mod},i})^2 \\ &= \sum_{i=1}^n \varepsilon_i^2 + \sum_{i=1}^n y_{\text{mod},i}^2 + \sum_{i=1}^n 2\varepsilon_i y_{\text{mod},i}. \end{aligned} \tag{7.9}$$

The last term is  $\boldsymbol{\varepsilon} \cdot \mathbf{y}_{\text{mod}}$ , which we've seen geometrically is zero. We now easily show it algebraically: since SSE is minimized w.r.t. all the model parameters  $b_m$ , its derivative w.r.t. each of them is zero. I.e., for each  $k$ :

$$\frac{\partial SSE}{\partial b_k} = 0 = \frac{\partial}{\partial b_k} \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n 2\varepsilon_i \frac{\partial}{\partial b_k} \varepsilon_i = 2 \sum_{i=1}^n \varepsilon_i \frac{\partial}{\partial b_k} \left( y_i - \sum_{m=1}^p b_m f_m(x_i) \right).$$

In this equation, all the  $y_i$  are constant. The only term that survives the partial derivative is where  $m = k$ . Dividing by  $-2$ , we get:

$$0 = \sum_{i=1}^n \varepsilon_i \frac{\partial}{\partial b_k} b_k f_k(x_i) = \sum_{i=1}^n \varepsilon_i f_k(x_i) \quad \Leftrightarrow \quad \boldsymbol{\varepsilon} \cdot \mathbf{f}_m = 0. \quad (7.10)$$

Therefore, the last term in (7.9) drops out, leaving the SSQ identity.

### The ANOVA Sum-of-Squares Identity

It is often the case that the DC offset in a set of measurements is either unmeasurable, or not relevant. This leads to **ANalysis Of Variance (ANOVA)**, or analysis of how the data varies from its own average. In the ANOVA case, the sum-of-squares identity is modified: we subtract the data average  $\bar{y}$  from both the  $y_i$  and the  $y_{mod,i}$ :

$$(ANOVA) \quad SST = SSA + SSE : \quad \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_{mod,i} - \bar{y})^2 + \sum_{i=1}^n (y_i - y_{mod,i})^2. \quad (7.11)$$

This has an important consequence which is often overlooked: the ANOVA sum-of-squares identity *holds only if the model includes a DC offset (constant) fit parameter*, which we call  $b_0$ .

**Example:**  $n = 3, p = 2$ : We again consider the data of Figure 7.7c:  $(-1, -1)$ ,  $(0, 0.3)$ , and  $(1, 1)$ . We now use the ANOVA sum-of-squares, which is allowed because we have a  $b_0$  (DC offset) in the model:

$$y(x) = b_0 + b_1 x.$$

Our ANOVA sum-of-squares identity (7.11) is, using  $\bar{y} = 0.1$ :

$$\underbrace{(-1.1)^2 + 0.2^2 + 0.9^2}_{SST} = \underbrace{\left( (-1)^2 + 0^2 + 1^2 \right)}_{SSA} + \underbrace{\left( (-0.1)^2 + 0.2^2 + (-0.1)^2 \right)}_{SSE} \quad \rightarrow \quad 2.06 = 2 + 0.06.$$

The ANOVA sum-of-squares identity holds for any linear least-squares fit that includes a DC offset fit parameter (and also in the special case that the sum of residuals (not squared) = 0).

With no DC offset parameter in the model, in general, the ANOVA sum-of-squares identity fails.

We prove the ANOVA SSQ identity (often called just “the sum of squares identity”) similarly to our proof of the raw SSQ identity. We start by subtracting and adding  $y_{mod,i}$  to each term:

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n \left( (y_i - y_{mod,i}) + (y_{mod,i} - \bar{y}) \right)^2 = \sum_{i=1}^n \left( \varepsilon_i + (y_{mod,i} - \bar{y}) \right)^2 \\ &= \sum_{i=1}^n \varepsilon_i^2 + \sum_{i=1}^n (y_{mod,i} - \bar{y})^2 + \sum_{i=1}^n 2\varepsilon_i (y_{mod,i} - \bar{y}) \\ &= \sum_{i=1}^n \varepsilon_i^2 + \sum_{i=1}^n (y_{mod,i} - \bar{y})^2 + \sum_{i=1}^n 2\varepsilon_i y_{mod,i} + 2\bar{y} \sum_{i=1}^n \varepsilon_i. \end{aligned}$$

Compared to the raw SSQ proof, there is an extra 4<sup>th</sup> term. The 3<sup>rd</sup> term is zero, as before, because  $\boldsymbol{\varepsilon}$  is shortest when it is perpendicular to the model. The 4<sup>th</sup> term is zero when the sum of the residuals is zero. This might happen by chance (but don’t count on it). However, it is guaranteed if we include a DC offset parameter  $b_0$  in the model. Recall that the constant  $b_0$  is equivalent to a fit function  $f_0(x) = 1$ . We know from the raw SSQ proof that for every  $k$ :

$$\boldsymbol{\varepsilon} \cdot \mathbf{b}_k \equiv \sum_{i=1}^n \varepsilon_i f_k(x_i) = 0 \quad \Rightarrow \quad \sum_{i=1}^n \varepsilon_i f_0(x_i) = \sum_{i=1}^n \varepsilon_i = 0.$$

QED.

The necessary and sufficient condition for the ANOVA SSQ identity to hold is that the sum of the residuals is zero. A sufficient condition (and the most common) is that the fit model contains a constant (DC offset) fit parameter  $b_0$ .

### The Failure of the ANOVA Sum-of-Squares Identity

The ANOVA sum-of-squares identity fails when the sum of the residuals is not zero:

$$\sum_{i=1}^n \varepsilon_i \neq 0 \quad \Rightarrow \quad (\text{ANOVA}) \text{ SST} \neq \text{SSA} + \text{SSE} .$$

(We proved this when we proved the ANOVA SSQ identity.) This pretty much mandates including a  $b_0$  parameter, which guarantees the sum of the residuals is zero. You might think this is no problem, because everyone probably already has a  $b_0$  parameter; however, the traditional Lomb-Scargle algorithm [Sca 1982] fails to include a  $b_0$  parameter, and therefore all of its statistics are incorrect. The error is worse for small sample sizes, and better for large ones.

As an example of the failure of the sum-of-squares identity, consider again the data of Figure 7.7a:  $n = 2$  measurements, (0, 1), and (1, 2). As before, we fit the raw data to  $y = b_1x$ , and the best-fit is still  $b_1 = 2$ . We now incorrectly try the ANOVA sum-of-squares identity, with  $\bar{y} = 1.5$ , and find it fails:

$$\underbrace{\left(-\frac{1}{2}\right)^2}_{\text{SST}} + \underbrace{\left(\frac{1}{2}\right)^2}_{\text{SSA}} \stackrel{?}{=} \underbrace{\left((-1.5)^2 + 0.5^2\right)}_{\text{SSA}} + \underbrace{(1^2 + 0^2)}_{\text{SSE}} \rightarrow \frac{1}{2} \neq 2.5 + 1 .$$

For another example, consider again the  $n = 3$  data from earlier: (-1, -1), (0, 0.3), and (1, 1). If we fit with just  $y = b_1x$ , we saw already that  $b_1 = 1$  (Figure 7.7b). As expected, because there is no constant fit parameter  $b_0$ , the sum of the residuals is not zero:

$$\sum_{i=1}^n \varepsilon_i \equiv \sum_{i=1}^n (y_i - y_{\text{mod},i}) = 0 + 0.3 + 0 \neq 0 .$$

Therefore, the ANOVA sum-of-squares identity fails:

$$\underbrace{(-1.1)^2 + 0.2^2 + 0.9^2}_{\text{SST}} \stackrel{?}{=} \underbrace{\left((-1.1)^2 + (-0.1)^2 + 0.9^2\right)}_{\text{SSA}} + \underbrace{(0^2 + 0.3^2 + 0^2)}_{\text{SSE}} \rightarrow 2.06 \neq 2.03 + 0.09 .$$

In the above two examples, the fit function had no DC component, so you might wonder if including such a fit function would restore the ANOVA SSQ identity. It doesn't, because the condition for the ANOVA SSQ identity to hold is that the sum of residuals is zero. To illustrate, we add a fit function,  $(x^2 + 1)$  with a nonzero DC (average) value, so our model is this:

$$y_{\text{mod}}(x) = b_1x + b_2(x^2 + 1) .$$

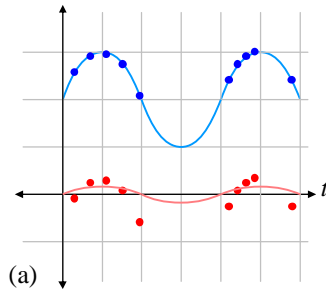
The best fit is  $b_1 = 1$  (as before), and  $b_2 = 0.0333$  (from correlation). Then  $y_{\text{mod},i} = (-0.933, 0.0333, 1.0667)$ , and:

$$\underbrace{(-1.1)^2 + 0.2^2 + 0.9^2}_{\text{SST}} \stackrel{?}{=} \underbrace{\left((-1.033)^2 + (-0.0667)^2 + 0.967^2\right)}_{\text{SSA}} + \underbrace{\left((-0.0667)^2 + 0.267^2 + (0.0667)^2\right)}_{\text{SSE}} \rightarrow 2.06 \neq 2.007 + 0.08 .$$

### Subtracting DC Before Analysis

A common method of trying to avoid problems of DC offset is to simply subtract the average of the data before fitting to it. This generally fails to solve the DC problem (though it is often advisable for improved numerical accuracy in calculations). Subtracting DC makes  $\bar{y} = 0$ , so the ANOVA SSQ identity

is the *same* as the raw SSQ identity, and the raw identity always holds. However, subtracting DC does *not* give an optimal fit when the fit functions have a DC offset over the  $\{x_i\}$ . The traditional Lomb-Scargle analysis [Sca 1982] has this error. The only solution is to use a 3-parameter fit: a constant, a cosine component, and a sine component [Zeich 2009].



**Figure 7.9** (a) The top curve (blue) shows a cosine fit to data points. The bottom curve (red) shows the same frequency fit to DC-subtracted data, and is a much worse fit.

Figure 7.9 shows an example of the failure of DC-subtraction to fix the problem, and how DC-subtraction can lead to a much worse fit. Therefore:

We must include the constant  $b_0$  parameter both to enable the other parameters to be properly fit, and to enable Analysis of Variance with the SSQ identity.

In general, any fit parameter that we must include in the model, but whose value we actually don't need, is called a **nuisance parameter**.  $b_0$  is probably the most common nuisance parameter in data analysis.

### Fitting to Orthonormal Functions

For  $p$  orthonormal fit functions, each  $b_m$  can be found by a simple inner product:

$$y_{\text{mod}}(x) \equiv \sum_{m=1}^p b_m f_m(x), \quad \mathbf{f}_j \cdot \mathbf{f}_k = \delta_{jk} \quad \Rightarrow \quad b_m \equiv \mathbf{f}_m \cdot \mathbf{y}.$$

As examples, this is how Fourier Transform coefficients are found, and usually how we find components of a ket in quantum mechanics.

### Hypothesis Testing with the Sum of Squares Identity

A big question for some data analysts is, “Is there a signal in my data?” For example, “Is the star’s intensity varying periodically?” One approach to answering this question is to fit for the signal you expect, and then test the probability that the fit is just noise. This is a simple form of **Analysis of Variance** (ANOVA). This type of hypothesis is widely used throughout science, e.g. astronomers use this significance test in Lomb-Scargle and Phase Dispersion Minimization periodograms.

To make progress in determining if a signal is present, we will test the hypothesis:

$$H_0: \text{there is no signal, i.e. our data is pure noise.}$$

This is called the **null hypothesis**, because we usually define it to be a hypothesis that nothing interesting is in our data, e.g. there is no signal, our drug doesn’t cure the disease, the two classes are performing equally well, etc.

After our analysis, we make one of two conclusions: either we *reject*  $H_0$ , or we *fail to reject* it. It is crucial to be crystal clear in our logic here. If our analysis shows that  $H_0$  is unlikely to be true, then we reject  $H_0$ , and take it to be false. We also quantify our confidence level in rejecting  $H_0$ , typically 95% or better. Rejecting  $H_0$  means there *is* a signal, i.e. our data is *not* pure noise. Note that rejecting  $H_0$ , by itself, tells us nothing about the nature of the signal that we conclude is present. In particular, it may or may not match the model we fitted for (but it certainly must have some correlation with our model).

However, if our analysis says  $H_0$  has even a fair chance of being true (typically  $> 5\%$ ), then we do not reject it.

Failing to reject  $H_0$  is *not the same* as accepting it. Failing to reject means either (a)  $H_0$  is true; or (b)  $H_0$  is false, but our data are insufficient to show that confidently.

This point cannot be over-emphasized.

Notice that scientists are a conservative lot: if we claim a detection, we want to be highly confident that our claim is true. It wouldn't do to have scientists crying "wolf" all the time, and being wrong a lot. The rule of thumb in science is, "If you are not highly confident, then don't make a claim." You can, however, say that your results are intriguing, and justify further investigation.

## Introduction to Analysis of Variance (ANOVA)

ANOVA addresses the question: Why don't all my measurements equal the average? The "master equation" of ANOVA is the sum of squares identity (see The Sum-of-Squares Identity section):

$$SST = SSA + SSE \quad \text{where} \quad \begin{aligned} SST &\equiv \text{total sum of squared variation} \\ SSA &\equiv \text{modeled sum of squared variation} \\ SSE &\equiv \text{residual sum of squared variation} \end{aligned}$$

This equation says that in our data, the total of "differences" from the average is the measured differences from the model, plus the unmodeled residuals. Specifically, the total sum of squared differences ( $SST$ ) equals the modeled sum of squared differences ( $SSA$ ) plus the residual (unmodeled + noise) sum of squared differences ( $SSE$ ).

As shown earlier, for a least-squares linear fit, the master equation (the SSQ identity) requires no statistics or assumptions of any kind (normality, independence, ...).

[ANOVA is identical to least-squares linear regression (fitting) to the "categorical variables." More later.]

To test a hypothesis, we must consider that our data is only one set of many possible sets that might have been taken, each with different noise contributions,  $\varepsilon_i$ . Recall that when considered over an ensemble of hypothetical data sets, all the fit parameters  $b_m$ , as well as  $SST$ ,  $SSA$ , and  $SSE$  are random variables. It is in this sense that we speak of their statistical properties.

For concreteness, consider a time sequence of data, such as a light curve with pairs of times and intensities,  $(t, s)$ . Why do the measured intensities *vary* from the average? There are conceptually three reasons:

- We have an accurate *model*, which predicts deviations from the average.
- The system under study is more complex than our model, so there are *unmodeled*, but systematic, deviations.
- There is noise in the measurement (which by definition, cannot be modeled).

However, mathematically we can distinguish only two reasons for variation in the measurements: either we predict the variation with a model, or we don't, i.e. modeled effects, and unmodeled effects. Therefore, in practice, the 2<sup>nd</sup> and 3<sup>rd</sup> bullets above are combined into **residuals**: unmodeled variations in the data, which includes both systematic physics and measurement noise.

This section requires a conceptual understanding of vector decomposition into both orthonormal and non-orthonormal basis sets.

## The Temperature of Liberty

As prerequisite to hypothesis testing, we must consider a number of properties of the fit coefficients  $b_k$  that occur when we apply linear regression to measurements  $\mathbf{y}$ . We then apply these results to the case when the "null hypothesis" is true: there is no signal (only noise). We proceed along these lines:



- A look ahead to our goal.
- The distribution of orthonormal fit coefficients,  $b_m$ .
- The non-correlation of orthonormal fit coefficients in pure noise.
- The model sum-of-squares (SSA).
- The residual sum-of-squares (SSE) in pure noise.

**A Look Ahead to the Result Needed for Hypothesis Testing**

To better convey where we are headed, the following sections will prove the degrees-of-freedom decomposition of the sum-of-squares (SSQ) identity:

$$(raw) \quad SST = SSA + SSE \quad \rightarrow \quad \underbrace{\mathbf{y}^2}_{dof=n} = \underbrace{\mathbf{y}_{mod,i}^2}_{dof=p} + \underbrace{\boldsymbol{\epsilon}^2}_{dof=n-p} .$$

We already proved the SSQ identity holds for any least-squares linear fit (regardless of the distribution of SSE). To perform hypothesis testing, we must further know that for pure noise, the  $n$  degrees of freedom (dof) of SST also separate into  $p$  dof in SSA, and  $n - p$  dof in SSE.

For the ANOVA SSQ identity, the subtraction of the average reduces the dof by 1, so the dof partition as:

$$(ANOVA) \quad SST = SSA + SSE \quad \rightarrow \quad \underbrace{(\mathbf{y} - \bar{\mathbf{y}})^2}_{dof=n-1} = \underbrace{(\mathbf{y}_{mod,i} - \bar{\mathbf{y}})^2}_{dof=p-1} + \underbrace{\boldsymbol{\epsilon}^2}_{dof=n-p} .$$

**Distribution of Orthogonal Fit Coefficients in the Presence of Pure Noise**

We have seen that if a fit function is orthogonal to all other fit functions, then its fit coefficient is given by a simple correlation. I.e., for a given  $k$ :

$$\mathbf{f}_k \cdot \mathbf{f}_j = 0 \text{ for all } j \neq k \quad \Rightarrow \quad b_k = \frac{\mathbf{f}_k \cdot \mathbf{y}}{\mathbf{f}_k \cdot \mathbf{f}_k} \equiv \frac{\sum_{i=1}^n f_k(x_i) y_i}{\sum_{i=1}^n f_k(x_i)^2} . \tag{7.12}$$

We now further restrict ourselves to a normalized (over the  $\{x_i\}$ ) fit-function, so that:

$$\sum_{i=1}^n f_k(x_i)^2 = 1 \quad \Rightarrow \quad b_k = \sum_{i=1}^n f_k(x_i) y_i .$$

We now consider an ensemble of sample sets of noise, each with the same set of  $\{x_i\}$ , and each producing a random  $b_k$ . In other words, the  $b_k$  are RVs over the set of possible sample-sets. Therefore, in the presence of pure noise, we can easily show that  $\text{var}(b_k) = \text{var}(y) \equiv \sigma^2$ . Recall that the variance of a sum (of uncorrelated RVs) is the sum of the variances, and the variance of  $k$  times an RV =  $k^2 \text{var}(\text{RV})$ . All the values of  $f_k(x_i)$  are constants, and  $\text{var}(y_i) = \text{var}(y) \equiv \sigma^2$ ; therefore from (7.12):

$$\text{var}(b_k) = \underbrace{\left( \sum_{i=1}^n f_k(x_i)^2 \right)}_1 \text{var}(y_i) = \sigma^2 .$$

This is a remarkable and extremely useful result:

In pure noise, for a normalized fit-function orthogonal to all others, the variance of its least-squares linear fit coefficient is that of the noise, regardless of the noise PDF.

At this point, the noise need not be zero-mean. In fact:

$$\langle b_k \rangle = \left( \sum_{i=1}^n f_k(x_i) \right) \underbrace{\langle y_i \rangle}_{\mu_y}$$

Since the sum has no simple interpretation, this equation is most useful for showing that if the noise is zero-mean, then  $b_k$  is also zero-mean:  $\langle b_k \rangle = 0$ . However, if the fit-function  $f_k$  taken over the  $\{x_i\}$  happens to be zero mean, then the summation is zero, and even for non-zero mean noise, we again have  $\langle b_k \rangle = 0$ .

Similarly, any weighted sum of gaussian RVs is a gaussian; therefore, if the  $y_i$  are gaussian (zero-mean or not), then  $b_k$  is also gaussian.

**Non-correlation of Orthogonal Fit Coefficients in Pure Noise**

We now consider the correlation between two fit coefficients,  $b_k$  and  $b_m$  (again, over multiple samples (sample sets) of noise), when the fit-functions  $f_k$  and  $f_m$  are orthogonal to each other, and to all other fit-functions. We show that the covariance  $\text{cov}(b_k, b_m) = 0$ , and so the coefficients are uncorrelated. For convenience, we take  $f_k$  and  $f_m$  to be normalized:  $\mathbf{f}_k^2 = \mathbf{f}_m^2 = 1$ . We start with the formula for a fit-coefficient of a fit-function that is orthogonal to all others, (7.12), and use our algebra of statistics:

$$\text{cov}(b_k, b_m) = \text{cov}(\mathbf{f}_k \cdot \mathbf{y}, \mathbf{f}_m \cdot \mathbf{y}) = \text{cov} \left( \sum_{i=1}^n f_k(x_i) y_i, \sum_{j=1}^n f_m(x_j) y_j \right)$$

Again, all the  $f_k$  and  $f_m$  are constants, so they can be pulled out of the  $\text{cov}()$  operator:

$$\text{cov}(b_k, b_m) = \sum_{i=1}^n \sum_{j=1}^n f_k(x_i) f_m(x_j) \text{cov}(y_i, y_j)$$

As always, the  $y_i$  are independent, and therefore uncorrelated. Hence, when  $i \neq j$ ,  $\text{cov}(y_i, y_j) = 0$ , so only the  $i = j$  terms survive, and the double sum collapses to a single sum. Also,  $\text{cov}(y_i, y_i) = \text{var}(y_i) = \sigma^2$ , which is a constant:

$$\text{cov}(b_k, b_m) \propto \underbrace{\sigma^2 \sum_{i=1}^n f_k(x_i) f_m(x_i)}_0 = 0 \quad (f_k \text{ \& } f_m \text{ are orthogonal})$$

This is true for arbitrary distributions of  $y_i$ , even if the  $y_i$  are nonzero-mean.

In pure noise of arbitrary distribution, for fit-functions orthogonal to all others, the  $\{b_k\}$  are uncorrelated.

**The Total Sum-of-Squares (SST) in Pure Noise**

The total sum of squares is:

raw:  $SST = \mathbf{y} \cdot \mathbf{y} = \sum_{i=1}^n y_i^2$

ANOVA:  $SST = (\mathbf{y} - \bar{y})^2 = \sum_{i=1}^n (y_i - \bar{y})^2$ , where  $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$

For zero-mean gaussian noise, the raw  $SST$  (taken over an ensemble of samples) satisfies the definition of a scaled  $\chi^2$  RV with  $n$  degrees of freedom (dof), i.e.  $SST/\sigma^2 \in \chi^2_n$ . As is well-known, the ANOVA  $SST$ , by subtracting off the sample average, reduces the dof by 1, so ANOVA  $SST/\sigma^2 \in \chi^2_{n-1}$ .

### The Model Sum-of-Squares (SSA) in Pure Noise

We're now ready for the last big step: to show that in pure noise, the model sum-of-squares (SSA) has  $p$  degrees of freedom. The model can be thought of as a vector,  $\mathbf{y}_{\text{mod}} = \{y_{\text{mod},i}\}$ , and the basis functions for that vector are the fit-functions evaluated at the sample points,  $\mathbf{f}_m \equiv \{f_m(x_i)\}$ . Then:

$$\mathbf{y}_{\text{mod}} = \sum_{m=1}^p b_m \mathbf{f}_m .$$

The  $\mathbf{f}_m$  may be oblique (non-orthogonal), and of arbitrary normalization. However, for any model vector space spanned by  $\mathbf{y}_{\text{mod}}$ , there exists an orthonormal basis in which it may be written:

$$\mathbf{y}_{\text{mod}} = \sum_{m=1}^p c_m \mathbf{g}_m \text{ where } \mathbf{g}_m \equiv \text{orthonormal basis, } c_m \equiv \text{coefficients in the } \mathbf{g} \text{ basis} . \quad (7.13)$$

We've shown that since the  $\mathbf{g}_m$  are orthonormal, the  $c_m$  are uncorrelated, with  $\text{var}(c_m) = \sigma^2$ . Now consider  $\mathbf{y}_{\text{mod}}^2$  written as a summation:

$$\mathbf{y}_{\text{mod}}^2 = \sum_{i=1}^n \left( \sum_{m=1}^p c_m g_m(x_i) \right)^2 .$$

Since the  $g_m$  are orthogonal, all the cross terms in the square are zero. Then reversing the order of summation gives:

$$\mathbf{y}_{\text{mod}}^2 = \sum_{m=1}^p \sum_{i=1}^n (c_m g_m(x_i))^2 = \sum_{m=1}^p c_m^2 \underbrace{\sum_{i=1}^n (g_m(x_i))^2}_1 = \sum_{m=1}^p c_m^2 . \quad (7.14)$$

Therefore,  $\mathbf{y}_{\text{mod}}^2$  is the sum of  $p$  uncorrelated RVs (the  $c_m^2$ ). Using the general formula for the average of the square of an RV (7.2):

$$\langle c_m^2 \rangle = \langle c_m \rangle^2 + \text{var}(c_m) = \langle c_m \rangle^2 + \sigma^2 \quad \Rightarrow \quad \langle \mathbf{y}_{\text{mod}}^2 \rangle = \left( \sum_{m=1}^p \langle c_m \rangle^2 \right) + p\sigma^2 .$$

This is true for any distribution of noise, even non-zero-mean. In general, there is no simple formula for  $\text{var}(\mathbf{y}_{\text{mod}}^2)$ .

If the noise is zero-mean, then each  $\langle c_m \rangle = 0$ , and the above reduces to:

$$\langle \mathbf{y}_{\text{mod}}^2 \rangle = p\sigma^2 \quad (\text{zero-mean noise}) .$$

If the noise is zero-mean gaussian, then the  $c_m$  are zero-mean uncorrelated joint-gaussian RVs. This is a well-known condition for independence [ref ??], so the  $c_m$  are independent, gaussian, with variance  $\sigma^2$ . Then (7.14) tells us that, by definition,  $\mathbf{y}_{\text{mod}}^2$  is a scaled chi-squared RV with  $p$  degrees of freedom:

$$\text{(raw)} \quad \frac{\mathbf{y}_{\text{mod}}^2}{\sigma^2} \equiv \frac{SSA}{\sigma^2} \in \chi_p^2 \quad (\text{zero-mean gaussian noise}) .$$

We developed this result using the properties of the orthonormal basis, but our model  $\mathbf{y}_{\text{mod}}$ , and therefore  $\mathbf{y}_{\text{mod}}^2$ , are identical in *any* basis. Therefore, the result holds for any  $p$  fit-functions that span the same model space, even if they are oblique (i.e. overlapping) and not normalized.

For the ANOVA SSQ identity, a similar analysis shows that the constraint of  $\bar{y}$  removes one degree of freedom from SSA, and therefore, for zero-mean noise:

$$\langle (\mathbf{y}_{\text{mod}} - \bar{y})^2 \rangle = (p-1)\sigma^2 \quad (\text{zero-mean noise}) .$$

For zero-mean gaussian noise, then:

$$\frac{(\mathbf{y}_{\text{mod}} - \bar{y})^2}{\sigma^2} \equiv \frac{SSA}{\sigma^2} \in \chi^2_{p-1} \quad (\text{ANOVA SSQ, zero-mean gaussian noise}).$$

If instead of pure noise, we have a signal that correlates to some extent with the model, then  $(\mathbf{y}_{\text{mod}} - \bar{y})^2$  will be bigger, on average, than  $(p - 1)\sigma^2$ . That is, the model will explain some of the variation in the data, and therefore the model sum-of-squares will (on average) be bigger than just the noise (even non-gaussian noise). :

$$\langle (\mathbf{y}_{\text{mod}} - \bar{y})^2 \rangle \equiv \langle SSA \rangle > (p-1)\sigma^2 \quad (\text{signal + zero-mean noise}).$$

### The Residual Sum-of-Squares (SSE) in Pure Noise

We determine the distribution of *SSE* in pure noise from the following:

- For least-squares linear fits:  $SST = SSA + SSE$ .
- From our analysis so far, in pure gaussian zero-mean noise:  
 $SST / \sigma^2 \in \chi^2_{n-1}$ ,  $SSA / \sigma^2 \in \chi^2_{p-1}$ .
- From the definition of  $\chi^2_v$ , the sum of independent  $\chi^2$  RVs is another  $\chi^2$  RV, and the dof add.

These are sufficient to conclude that  $SSE/\sigma^2$  must be  $\chi^2_{n-p}$ , and must be independent of *SSA*. [I'd like to show this separately from first principles??]:

$$SSE / \sigma^2 \in \chi^2_{n-p} \quad (\text{for pure gaussian zero-mean noise}).$$

### The F-test: The Decider for Zero Mean Gaussian Noise

In the sections on linear fitting, our results are completely general, and we made no assumptions at all about the nature of the residuals. In the more recent results under hypothesis testing, we have made the minimum assumptions possible, to have the broadest applicability possible. However:

To do quantitative hypothesis testing,  
we must know something about the residual distribution in our data.

One common assumption is that our noise is zero-mean gaussian. Then we can quantitatively test if our data are pure noise, and establish a level of confidence (e.g., 98%) in our conclusion. Later, we show how to use simulations to remove the restriction to gaussian noise, and establish confidence bounds for any distribution of residuals.

For zero-mean pure gaussian noise only: we have shown that the *raw*  $(SSA/\sigma^2) \in \chi^2_p$ . We have also indicated that for ANOVA:

$$\left. \begin{aligned} SST / \sigma^2 &\in \chi^2_{n-1} \\ SSA / \sigma^2 &\in \chi^2_{p-1} \\ SSE / \sigma^2 &\in \chi^2_{n-p} \end{aligned} \right\} \Rightarrow \begin{aligned} \sigma^2 &= \langle SST / (n-1) \rangle \\ \sigma^2 &= \langle SSA / (p-1) \rangle \\ \sigma^2 &= \langle SSE / (n-p) \rangle \end{aligned}$$

Furthermore, *SSA* and *SSE* are statistically independent, and each provides an estimate of the noise variance  $\sigma^2$ .

[Note that the *difference* between two *independent*  $\chi^2$  RVs has no simple distribution. This means that *SST* is *correlated* with *SSA* in just the right way so that  $(SST - SSA) = SSE$  is  $\sigma^2\chi^2$  distributed with  $p - 1$  dof; similarly *SST* is correlated with *SSE* such that  $(SST - SSE) = SSA \in \sigma^2\chi^2$  with  $n - p$  dof.]

We can take the ratio of the two independent estimates of  $\sigma^2$ , and in pure noise, we should get something close to 1:

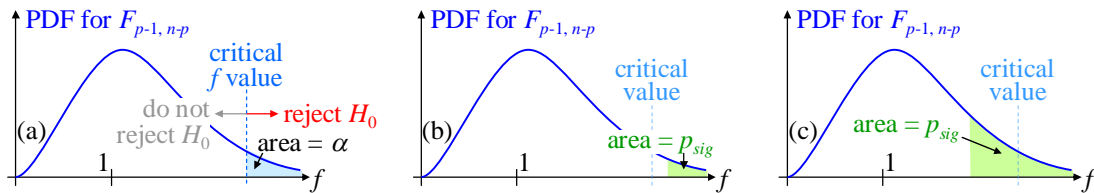
$$f \equiv \frac{SSA/(p-1)}{SSE/(n-p)} \approx 1 \quad (\text{in pure noise}).$$

Of course, this ratio is itself a random variable, and will vary from sample set to sample set. The distribution of this RV is the **Fisher–Snedecor F-distribution**. It is the distribution of the ratio of two reduced- $\chi^2$  parameters. Its closed-form is not important, but its general properties are. First, the distribution depends on both the numerator and denominator degrees of freedom, so  $F$  is a two-parameter family of distributions, denoted here as  $F(\text{dof num}, \text{dof denom}; f)$ . (Some references use  $F$  to denote the CDF, rather than PDF.)

If our test value  $f$  is much larger than 1, we might suspect that  $H_0$  is false: we actually have a signal. We establish this quantitatively with a one-sided  $F$ -test, at the  $\alpha$  level of significance (Figure 7.10):

$$f > \text{critical\_value}[F_{p-1, n-p}; \alpha] \Rightarrow \text{reject } H_0.$$

If  $f >$  critical value, then it is unlikely to be the result of pure noise. We therefore reject  $H_0$  at the  $\alpha$  level of significance, or equivalently, at the  $(1 - \alpha)$  level of confidence.



**Figure 7.10** One-sided F-test for the null hypothesis,  $H_0$ . (a) Critical  $f$  value; (b) statistically significant result; (c) not statistically significant result.

### Coefficient of Determination and Correlation Coefficient

We hear a lot about the correlation coefficient,  $\rho$ , but it's actually fairly useless. However, its square ( $\rho^2$ ) is the **coefficient of determination**, and is much more meaningful: it tells us the fraction of measured variation “explained” by a straight-line fit to the predictor  $f_1(x)$ . This is sometimes useful as a measure of the effectiveness of the model.  $\rho^2$  is a particular use of the linear regression we have already studied.

First consider a (possibly infinite) *population* of  $(x, y)$  pairs. Typically,  $x$  is an independent variable, and  $y$  is a measured dependent variable. (We mention a slightly different use for  $\rho^2$  at the end.) We often think of the fit function as  $f_1(x) = x$  (which we use as our example), but as with all linear regression, the fit-function is *arbitrary*. Recall the sum-of-squares definitions of  $SST$ ,  $SSA$ , and  $SSE$  (7.11) We *define* the coefficient of determination in linear-fit terms, as the fraction of  $SST$  that is determined by the best-fit model. This is also the ratio of population variances of a least-squares fit:

$$\rho^2 \equiv \frac{SSA}{SST} \equiv \frac{\text{var}(y_{\text{mod}})}{\text{var}(y)} \quad \text{where } y_{\text{mod}}(x) \equiv b_0 + b_1 x \quad (\text{population}).$$

Note that for the variance of the straight line  $y_{\text{mod}}$  to be defined, the domain of  $x$  must be finite, i.e.  $x$  must have finite lower and upper bounds. For experimental data, this requirement is necessarily satisfied.

Now consider a *sample* of  $n$   $(x, y)$  pairs. It is a straightforward application of our linear regression principles to estimate  $\rho^2$ . We call the estimate the **sample coefficient of determination**,  $r^2$ , and define it analogously to the population parameter:

$$r^2 \equiv \frac{SSA}{SST} \quad (\text{sample coefficient of determination}) \quad [\text{Myers 1986 2.20 p28}]$$

$$\text{where } SSA \equiv \sum_{i=1}^n (y_{\text{mod},i} - \bar{y})^2, \quad SST \equiv \sum_{i=1}^n (y_i - \bar{y})^2.$$

Note that the number of fit parameters is  $p = 2$  ( $b_0$  and  $b_1$ ). Therefore  $SSA$  has  $p - 1 = 1$  degree of freedom (dof), and  $SST$  has  $n - 1$  dof.

[The sample correlation coefficient is just  $r$  (with a sign given below):

$$|r| \equiv \sqrt{r^2} \equiv \sqrt{SSA/SST} \quad (\text{sample correlation coefficient}).$$

For multiple regression (i.e., with multiple “predictors”, where  $p \geq 3$  but one is the constant  $b_0$ ), we define  $r$  always  $\geq 0$ . In the case of single regression to one predictor (call it  $x$ ,  $p = 2$  but still one is the constant  $b_0$ ),  $r > 0$  if  $y$  increases with the predictor  $x$ , and  $r < 0$  if  $y$  decreases with increasing  $x$ .

For simplicity, we start with a sample where  $\bar{x} = \bar{y} = 0$ . At the end, we easily extend the result to the general case where either or both averages are nonzero. If  $\bar{x} = 0$ , then  $f_1$  is orthogonal to the constant  $b_0$ , and we can find  $b_1$  by a simple correlation, including normalization of  $f_1$  (see linear regression, earlier):

$$b_1 = \frac{\sum_{i=1}^n f_1(x_i) y_i}{\sum_{i=1}^n f_1(x_i)^2} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} = \frac{n \langle xy \rangle}{n \sigma_x^2} = \frac{\langle xy \rangle}{\sigma_x^2}.$$

With  $b_1$  now known, we can compute  $SSA$  (recalling that  $\bar{y} = b_0 = 0$  for now):

$$SSA = \sum_{i=1}^n (y_{\text{mod},i} - \bar{y})^2 = \sum_{i=1}^n (b_1 x_i)^2 = b_1^2 \sum_{i=1}^n x_i^2 = \left( \frac{\langle xy \rangle}{\sigma_x^2} \right)^2 n \sigma_x^2 = n \frac{\langle xy \rangle^2}{\sigma_x^2}.$$

$SST$  is, with  $\bar{y} = 0$ :

$$SST = \sum_{i=1}^n y_i^2 = n \sigma_y^2.$$

Then:

$$r^2 \equiv \frac{SSA}{SST} = \frac{n \langle xy \rangle^2 / \sigma_x^2}{n \sigma_y^2} = \frac{\langle xy \rangle^2}{\sigma_x^2 \sigma_y^2} \quad \Rightarrow \quad r \equiv \frac{\langle xy \rangle}{\sigma_x \sigma_y} = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\left( \sum_{i=1}^n x_i^2 \right) \left( \sum_{i=1}^n y_i^2 \right)}}.$$

Since  $\bar{y}$  was known exactly, and not estimated from the sample,  $SST$  has  $n$  dof.

To generalize to nonzero  $\bar{x}$  and  $\bar{y}$ , we note that we can transform  $x \rightarrow x - \bar{x}$ , and  $y \rightarrow y - \bar{y}$ . These are simple shifts in  $(x, y)$  position, and have no effect on the fit line slope or the residuals. These new random variables are zero-mean, so our simplified derivation applies, with one small change:  $\bar{y}$  is estimated from the sample, so that removes 1 dof from  $SST$ :  $SST$  has  $n - 1$  dof. Then:

$$r^2 \equiv \frac{SSA}{SST} = \frac{\langle (x-\bar{x})(y-\bar{y}) \rangle^2}{\sigma_x^2 \sigma_y^2} \quad \text{where} \quad \langle (x-\bar{x})(y-\bar{y}) \rangle^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

$$\sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \sigma_y^2 \text{ similar.}$$

Note that another common notation is:

$$r^2 \equiv \frac{SSA}{SST} = \frac{S_{xy}^2}{S_{xx} S_{yy}} \quad \text{where} \quad S_{xy} \equiv \langle (x-\bar{x})(y-\bar{y}) \rangle, \quad S_{xx} \equiv \sigma_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_{yy} \text{ similar.}$$

**Distribution of  $r^2$ :** Similarly to what we have seen with testing other fit parameters, to test the hypothesis that  $r^2 > 0$ , we first consider the distribution of  $r^2$  in pure noise. For pure zero-mean gaussian noise,  $r^2$  follows a beta distribution with 1 and  $n-1$  degrees of freedom (dof) [ref ??]. We can use the usual one-sided test at the  $\alpha$  significance threshold: if

$$p_{sig} = 1 - \text{cdf}_{beta}(r^2) > \text{critical\_value}[beta(1, n-1); \alpha] \quad (\text{gaussian}), \quad (7.15)$$

then we reject the null hypothesis  $H_0$ , and accept that  $r^2$  is probably  $> 0$ , at the  $p_{sig}$  level of significance.

However:

The beta distribution is difficult to use, since it crams up near 1, and many computer implementations are unstable in the critical region where we need it most. Instead, we can use an equivalent  $F$  test, which is easy to interpret, and numerically stable.

Again applying our results from linear regression, we recall that:

$$SST = SSA + SSE \quad \Rightarrow \quad f \equiv \frac{SSA}{SSE} = \frac{\rho}{1-\rho} \in F_{1, n-1}.$$

Then for pure noise,  $f \approx 1$ . If  $f \gg 1$ , then  $r^2$  is probably  $> 0$ , with significance given by the standard 1-sided  $F$  test ( $\alpha$  is our threshold for rejecting  $H_0$ ):

$$p_{sig} = 1 - \text{cdf}_F(f) > \text{critical\_value}[F_{1, n-1}; \alpha].$$

Note that the significance  $p_{sig}$  here is *identical* to the significance from the beta function (7.15), but using the  $F$  distribution is usually an easier way to compute it.

**Alternative interpretation of  $x$  and  $y$ :** There is another way that  $\rho^2$  can be used, depending on the nature of your data. Instead of  $x$  being an independent variable and  $y$  being corresponding measured values, it may be that *both*  $x$  and  $y$  are RVs, with some interdependence. Then, much like  $\bar{y}$  is a population parameter of a single random variable  $y$ ,  $\rho^2$  is a population parameter of two *dependent* random variables,  $x$  and  $y$ , and their joint density function. Either way, we *define* the coefficient of determination in linear-fit terms, as a ratio of population variances of a least-squares fit of  $y$  to  $x$ . (We ignore here the question of the dof in  $\sigma_x^2$ .)

## Uncertainty Weighted Data

When taking data, our measurements often have varying uncertainty: some measurements are “better” than others. We can still find an average, but what is the *best* average, and what is its uncertainty? These questions extend to almost all of the statistics we’ve covered so far: sample average and variance, fitting, etc. In general, if you have a set of estimates of a parameter, but each estimate has a *different* uncertainty, how do you combine the estimates for the most reliable estimate of the parameter? Intuitively, estimates with smaller uncertainty should be given more weight than estimates with larger uncertainty. But exactly how much?

Each topic in this section assumes you thoroughly understand the unweighted case before delving into the weighted case.

Throughout this section, we consider data triples of the form  $(x_i, y_i, u_i)$ , where  $x_i$  are the independent variables,  $y_i$  are the measured variables, and  $u_i$  are the  $1\sigma$  uncertainties of each measurement. We *define* the uncertainty as variations that *cannot* be modeled in detail, though their PDF or other statistics may be known.

Formulas with uncertainties are *not* simply the unweighted formulas with weights thrown into the “obvious” places.

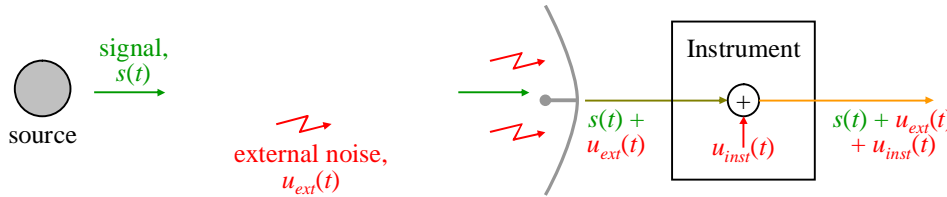
Examples of the failure of the “obvious” adjustments to formulas for uncertainty-weighted data are the unbiased estimate of a population  $\sigma^2$  from a sample (detailed below), and the Lomb-Scargle detection parameter.

### Be Sure of Your Uncertainty

We must carefully define what we mean by “uncertainty”  $u_i$ . Figure 7.11 depicts a typical measurement, with two separate sources of noise: external ( $u_{ext}$ ), and instrumental ( $u_{inst}$ ). The model experiment could be an astronomical one, spread over millions of light-years, or it could be a table top experiment. The external noise might be background radiation, CMB, thermal noise, etc. The instrument noise is the inevitable variation in any measurement system. One can often calibrate the instrument, and determine  $u_{inst}$ . Sometimes, one can measure  $u_{ext}$ , as well. However, for purposes of this chapter, we *define* our **uncertainty**  $u_i$  as:

$$u_i \equiv \text{all of the noise outside of the desired signal, } s(t).$$

Our results depend on this.



**Figure 7.11** A typical measurement includes two sources of noise.

### Average of Uncertainty Weighted Data

We give the formula for the uncertainty-weighted average of a sample, and the uncertainty of that average. Consider a sample of  $n$  uncertainty weighted measurements, say  $(t_i, y_i, u_i)$ , where  $t_i$  is time,  $y_i$  is the measurement, and  $u_i$  is the  $1\sigma$  uncertainty in  $y_i$ . How should we best estimate the population average from this sample? If we *assume* the estimator is a weighted average (as opposed to RMS or something else), we now show that we should weight each  $y_i$  by  $u_i^{-2}$ . The general formula for a weighted average is:

$$\bar{y} = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} . \tag{7.16}$$

The variance (over an ensemble of samples) of this weighted average, where the weights are constants, is (recall that uncorrelated variances add):

$$\text{var}(\bar{y}) = \frac{\sum_{i=1}^n w_i^2 u_i^2}{\left(\sum_{i=1}^n w_i\right)^2} . \tag{7.17}$$



Note that because of the normalization factor in the denominator, both  $\bar{y}$  and its variance are independent of any multiplicative constant in the weights (scale invariance): e.g., doubling all the weights has no effect on either. However, we want to choose a set of weights to give  $\bar{y}$  the minimum variance possible. Therefore the derivative of the above variance with respect to any weight,  $w_k$ , is zero. Using the quotient rule for derivatives:

$$\frac{\partial \text{var}(\bar{y})}{\partial w_k} = \frac{\left( \sum_{i=1}^n w_i \right)^2 \cancel{w_k} u_k^2 - \left( \sum_{i=1}^n w_i^2 u_i^2 \right) \cancel{\left( \sum_{i=1}^n w_i \right)}}{\left( \sum_{i=1}^n w_i \right)^2} = 0 \quad \left( dUV = \frac{VdU - UdV}{V^2} \right)$$

$$w_k = \frac{\left( \sum_{i=1}^n w_i^2 u_i^2 \right) \left( \sum_{i=1}^n w_i \right)}{\left( \sum_{i=1}^n w_i \right)^2 u_k^2}.$$

Since the weights are scale invariant, the only dependence that matters is that  $w_k \propto u_k^{-2}$ . Therefore, we take the simplest form, and define:

$$w_i \equiv u_i^{-2} \quad (\text{raw weights}).$$

For a least-squares estimate of the population average, we weight each measurement by the inverse of the uncertainty squared (inverse of the *measurement variance*).

As expected, large uncertainty points are weighted less than small uncertainty points. Our derivation applies to *any* measurement error distribution; in particular, errors *need not* be gaussian. The least-squares weighted average is well-known [Myers 1986 p171t]. [Note that we have *not* proved that a weighted average is necessarily the optimum form for an average, but it is. (I suspect this can be proved with calculus of variations, but I've never seen it done.)]

Given these optimum weights, we can now write the uncertainty of  $\bar{y}$  more succinctly. For convenience, we define:

$$W \equiv \sum_{i=1}^n u_i^{-2}, \quad V_1 \equiv \sum_{i=1}^n w_i \quad (\text{a normalization factor}), \quad V_2 \equiv \sum_{i=1}^n w_i^2.$$

Note that  $W$  is defined to be independent of weight scaling,  $V_1$  scales with the weights, and  $V_2$  scales with the square of the weights. Then from eq. (7.17), the variance of  $\bar{y}$  is:

$$\text{var}(\bar{y}) = \frac{\sum_{i=1}^n w_i^2 u_i^2}{V_1^2} \quad \text{Use: } u_i^2 = w_i^{-1}: \quad \text{var}(\bar{y}) = \frac{\sum_{i=1}^n w_i}{V_1^2} = \frac{V_1}{V_1^2} = \frac{1}{V_1}. \quad (7.18)$$

This variance must be scale invariant, but  $V_1$  scales. We chose a scale when we used  $u_i^2 = w_i^{-1}$ , for which  $V_1 = W$ .  $W$  is scale invariant, therefore the scale invariant result is:

$$\text{var}(\bar{y}) = \frac{1}{W}, \quad \text{and} \quad U(\bar{y}) \equiv \text{dev}(\bar{y}) = \sqrt{\text{var}(\bar{y})} = \frac{1}{\sqrt{W}}.$$

The weights,  $w_i$ , as we have defined them, have units of [measurement]<sup>-2</sup>.

Note that the weighted uncertainty of  $\bar{y}$  reduces to the well-known unweighted uncertainty when all the uncertainties are equal, say  $u$ :

$$\text{var}(\bar{y}) = \frac{1}{W} = \left( \sum_{i=1}^n u^{-2} \right)^{-1} = \frac{u^2}{n} \quad \Rightarrow \quad U(\bar{y}) = \frac{u}{\sqrt{n}}.$$

### Variance and Standard Deviation of Uncertainty Weighted Data

**Handy numerical identity:** When computing unweighted standard deviations, we simplify the calculation using the handy identity:

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 \quad \text{or} \quad \sum_{i=1}^n y_i^2 - \frac{(\sum y_i)^2}{n}.$$

What is the equivalent identity for *weighted* sums of squared deviations? We derive it here:

$$\begin{aligned} \sum_{i=1}^n w_i (y_i - \bar{y})^2 &= \sum_{i=1}^n w_i (y_i^2 - 2y_i\bar{y} + \bar{y}^2) && \text{Use: } \sum_{i=1}^n w_i y_i = V_1 \bar{y} \\ &= \sum w_i y_i^2 - 2V_1 \bar{y}^2 + V_1 \bar{y}^2 && (7.19) \\ &= \sum w_i y_i^2 - V_1 \bar{y}^2 \quad \text{or} \quad \sum w_i y_i^2 - \frac{(\sum w_i y_i)^2}{V_1}. \end{aligned}$$

We note a general pattern that in going from an unweighted formula to the equivalent weighted formula: the number  $n$  is often replaced by the number  $V_1$ , and all the summations include the weights.

**Weighted sample variance:** We now find an unbiased weighted sample variance; unbiased means that over many samples (sets of individual values), the sample variance averages to the population variance. In other words, it is an unbiased estimate of the population variance. We first state the result:

$$s^2 = \frac{\sum_{i=1}^n w_i (y_i - \bar{y})^2}{V_1 - V_2 / V_1}.$$

We prove below that this is an unbiased estimator.

Many references give incorrect formulas for the weighted sample variance;  
in particular, it is *not* just  $(1/V_1) \sum w_i (y_i - \bar{y})^2$ .

Because the weights are arbitrary,  $s^2$  does *not* exactly follow a scaled  $\chi^2$  distribution. However, if the uncertainties are not too disparate, we can approximate  $s^2$  as being  $\chi^2$  with  $(n-1)$  dof [ref??].

For computer code, we often use the weighted sum-of-squared deviations identity (7.19) to simplify the calculation:

$$s^2 = \frac{\sum_{i=1}^n w_i (y_i - \bar{y})^2}{V_1 - V_2 / V_1} = \frac{\sum_{i=1}^n w_i y_i^2 - \left( \sum_i w_i y_i \right)^2 / V_1}{V_1 - V_2 / V_1} \quad \text{or} \quad \frac{V_1 \sum_{i=1}^n w_i y_i^2 - \left( \sum_i w_i y_i \right)^2}{V_1^2 - V_2}.$$

We now prove that over many sample sets, the statistic  $s^2$  averages to the true population  $\sigma^2$ . (We use our statistical algebra.) Without loss of generality, we take the population average to be zero, because we

can always shift a random variable by a constant amount to make its (weighted) average zero, without affecting its variance. Then the population variance becomes:

$$\sigma^2 = \langle Y^2 \rangle - \langle Y \rangle^2 \quad \rightarrow \quad \sigma^2 = \langle Y^2 \rangle.$$

We start by guessing (as we did for unweighted data) that the *weighted* average squared-deviation is an estimate for  $\sigma^2$ . For a single given sample, the simple weighted average of the squared-deviations from  $\bar{y}$  is (again using (7.19)):

$$q^2 \equiv \frac{\sum_{i=1}^n w_i (y_i - \bar{y})^2}{\sum w_i} = \frac{\sum w_i y_i^2 - V_1 \bar{y}^2}{V_1} = \frac{\sum w_i y_i^2}{V_1} - \bar{y}^2 \tag{7.20}$$

Is this unbiased? To see, we average over the ensemble of all possible sample sets (using the same weights). I.e., the weights, and therefore  $V_1$  and  $V_2$ , are constant over the ensemble average. The first term in (7.20) averages to:

$$\frac{\sum_{i=1}^n w_i y_i^2}{V_1} = \frac{\sum_{i=1}^n w_i}{V_1} \langle y_i^2 \rangle = \langle Y^2 \rangle = \sigma^2.$$

The second term in (7.20) averages to:

$$\langle \bar{y}^2 \rangle = \left\langle \left( \frac{\sum w_i y_i}{V_1} \right)^2 \right\rangle = \frac{1}{V_1^2} \left[ \left\langle \sum_{i=1}^n w_i^2 y_i^2 \right\rangle + \left\langle \sum_{i \neq j} w_i w_j y_i y_j \right\rangle \right].$$

Recall that the covariance, or equivalently the correlation coefficient, between any two independent random variables is zero. Then the last term is proportional to  $\langle y_i y_j \rangle$ , which is zero for the independent values  $y_i$  and  $y_j$ . Thus:

$$\langle \bar{y}^2 \rangle = \frac{1}{V_1^2} V_2 \langle Y^2 \rangle = \frac{V_2}{V_1^2} \sigma^2 \quad \Rightarrow \quad \langle q^2 \rangle = \sigma^2 - \frac{V_2}{V_1^2} \sigma^2 = \left( 1 - \frac{V_2}{V_1^2} \right) \sigma^2.$$

Finally, the unbiased estimate of  $\sigma^2$  simply divides out the prefactor:

$$\langle s^2 \rangle = \frac{\langle q^2 \rangle}{1 - V_2 / (V_1^2)} \quad \Rightarrow \quad s^2 \equiv \frac{\sum_{i=1}^n w_i (y_i - \bar{y})^2}{V_1 - V_2 / V_1}, \tag{7.21}$$

as above. Note that we have shown that  $s^2$  is unbiased, but we have *not* shown that  $s^2$  is the *least-squares* estimator, nor that it is the *best* (minimum variance) unbiased estimator. But it is [ref??].

Also, as always, the sample standard deviation  $s \equiv \sqrt{s^2}$  is *biased*, because the square root of an average is not the average of the square roots. Since we are concerned most often with bias in the variance, and rarely with bias in the standard deviation, we don't bother looking for an unbiased estimator for  $\sigma$ , the population standard deviation.

**Distributed of weighted  $s^2$ :** Since  $s^2$  derives from a weighted sum of squares, it is *not*  $\chi^2$  distributed, and therefore we cannot associate any degrees of freedom with it. However, for large  $n$ , and not too disparate uncertainties  $u_i$ , we can *approximate* the weighted  $s^2$  as having a  $\chi^2_{n-1}$  distribution (like the unweighted  $s^2$  does).

### Normalized weights

Some references normalize the weights so that they sum to 1, in which case they are dimensionless:

$$W \equiv \sum_{i=1}^n u_i^{-2}, \quad \text{and} \quad w_i \equiv \frac{u_i^{-2}}{W} \quad (\text{normalized, dimensionless weights}).$$

This makes  $V_1 \equiv 1$  (dimensionless), and therefore  $V_1$  does not appear in any formulas. ( $V_2$  must still be computed from the normalized weights.) Both normalizations are found in the literature, so it is helpful to be able to switch between the two.

As an example of how formulas are changed, consider a chi-squared goodness-of-fit parameter. Its form is, in both raw and normalized weights:

$$(\text{raw}) \quad \chi^2 = \sum_{i=1}^n w_i (y_i - y_{\text{mod},i})^2 \quad \rightarrow \quad \chi^2 = W \sum_{i=1}^n w_i (y_i - y_{\text{mod},i})^2 \quad (\text{normalized}).$$

Other similar modifications appear in other formulas. In general, we can say:

$$w_i^{\text{raw}} \rightarrow W w_i^{\text{norm}}, \quad V_1 \rightarrow W, \quad V_2^{\text{raw}} \rightarrow W^2 V_2^{\text{norm}}, \quad \text{and}$$

$$w_i^{\text{norm}} \rightarrow \frac{w_i^{\text{raw}}}{V_1}, \quad W \rightarrow V_1, \quad V_2^{\text{norm}} \rightarrow \frac{V_2^{\text{raw}}}{V_1^2}.$$

We use the first set of transforms to take formulas from raw to normalized, and the second set of transforms to take formulas from normalized to raw. As another example, we transform the raw formula for  $s^2$ , eq. (7.21), to normalized:

$$(\text{raw}) \quad s^2 \equiv \frac{\sum_{i=1}^n w_i (y_i - \bar{y})^2}{V_1 - V_2 / V_1} \quad \rightarrow \quad s^2 = \frac{W \sum_{i=1}^n w_i (y_i - \bar{y})^2}{W - W^2 V_2 / W} = \frac{\sum_{i=1}^n w_i (y_i - \bar{y})^2}{1 - V_2} \quad (\text{normalized}).$$

To go back (from the normalized  $s^2$  to raw), we take  $W \rightarrow V_1$  (if  $W$  were there),  $w_i \rightarrow w_i / V_1$ , and  $V_2 \rightarrow V_2 / V_1^2$ .

For now, the raw, dimensionful weights give us a handy check of units for our formulas, so we continue to use them in most places.

### Numerically Convenient Weights

It is often convenient to perform preliminary calculations by ignoring the measurement uncertainties  $u_i$ , and using unweighted formulas. We might even do such estimates mentally. Later, more accurate calculations may be done which include the uncertainties. It is often convenient to compare the preliminary unweighted values with the weighted values, especially for intermediate steps in the analysis, e.g. during debugging of analysis code. However, unnormalized weights,  $w_i = u_i^{-2}$ , have arbitrary magnitudes that lead to intermediate values with no simple interpretation, and that are not directly comparable to the unweighted estimates. Therefore, it is often convenient to scale the weights so that intermediate results have the same scale as unweighted results. The unweighted case is equivalent to all weights being 1, with a sum of  $n$ . We can scale our uncertainty weights to the same sum, i.e.  $n$ , or equivalently, we scale our weights to an average of 1:

$$\sum_{i=1}^n 1 = n \quad (\text{unweighted}) \quad \Rightarrow \quad (\text{weighted}) \quad \sum_{i=1}^n w_i = n \quad \text{and therefore} \quad w_i = \frac{n}{W} u_i^{-2}.$$

With this weight scaling, “quick and dirty” calculations are easily compared to more accurate fully-weighted intermediate (debug) results.

## Transformation to Equivalent Homoskedastic Measurements

We expect that the homoskedastic case (all measurements have the same uncertainty,  $\sigma$ ) is simpler, and possibly more powerful than the heteroskedastic case (each measurement has its own uncertainty,  $u_i$ ). Furthermore, many computer regression libraries cannot handle heteroskedastic data. Fortunately, for the purpose of linear regression, there is a simple transformation from heteroskedastic measurements to an equivalent set of homoskedastic measurements. This not only provides theoretical insight, but is very useful in practice: it allows us to use many (but not all) of the homoskedastic libraries by transforming to the equivalent homoskedastic measurements, and operating on the transformed data.

To perform the transformation, we choose an arbitrary uncertainty to act as our new, equivalent *homoskedastic* uncertainty  $\sigma$ . As a convenient choice, we might choose the smallest of all the measurement uncertainties  $u_{min}$  to be our equivalent homoskedastic uncertainty  $\sigma$ , or perhaps the RMS( $u_i$ ). (Recall that  $u_i$  is defined as *all* of the measurement error, both internal and external.) Then we define a new set of equivalent “measurements”  $(x_i, y_i, u_i) \rightarrow (x'_i, y'_i, \sigma)$  according to:

$$y'_i = \frac{\sigma}{u_i} y_i, \quad x'_{mi} = x_{mi} \frac{\sigma}{u_i}.$$

We can now use all of the homoskedastic procedures and calculations for linear regression on the new, equivalent “measurements.” Note that we have scaled both the predictors  $x_{mi}$ , and the measurements  $y_i$ , by the ratio of our chosen  $\sigma$  to the original uncertainty  $u_i$ . Measurements with smaller uncertainties than  $\sigma$  get scaled “up” (bigger), and measurements with larger uncertainties than  $\sigma$  get scaled “down” (smaller).

If the original noise added into each sample was independent (as we usually assume), then multiplying the  $y_i$  by constants also yields independent noise samples, so the property of independent noise is preserved in the transformation.

Figure 7.12 shows an example transformation graphically, and helps us understand why it works. Consider 3 heteroskedastic measurements:

$$(1.0, 0.5, 0.1), \quad (1.6, 0.8, 0.2), \quad (2.0, 1.0, 0.3) \quad (\text{original measurements}).$$

We choose our worst uncertainty, 0.3, as our equivalent homoskedastic  $\sigma$ . Then our equivalent measurements become:

$$(3.0, 1.5, 0.3), \quad (2.4, 1.2, 0.3), \quad (2.0, 1.0, 0.3) \quad (\text{equivalent measurements}).$$

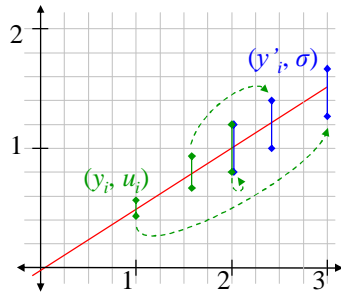
Figure 7.12 illustrates that an uncertainty of 0.3 at  $x'_1 = 3.0$  is equivalent to an uncertainty of 0.1 at  $x_1 = 1.0$ , because the  $x'$  point “tugs on” the slope of the line with the same contribution to  $\chi^2$ , the square of  $(y_{\text{mod},i} - y_i)/u_i$ . In terms of sums of squares, the transformation equates *every* term of the sum:

$$\frac{y_{\text{mod}}(x_i) - y_i}{u_i} = \frac{y_{\text{mod}}(x'_i) - y'_i}{\sigma}, \quad \forall i \quad \Rightarrow \quad \sum_{i=1}^n \left( \frac{y_{\text{mod}}(x_i) - y_i}{u_i} \right)^2 = \sum_{i=1}^n \left( \frac{y_{\text{mod}}(x'_i) - y'_i}{\sigma} \right)^2.$$

The transformation coefficients are dimensionless, so the units of the transformed quantities are the same as the originals. Note that:

The regression coefficients  $b_k$ , and their covariances, are *unchanged* by the transformation to equivalent homoskedastic measurements, but the model *values*  $y'_{\text{mod},i} = y_{\text{mod}}(x'_i)$  change because the predictors  $x'_i$  are transformed from the original  $x_i$ .

Equivalently, the *predictions* of the transformed model are *different* than the predictions of the original model. The uncertainties in the  $b_m$  are given by the standard homoskedastic formulas with  $\sigma$  as the measurement uncertainties, and the covariance matrix  $\text{var}(\mathbf{b})$  is also preserved by the transformation.. These considerations show that *SST*, *SSA*, and *SSE* are *not* preserved in the transformation.



**Figure 7.12** The model  $y_{\text{mod}}$  vs. the original and the equivalent homoskedastic measurements.

In matrix form, the transformation is:

$$\mathbf{T} \equiv \begin{bmatrix} \frac{\sigma}{u_1} & & & \\ & \frac{\sigma}{u_2} & & \\ & & \ddots & \\ & & & \frac{\sigma}{u_n} \end{bmatrix}, \quad \mathbf{y}' = \mathbf{T}\mathbf{y}, \quad \mathbf{x}' = \mathbf{T} \mathbf{x}.$$

$\begin{matrix} \underline{\mathbf{T}} & \underline{\mathbf{y}'} & \underline{\mathbf{x}'} \\ n \times n & n \times p & n \times p \end{matrix}$

The transformed data are *only* equivalent for the purpose of linear regression, and its associated capabilities, such as prediction, correlation coefficients, etc.

To illustrate this, the standard sample average is a linear fit to a constant function  $f_0(t) = 1$ . Therefore, the weighted sample average *is* given by the unweighted average of the transformed measurements. Proof TBS??. Note that the transformed function,  $f'_0(t)$  is *not* constant.

In contrast, note that the heteroskedastic population variance estimate (eq. (7.21)),

$$s^2 \equiv \frac{\sum_{i=1}^n w_i (y_i - \bar{y})^2}{V_1 - V_2 / V_1}$$

is *not* a linear fit. That's why it requires this odd-looking formula, and is *not* given by the common homoskedastic variance estimate,  $s^2 = \sum (y_i - \bar{y})^2 / (n - 1)$ , applied to the transformed data.

As another example, the standard Lomb-Scargle algorithm doesn't work on transformed data. Although it is *essentially* a simultaneous fit to a cosine and a sine, it relies on a nonlinear computation of the orthogonalizing time offset,  $\tau$ . This fails for the transformed data.

**Orthogonality is preserved:** If two predictors are orthogonal w.r.t. the weights, then the transformed predictors are also orthogonal:

$$\sum_{i=1}^n w_i x_{ki} x_{mi} = 0 \Rightarrow \sum_{i=1}^n x'_{ki} x'_{mi} = \sigma^2 \sum_{i=1}^n \frac{x'_{ki}}{u_i} \frac{x'_{mi}}{u_i} = \sigma^2 \underbrace{\sum_{i=1}^n w_i x_{ki} x_{mi}}_0 = 0.$$

## Linear Regression with Individual Uncertainties

We have seen that for data with *constant* uncertainties, we fit it to a model using the criterion of least-squared residual. If instead we have individual uncertainties  $(y_i, u_i)$ , we commonly use least-chi-squared residuals. That is, we fit the model coefficients  $(b_k)$  to minimize:

$$SSE \equiv \chi^2 \equiv \sum_{i=1}^n \frac{\varepsilon_i^2}{u_i^2} \quad \text{where} \quad \varepsilon_i \equiv \text{residual} \equiv y_i - y_{\text{mod},i} .$$

For gaussian residuals, least-chi-squared fits yields maximum likelihood fit coefficients. For non-gaussian residuals, least-chi-squared is as good a criterion as any.

However, there are many statistical formulas that need updating for uncertainty-weighted data. Often, we need an exact closed-form formula for a weighted-data statistical parameter. For example, computing an iterative approximate fit to data can be prohibitively slow, but a closed-form formula may be acceptable (e.g., periodograms). Finding such exact formulas in the literature is surprisingly hard.

Even though we've described the transformation to linear equivalent measurements, it is often more convenient to compute results directly from the original measurements and uncertainties.

We discuss and analyze some direct weighted-regression computations here. As in the earlier unweighted analysis, we clearly identify the scope of applicability for each formula. And as always, understanding the methods of analyzing and deriving these statistics is essential to developing your own methods for processing new situations.

This section assumes a thorough understanding of the similar unweighted sections. Many of our derivations follow the unweighted ones, but may be briefer here.

The first step of linear regression with individual uncertainties is summarized in [Bev p117-118], oddly in the chapter "Least-Squares Fit to a Polynomial," even though it applies to *all* fit functions (*not* just polynomials). We summarize here the results. The linear model is the same as the unweighted case: given  $p$  functions we wish to fit to  $n$  data points, the simplified model is:

$$y_{\text{mod}}(x) = \sum_{m=1}^p b_m f_m(x) = b_1 f_1(x) + b_2 f_2(x) + \dots + b_p f_p(x) \quad [\text{Bev 7.3 p117}] .$$

Each measurement is a triple of independent variable, dependent variable, and measurement uncertainty,  $(x_i, y_i, u_i)$ . As before, the predictors do not have to be functions of an independent variable (and in ANOVA, they are not); we use such functions only to simplify the presentation. We find the  $b_k$  by minimizing the  $\chi^2$  parameter:

$$SSE \equiv \chi^2 = \sum_{i=1}^n \frac{(y(x_i) - y_{\text{mod}}(x_i))^2}{u_i^2} = \sum_{i=1}^n \frac{\left( y(x_i) - \sum_{m=1}^p b_m f_m(x_i) \right)^2}{u_i^2} \quad [\text{Bev 7.5 p117}] .$$

For each  $k$  from 1 to  $p$ , we set the partial derivative,  $\partial \chi^2 / \partial b_k = 0$ , to get a set of simultaneous linear equations in the  $b_k$ :

$$\frac{\partial \chi^2}{\partial b_k} = 0 = \sum_{i=1}^n 2 \frac{\left( y_i - \sum_{m=1}^p b_m f_m(x_i) \right) (-f_k(x_i))}{u_i^2}, \quad k = 1, 2, \dots, p .$$

Dividing out the  $-2$ , and simplifying:

$$0 = \sum_{i=1}^n \frac{\left( y_i - \sum_{m=1}^p b_m f_m(x_i) \right) f_k(x_i)}{u_i^2}, \quad k = 1, 2, \dots, p .$$

Moving the constants to the LHS, we get a linear system of equations in the sought-after  $b_k$ :

$$\sum_{i=1}^n y_i \frac{f_k(x_i)}{u_i^2} = \sum_{i=1}^n \sum_{m=1}^p \frac{b_m f_m(x_i)}{u_i^2} f_k(x_i) = \sum_{m=1}^p b_m \sum_{i=1}^n \frac{f_m(x_i)}{u_i^2} f_k(x_i) \quad k = 1, 2, \dots, p.$$

### Linear Regression With Uncertainties and the Sum-of-Squares Identity

As with unweighted data, the weighted sum-of-squares (SSQ) identity is the crucial underpinning of **weighted linear regression** (aka “generalized linear regression”). For simplicity, we start with fitting to a single function, called  $f_k(x)$  (for generality). Before considering uncertainties, recall our *unweighted* sum-of-squares identity in vector form:

$$\begin{aligned} \text{(raw) } SST = SSA + SSE : \quad \mathbf{y}^2 = \mathbf{y}_{\text{mod}}^2 + \boldsymbol{\varepsilon}^2 \quad \text{where} \quad \boldsymbol{\varepsilon} \equiv \text{residual vector, } \mathbf{y}^2 \equiv \mathbf{y} \cdot \mathbf{y}, \text{ etc.} \\ \mathbf{y}_{\text{mod}} \equiv b_k \mathbf{f}_k + \boldsymbol{\varepsilon}. \end{aligned} \tag{7.22}$$

Recall that the dot products are real numbers. Also, by construction,  $\boldsymbol{\varepsilon}$  is orthogonal to  $\mathbf{f}_k$ ,  $\boldsymbol{\varepsilon} \cdot \mathbf{f}_k = 0$ , and the SSQ identity hinges on this.

We derive the weighted theory almost identically to the unweighted case. All of our vectors remain the same as before, and we need only redefine our dot product. The weighted dot-product weights each term in the sum by  $w_i$ :

$$\mathbf{a} \cdot \mathbf{b} \equiv \sum_{i=1}^n w_i a_i b_i, \quad w_i \propto u_i^{-2}, \quad \mathbf{a}^2 \equiv \mathbf{a} \cdot \mathbf{a}.$$

Such generalized inner products are common in mathematics and science. They retain all the familiar, useful properties; in particular, they are bilinear, and in this case, commutative. Then the weighted SSQ identity has exactly the same form as the unweighted case:

$$\text{(raw) } SST = SSA + SSE : \quad \mathbf{y}^2 = \mathbf{y}_{\text{mod}}^2 + \boldsymbol{\varepsilon}^2. \tag{7.23}$$

Note that  $SSE$  is the  $\chi^2$  parameter we minimize when fitting. Written explicitly as summations, the weighted SSQ identity is:

$$\text{(raw) } \underbrace{\sum_{i=1}^n w_i y_i^2}_{SST} = \underbrace{\sum_{i=1}^n w_i (b_k f_k(x_i))^2}_{SSA} + \underbrace{\sum_{i=1}^n w_i (y_i - b_k f_k(x_i))^2}_{SSE} \quad [\text{Schwa 1998, eq 4 p832}].$$

If this identity still holds in the weighted case, then most of our previous (unweighted) work remains valid. We now show that it *does* hold. We start by noting that even in the weighted case,  $\boldsymbol{\varepsilon} \cdot \mathbf{f}_k = 0$ . The proof comes from the fact that  $SSE$  is a minimum w.r.t. all the  $b_k$ :

$$\begin{aligned} \frac{\partial SSE}{\partial b_k} = 0 &= \frac{\partial}{\partial b_k} \sum_{i=1}^n w_i \varepsilon_i^2 = \sum_{i=1}^n 2w_i \varepsilon_i \frac{\partial}{\partial b_k} \varepsilon_i = 2 \sum_{i=1}^n w_i \varepsilon_i \frac{\partial}{\partial b_k} (y_i - b_k f_k(x_i)) \\ 0 &= \sum_{i=1}^n w_i \varepsilon_i f_k(x_i) \equiv \boldsymbol{\varepsilon} \cdot \mathbf{f}_k. \end{aligned}$$

Therefore, per (7.9), the *weighted* sum-of-squares identity holds.

Generalizing to  $p$  fit functions requires simply including a summation from 1 to  $p$ . This would make the sum-of-squares identity a little hard to read, so we separate out the “model” functions:



$$y_{\text{mod}}(x) \equiv \sum_{m=1}^p b_m f_m(x) \Rightarrow$$

$$\text{(raw)} \quad \sum_{i=1}^n w_i y_i^2 = \sum_{i=1}^n w_i y_{\text{mod}}(x_i)^2 + \sum_{i=1}^n w_i (y_i - y_{\text{mod}}(x_i))^2.$$

Also as before, if we include a constant  $b_0$  fit parameter, then the ANOVA SSQ identity holds:

$$\text{ANOVA:} \quad \sum_{i=1}^n w_i (y_i - \bar{y})^2 = \sum_{i=1}^n w_i (y_{\text{mod}}(x_i) - \bar{y})^2 + \sum_{i=1}^n w_i (y_i - y_{\text{mod}}(x_i))^2.$$

Recall that  $\bar{y}$  is the *weighted* average (7.16).

### Distribution of Weighted Orthogonal Fit Coefficients in Pure Noise

As in the unweighted case, in hopes of hypothesis testing, we need the distribution of the  $b_k$  in pure noise (no signal). Here again, if a fit function is orthogonal (w.r.t the weights) to all other fit functions, then its (least-chi-squared) fit coefficient is given by a simple correlation. I.e., for a given  $k$ :

$$\mathbf{f}_k \cdot \mathbf{f}_j = 0 \text{ for all } j \neq k \quad \Rightarrow \quad b_k = \frac{\mathbf{f}_k \cdot \mathbf{y}}{\mathbf{f}_k \cdot \mathbf{f}_k} \equiv \frac{\sum_{i=1}^n w_i f_k(t_i) y_i}{\sum_{i=1}^n w_i f_k(t_i)^2}.$$

For convenience, we now further restrict ourselves to a normalized (over the  $\{t_i\}$ ) fit-function, though this imposes no real restriction, since any function is easily normalized by a scale factor. Then:

$$\sum_{i=1}^n w_i f_k(t_i)^2 = 1 \quad \Rightarrow \quad b_k = \sum_{i=1}^n w_i f_k(t_i) y_i. \tag{7.24}$$

Now consider an ensemble of samples (sets) of noise, each with the same set of  $\{(t_i, u_i)\}$ , and each producing a random  $b_k$ . In other words, the  $b_k$  are RVs over the set of possible samples. We now find  $\text{var}(b_k)$  and  $\langle b_k \rangle$ . Recall that the variance of a sum (of uncorrelated RVs) is the sum of the variances, and the variance of  $k$  times an RV =  $k^2 \text{var}(\text{RV})$ . All the values of  $w_i$  and  $f_k(t_i)$  are constants, and  $\text{var}(y_i) \equiv u_i^2 = w_i^{-1}$ ; therefore taking the variance of (7.12):

$$\text{var}(b_k) = \sum_{i=1}^n w_i^2 f_k(t_i)^2 \underbrace{\text{var}(y_i)}_{w_i^{-1}} = \sum_{i=1}^n w_i f_k(t_i)^2 = 1. \tag{7.25}$$

This is different than the unweighted case, because the noise variance  $\sigma^2$  has been incorporated into the weights, and therefore into the normalization of the  $f_k$ .

In pure noise, for a normalized fit-function orthogonal to all others, using raw weights, the variance of its least-chi-squared linear fit coefficient is 1, regardless of the noise PDF.

We now find the average  $\langle b_k \rangle$ . Taking the ensemble average of (7.12):

$$\langle b_k \rangle = \left( \sum_{i=1}^n w_i f_k(x_i) \right) \underbrace{\langle y_i \rangle}_{\mu_y}.$$

Since the sum has no simple interpretation, this equation is most useful for showing that if the noise is zero-mean, then  $b_k$  is also zero-mean:  $\langle b_k \rangle = 0$ . However, if the summation happens to be zero, then even for non-zero mean noise, we again have  $\langle b_k \rangle = 0$ .

Furthermore, any weighted sum of gaussian RVs is a gaussian; therefore, if the  $y_i$  are gaussian (zero-mean or not), then  $b_k$  is also gaussian.

**Non-Correlation of Weighted Orthogonal Fit Coefficients in Pure Noise**

We now consider the correlation between two fit coefficients,  $b_k$  and  $b_m$  (again, over multiple samples (sets) of noise), when the fit-functions  $f_k$  and  $f_m$  are orthogonal to each other, and to all other fit-functions. (From the homoskedastic equivalent measurements, we already know that  $b_k$  and  $b_m$  are uncorrelated. However, for completeness, we now show this fact directly from the weighted data.) For convenience, we take  $f_k$  and  $f_m$  to be normalized:  $\mathbf{f}_k^2 = \mathbf{f}_m^2 = 1$  (recall that our dot-products are weighted).

As in the unweighted case, we derive the covariance of  $b_k$  and  $b_m$  from the bilinearity of the  $\text{cov}()$  operator. We start with the formula for a fit-coefficient of a normalized fit-function that is orthogonal to all others, (7.12), and use our algebra of statistics:

$$\text{cov}(b_k, b_m) = \text{cov}(\mathbf{f}_k \cdot \mathbf{y}, \mathbf{f}_m \cdot \mathbf{y}) = \text{cov} \left( \sum_{i=1}^n w_i f_k(x_i) y_i, \sum_{j=1}^n w_j f_m(x_j) y_j \right).$$

Again, all the  $w_i$ ,  $w_j$ ,  $f_k$ , and  $f_m$  are constants, so they can be pulled out of the  $\text{cov}()$  operator:

$$\text{cov}(b_k, b_m) = \sum_{i=1}^n \sum_{j=1}^n w_i f_k(x_i) w_j f_m(x_j) \text{cov}(y_i, y_j).$$

As always, the  $y_i$  are independent, and therefore uncorrelated. Hence, when  $i \neq j$ ,  $\text{cov}(y_i, y_j) = 0$ , so only the  $i = j$  terms survive, and the double sum collapses to a single sum:

$$\text{cov}(b_k, b_m) = \sum_{i=1}^n w_i^2 f_k(x_i) f_m(x_i) \underbrace{\text{cov}(y_i, y_i)}_{w_i^{-1}}. \tag{7.26}$$

Now  $\text{cov}(y_i, y_i) = \text{var}(y_i) = u_i^2 = w_i^{-1}$ , so:

$$\text{cov}(b_k, b_m) = \sum_{i=1}^n w_i f_k(x_i) f_m(x_i) = \mathbf{f}_k \cdot \mathbf{f}_m = 0.$$

This is true for arbitrary distributions of  $y_i$ , even if the  $y_i$  are nonzero-mean.

In pure noise of arbitrary distribution, even for *weighted* fit-functions orthogonal to all others, the  $\{b_k\}$  are uncorrelated.

**The Weighted Total Sum-of-Squares (SST) in Pure Noise**

The weighted total sum of squares is:

raw:  $SST = \mathbf{y} \cdot \mathbf{y} = \sum_{i=1}^n w_i y_i^2$

ANOVA:  $SST = (\mathbf{y} - \bar{\mathbf{y}})^2 = \sum_{i=1}^n w_i (y_i - \bar{y})^2, \quad \text{where} \quad \bar{y} \equiv \frac{1}{V_1} \sum_{i=1}^n w_i y_i.$

For gaussian noise, in contrast to the unweighted case, the weighted  $SST$  (taken over an ensemble of samples) is *not* a  $\chi^2$  RV. It is a *weighted* sum of scaled  $\chi^2_1$ , which has no general PDF. However, we can often approximate its distribution as  $\chi^2$  with  $n$  dof (raw), or  $n - 1$  dof (ANOVA), especially when  $n$  is large.

**The Weighted Model Sum-of-Squares (SSA) in Pure Noise**

Recall that the model can be thought of as a vector,  $\mathbf{y}_{\text{mod}} = \{y_{\text{mod},i}\}$ , and the basis functions for that vector are the fit-functions evaluated at the sample points,  $\mathbf{f}_m \equiv \{f_m(t_i)\}$ . Then:

$$\mathbf{y}_{\text{mod}} = \sum_{m=1}^p b_m \mathbf{f}_m .$$

The  $\mathbf{f}_m$  may be oblique (non-orthogonal), and of arbitrary normalization. However, as in the unweighted case, there exists an orthonormal basis in which  $\mathbf{y}_{\text{mod}}$  may be written (just like eq. (7.13)):

$$\mathbf{y}_{\text{mod}} = \sum_{m=1}^p c_m \mathbf{g}_m \text{ where } \mathbf{g}_m \equiv \text{orthonormal basis, } c_m \equiv \text{coefficients in the } \mathbf{g} \text{ basis .}$$

We've shown that since the  $\mathbf{g}_m$  are orthonormal, the  $c_m$  are uncorrelated, with  $\text{var}(c_m) = 1$  (using raw weights). Then (recall that the dot-products are weighted):

$$SSA \equiv \mathbf{y}_{\text{mod}}^2 = \left( \sum_{m=1}^p c_m \mathbf{g}_m \right)^2 = \sum_{l=1}^p \sum_{m=1}^p c_m \mathbf{g}_m \cdot c_l \mathbf{g}_l .$$

By orthogonality, only terms where  $l = m$  are non-zero, so the double sum collapses to a single sum where  $l = m$ . The  $\mathbf{g}_m$  are normalized, so:

$$\mathbf{y}_{\text{mod}}^2 = \sum_{m=1}^p (c_m \mathbf{g}_m)^2 = \sum_{m=1}^p c_m^2 \frac{\mathbf{g}_m^2}{1} = \sum_{m=1}^p c_m^2 . \tag{7.27}$$

Therefore,  $\mathbf{y}_{\text{mod}}^2$  is the sum of  $p$  uncorrelated RVs (the  $c_m^2$ ). We find  $SSA \equiv \mathbf{y}_{\text{mod}}^2$  using the general formula for the average of the square of an RV (7.2):

$$\langle c_m^2 \rangle = \langle c_m \rangle^2 + \text{var}(c_m) = \langle c_m \rangle^2 + 1 \Rightarrow SSA \equiv \langle \mathbf{y}_{\text{mod}}^2 \rangle = \left( \sum_{m=1}^p \langle c_m \rangle^2 \right) + p . \tag{where var}(c_m)$$

This is true for any distribution of noise, even non-zero-mean. In general, there is no simple formula for  $\text{var}(\mathbf{y}_{\text{mod}}^2)$ .

If the noise is zero-mean, then each  $\langle c_m \rangle = 0$ , and the above reduces to:

$$\langle \mathbf{y}_{\text{mod}}^2 \rangle = p \quad (\text{zero-mean noise}) .$$

If the noise is zero-mean gaussian, then the  $c_m$  are zero-mean uncorrelated joint-gaussian RVs. This is a well-known condition for independence [ref ??], so the  $c_m$  are independent, gaussian, with variance 1 (see (7.25)). Then (7.27) tells us that, by definition,  $\mathbf{y}_{\text{mod}}^2$  is a chi-squared RV with  $p$  degrees of freedom:

$$\text{(raw) } \mathbf{y}_{\text{mod}}^2 \equiv SSA \in \chi_p^2 \quad (\text{zero-mean gaussian noise}) .$$

We developed this result using the properties of the orthonormal basis  $\mathbf{g}_m$ , but our model  $\mathbf{y}_{\text{mod}}$ , and therefore  $\mathbf{y}_{\text{mod}}^2$ , are identical in *any* basis. Therefore, the result holds for any  $p$  fit-functions that span the same model space, even if they are oblique (i.e. overlapping) and not normalized.

### The Residual Sum-of-Squares (SSE) in Pure Noise

For zero-mean gaussian noise, in the weighted case, we've shown that  $SSA$  is  $\chi^2$  distributed, but  $SST$  is not. Therefore,  $SSE$  is not, either. However, for large  $n$ , or for measurement uncertainties that are fairly consistent across the data set,  $SST$  and  $SSE$  are *approximately*  $\chi^2$  distributed, with the usual (i.e. equal uncertainty case) degrees of freedom assigned:

$$\underbrace{\sum_{i=1}^n w_i (y_i - \bar{y})^2}_{SST \text{ dof} \approx n-1} = \underbrace{\sum_{i=1}^n w_i (b_k f_k(x_i) - \bar{y})^2}_{SSA \text{ dof} = p-1} + \underbrace{\sum_{i=1}^n w_i (y_i - b_k f_k(x_i))^2}_{SSE \text{ dof} \approx n-p} \tag{zero-mean gaussian} .$$

## Hypothesis Testing a Model in Linear Regression with Uncertainties

The approximation that  $SST$  and  $SSE$  are almost  $\chi^2$  distributed allows the usual F-test as an *approximate* test for detection of a signal, i.e. testing whether the fit actually matches the presence of the model in the data. However, the F critical values will be approximate, and therefore so will the  $p$ -value. In many cases, numerical simulations (shuffle simulations) can provide more reliable critical values than the theoretical gaussian F critical values, for 2 reasons: even the theoretical F-values are only approximate (as described), and because the noise itself is often significantly non-gaussian.

We recommend numerical simulations (e.g., shuffling) to determine critical values, instead of the approximate (and often inapplicable) gaussian theory.

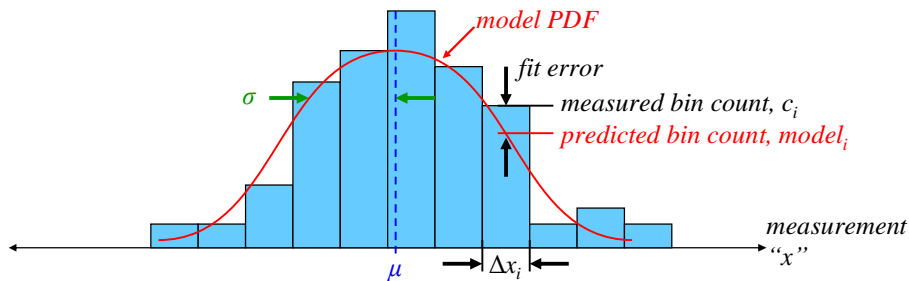
## Fitting To Histograms

Data analysis often requires fitting a function to binned data, that is, fitting a predicted probability distribution to a histogram of measured values. While such fitting is very commonly done, it is much less commonly understood. There are important subtleties often overlooked. This section assumes you are familiar with the binomial distribution, the  $\chi^2$  “goodness of fit” parameter (described earlier), and some basic statistics.

The general method for fitting a model to a histogram of data is this:

- Start with  $n$  data points (measurements), and a parameterized model for the PDF of those data.
- Bin the data into a histogram.
- Find the model parameters which “best fit” the data histogram

For example, a gaussian distribution is a 2-parameter model; the parameters are the average,  $\mu$ , and standard deviation,  $\sigma$ . If we believe our data should follow a gaussian distribution, and we want to know the  $\mu$  and  $\sigma$  of that distribution, we might bin the data into a histogram, and fit the gaussian PDF to it (Figure 7.13):



**Figure 7.13** Sample histogram with a 2-parameter model PDF ( $\mu$  and  $\sigma$ ). The fit model is gaussian in this example, but could be any pdf with any parameters.

Of course, realistically, there are better ways to estimate  $\mu$  and  $\sigma$  for a gaussian distribution, but the example illustrates the point of fitting to a histogram.

We must define “best fit.” Usually, we use the  $\chi^2$  (chi-squared) “goodness of fit” parameter as the figure of merit (FOM). The smaller  $\chi^2$ , the better the fit. Fitting to a histogram is a special case of general  $\chi^2$  fitting. Therefore, we need to know two things for each bin: (1) the predicted (model) count, and (2) the uncertainty in the measured count. We find these things in the next section.

(This gaussian fit is a simplified example. In reality, if we think the distribution is gaussian, we would compute the sample average and standard deviation directly, using the standard formulas. More on this later. In general, the model is more complicated, and there is no simple formula to compute the parameters. For now, we use this as an example because it is a familiar model to many.)

### Chi-squared For Histograms

We now develop the  $\chi^2$  figure of merit for fitting to a histogram. A **sample** is a set of  $n$  measurements (data points). In principle, we could take many samples of data. For each sample, there is one histogram, i.e., there is an infinite population of samples, each with its own histogram. But we have only one sample. The question is, how well does our *one* histogram represent the population of samples, and therefore, the population of data measurements.

To develop the  $\chi^2$  figure of merit for the fit, we must understand the statistics of a *single* histogram bin, from the population of all histograms that we might have produced from different samples. The key point is this: given a sample of  $n$  data points, and a particular histogram bin numbered  $i$ , each data point in the sample is either in the bin (with probability  $p_i$ ), or it's not (with probability  $(1 - p_i)$ ). Therefore, the count in the  $i^{\text{th}}$  histogram bin is binomially distributed, with some probability  $p_i$ , and  $n$  "trials." (See standard references on the binomial distribution if this is not clear.) Furthermore, this is true of *every* histogram bin:

The number of counts in each histogram bin is a binomial random variable.  
Each bin has its own probability,  $p_i$ , but all bins share the same number of trials,  $n$ .

Recall that a binomial distribution is a discrete distribution, i.e. it gives the probability of finding values of a whole-number random variable; in this case, it gives the probability for finding a given number of counts in a given histogram bin. The binomial distribution has two parameters:

- $p$  is the probability of a given data point being in the bin
- $n$  is the number of data points in the sample, and therefore the number of "trials" in the binomial distribution.

Recall that the binomial distribution has average, *c-bar*, and variance,  $\sigma^2$  given by:

$$\bar{c} = np, \quad \sigma^2 = np(1 - p) \quad (\text{binomial distribution})$$

For a large number of histogram bins,  $N_{bins}$ , the probability of being in a given bin is of order  $p \sim 1/N_{bins}$ , which is small. Therefore, we approximate

$$\sigma^2 \approx np(1) = \bar{c} \quad (N_{bins} \gg 1 \Rightarrow p \ll 1)$$

We find *c-bar* for a bin from the pdf model: typically, we assume the bins are narrow, and the probability of being in a bin is just

$$\text{Pr}(\text{being in bin } i) \equiv p_i \approx \text{pdf}_x(x_i) \Delta x_i$$

Then the model average ("expected") count is Pr(being in bin) times the number of data points,  $n$ :

$$\text{model}_i \approx n \text{pdf}_x(x_i) \Delta x_i \quad (\text{narrow bins})$$

$$\text{where } x_i \equiv \text{bin center}, \quad \Delta x_i \equiv \text{bin width}$$

$$\text{pdf}_x(x_i) \equiv \text{model pdf at bin center}$$

For example, for a gaussian histogram: 
$$\text{pdf}_x(\mu, \sigma; x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

However, one can use any more sophisticated method to properly integrate the PDF to find  $e$  for each bin.

We now know the two things we need for evaluating a general  $\chi^2$  goodness-of-fit parameter: for each histogram bin, we know (1) the model average count,  $\text{model}_i$ , and (2) the variance of the measured count, which is also approximately  $\text{model}_i$ . We now compute  $\chi^2$  for the model PDF (given a set of model parameters) in the usual way:

$$\chi^2 \equiv \sum_{i=1}^{N_{bins}} \frac{(c_i - model_i)^2}{model_i} \quad \text{where } c_i \equiv \text{the measured count in the } i^{th} \text{ bin}$$

$$model_i \equiv \text{the model average count in the } i^{th} \text{ bin}$$

If your model predicts a count of zero ( $model_i = 0$  for some  $i$ ), then  $\chi^2$  blows up. This is addressed below.

**Reducing the Effect of Noise**

To find the best-fit parameters, we take our given sample histogram, and try different values of the pdf(x) parameters (in this example,  $\mu$  and  $\sigma$ ) to find the combination which produces the minimum  $\chi^2$ .

Notice that the low count bins carry more weight than the higher count bins:  $\chi^2$  weights the terms by  $1/model_i$ . This reveals the first common misunderstanding:

A fit to a histogram is driven by the tails, not by the central peak. This is usually bad.

Tails are often the worst part of the model (theory), and often the most contaminated (percentage-wise) by noise: background levels, crosstalk, etc. Three methods help reduce these problems:

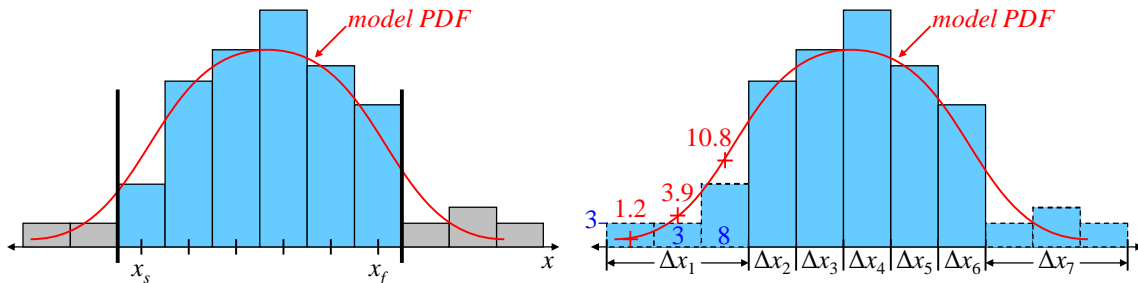
- limiting the weight of low-count bins
- truncating the histogram
- rebinning

**Limiting the weight:** The tails of the model distribution are often less than 1, and approach zero. This gives them extremely high weights compared to other bins. Since the model is probably inadequate at these low bin counts (due to noise, etc.), one can limit the denominator in the  $\chi^2$  sum to at least 1; this also avoids division-by-zero:

$$\chi^2 \equiv \sum_{i=1}^{N_{bins}} \frac{(c_i - model_i)^2}{d_i} \quad \text{where } d_i \equiv \begin{cases} model_i & \text{if } model_i > 1 \\ 1 & \text{otherwise} \end{cases}$$

This is an ad-hoc approach, and the minimum weight can be anything; it doesn't have to be 1. Notice, though, that this modified  $\chi^2$  value is still a monotonic function of the model parameters, which is critical for stable parameter fits (it avoids local minima, see "Practical Considerations" below).

**Truncating the histogram:** Alternatively, we can truncate the histogram on the left and right sides to those bins with a reasonable number of counts, substantially above the noise (below left). [Bev p110] recommends a minimum bin count of 10, based on a desire for gaussian errors. I don't think that matters much. In truth, the minimum count completely depends on the noise level.



Avoiding noisy tails by (left) truncating the histogram, or (right) rebinning.

Truncation requires renormalizing: we normalize the model within the truncated limits to the data count within those same limits:

$$\sum_{i=s}^f model_i = \sum_{i=s}^f n_{norm} pdf_X(x_i) \Delta x_i = \sum_{i=s}^f c_i \quad \text{where } s, f \text{ are the start and final bins to include}$$

$$\Rightarrow n_{norm} = \frac{\sum_{i=s}^f c_i}{\sum_{i=s}^f pdf_X(x_i) \Delta x_i}$$

You might think that we should use the model, not the data histogram, to choose our truncation limits. After all, why should we let sampling noise affect our choice of bins? This approach fails miserably, however, because our bin choices change as we vary our parameters in the hunt for the optimum  $\chi^2$ . Changing which bins are included in the FOM causes unphysical steps in  $\chi^2$  as we vary our parameters, making many local minima. This makes the fit unstable, and generally unusable. For stability: truncate your histogram based on the data, and keep it fixed during the parameter search.

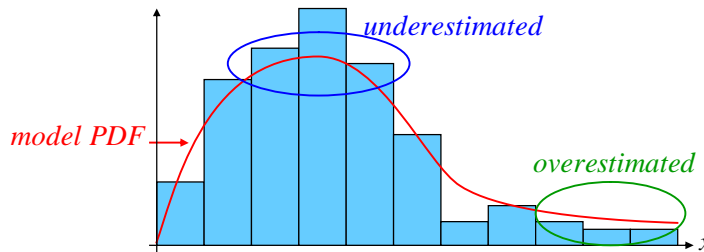
**Rebinning:** Alternatively, bins don't have to be of uniform width [Bev p175], so combining adjacent bins into a single, wider bin with higher count can help improve signal-to-noise ratio (SNR) in that bin (above right). Note that when rebinning, we evaluate the theoretical count as the sum of the original (narrow) bin theoretical counts. In the example of the diagram above right, the theoretical and measured counts for the new (wider) bin 1 are

$$model_1 = 1.2 + 3.9 + 10.8 = 15.9 \quad \text{and} \quad c_1 = 3 + 3 + 8 = 14$$

**Other Histogram Fit Considerations**

**Slightly correlated bin counts:** Bin counts are binomially distributed (a measurement is either in a bin, or it's not). However, there is a small negative correlation between any two bins, because the fact that a measurement lies in one bin means it doesn't lie in any other bin. Recall that the  $\chi^2$  parameter relies on *uncorrelated* errors between bins, so a histogram slightly violates that assumption. With a moderate number of bins (> ~15 ??), this is usually negligible.

**Overestimating the low count model:** If there are a lot of low-count bins in your histogram, you may find that the fit tends to overestimate the low-count bins, and underestimate the high-count bins (diagram below). When properly normalized, the sum of overestimates and underestimates must be zero: the sum of bin counts equals the sum of the model predicted counts.



$\chi^2$  is artificially reduced by overestimating low-count bins, and underestimating high-count bins.

But since low-count bins weigh more than high-count bins, and since an overestimated model reduces  $\chi^2$  (the model value  $model_i$  appears in the denominator of each  $\chi^2$  term), the overall  $\chi^2$  is reduced if low-count bins are overestimated, and high-count bins are underestimated.

This effect can only happen if your model has the freedom to “bend” in the way necessary: i.e., it can be a little high in the low-count regions, and simultaneously a little low in the high-count regions. Most realistic models have this freedom. If the model is reasonably good, this effect can cause reduced- $\chi^2$  to be consistently less than 1 (which should be impossible).

I don't know of a simple fix for this. It helps to limit the weight of low-count bins to (say) 1, as described above. However once again, the best approach is to minimize the number of low-count bins in your histogram.

**Noise not zero mean:** for counting experiments, such as those that fill in histograms with data, all bin counts are zero or positive. Any noise will add positive counts, and therefore noise cannot be zero-mean. If you know the pdf of the noise, then you can put it in the model, and everything should work out fine. However, if you have a lot of un-modeled noise, you should see that your reduced- $\chi^2$  is significantly greater than 1, indicating a poor fit. Some people have tried playing with the denominator in the  $\chi^2$  sum to try to get more “accurate” fit parameters in the presence of noise, but there is little theoretical justification for this, and it usually amounts to ad-hoc tweaking to get the answers you want.

**Non- $\chi^2$  figure of merit:** One does not have to use  $\chi^2$  as the fit figure of merit. If the model is not very good, or if there are problems as mentioned above, other FOMs might work better. The most common alternative is probably “least-squares,” which means minimizing the sum-squared-error:

$$SSE \equiv \sum_{i=1}^{N_{bins}} (c_i - model_i)^2 \quad (\text{sum-squared-error}).$$

This is like  $\chi^2$  where the denominator in each term in the sum is always 1.

---

## Guidance Counselor: Practical Considerations for Computer Code to Fit Data

Generic optimization algorithms are available off-the-shelf, e.g. [Numerical Recipes]. However, they are sometimes simplistic, and in the real world, often fail with arithmetic faults (overflow, underflow, domain-error, etc). The fault (no pun intended) lies not in their algorithm, but in their failure to tell you what you need to do to avoid such failures:

Your job is to write a bullet-proof figure-of-merit function.

This is harder than it sounds, but quite do-able with proper care.

As an example, I once wrote code to fit a sinusoid (frequency, amplitude, phase) to astronomical data: measures of a star's brightness at irregular times. That seems pretty simple, yet it was fraught with problems. The measurements were very noisy, which leads to lots of local minima. In some cases, the optimizer would choose an amplitude for the sinusoid that had a higher sum-of-squares than the sum-of-squares of the data! This amplitude is clearly “too big,” but it is hard to know ahead of time how big is “too big.” Furthermore, the “too big” threshold varies with the frequency and phase parameters, so you cannot specify ahead of time an absolute “valid range” for amplitude. Therefore, I had to provide “guiding errors” in my figure-of-merit function to “guide” the optimizer to a reasonable fit under all conditions.

Computer code for finding the best-fit parameters is usually divided into two pieces, one piece you buy, and one piece you have to write yourself:

- You buy a generic optimization algorithm, which varies parameters without knowledge of what they mean, looking for the minimum figure-of-merit (FOM). For each trial set of parameters, it calls your FOM function to compute the FOM as a function of the current trial parameters.
- You write the FOM function which computes the FOM as a function of the given parameters.

Generic optimizers usually minimize the figure-of-merit, consistent with the FOM being a “cost” or “error” that we want reduced. (If instead, you want to maximize a FOM, return its negative to the minimizer.)

Generic optimizers know nothing about your figure-of-merit (FOM) function, or its behavior, and your FOM usually knows nothing about the optimizer, or its algorithms.

If your optimizer allows you to specify valid ranges for parameters, *and* if your fit parameters have valid ranges that are independent of each other, then you don't need the methods here for your FOM function. If your optimizer (like many) does *not* allow you to limit the range of parameters, or if your



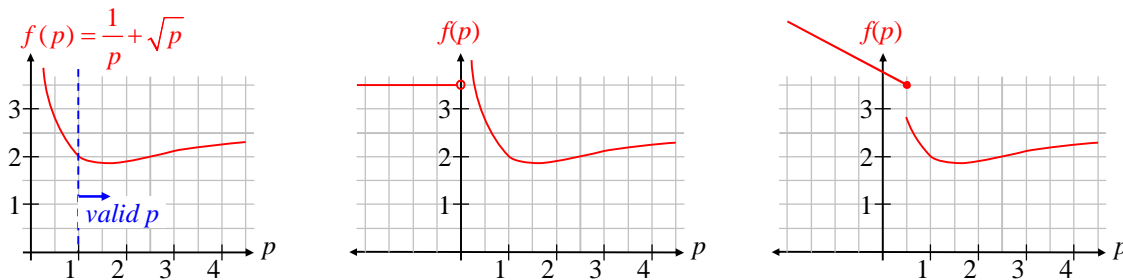
parameters have valid ranges that depend on each other, then you need the following methods to make a bullet-proof FOM. In either case, this section illustrates how many seemingly simple calculations can go wrong in unexpected ways.

A bullet-proof FOM function requires only two things:

- Proper validation of all parameters.
- A properly “bad” FOM for invalid parameters (a “guiding error”).

Guiding errors are similar to penalty functions, but they operate *outside* the valid parameter space, rather than inside it.

**A simple example:** Suppose you wish to numerically find the minimum of the figure-of-merit function below left. Suppose the physics is such that only  $p > 1$  is sensible.



(Left and middle) Bad figure-of-merit (FOM) functions. (Right) A bullet-proof FOM.

Your optimization-search algorithm will try various values of  $p$ , evaluating  $f(p)$  at each step, looking for the minimum. You might write your FOM function like this:

```
fom(p) = 1./p + sqrt(p)
```

But the search function knows nothing of  $p$ , or which values of  $p$  are valid. It may well try  $p = -1$ . Then your function crashes with a domain-error in the `sqrt()` function. You fix it with (above middle):

```
float fom(p)
    if(p < 0.) return 4.
    return 1./p + sqrt(p)
```

Since you know 4 is much greater than the true minimum, you hope this will fix the problem. You run the code again, and now it crashes with divide-by-zero error, because the optimizer tried  $p = 0$ . Easy fix:

```
float fom(p)
    if(p <= 0.) return 4.
    return 1./p + sqrt(p)
```

Now the optimizer crashes with an overflow error,  $p < -(\text{max\_float})$ . The big flat region to the left confuses the optimizer. It searches negatively for a value of  $p$  that makes the FOM increase, but it never finds one, and gets an overflow trying. Your flat value for  $p \leq 0$  is no good. It needs to grow upward to the left to provide guidance to the optimizer:

```
float fom(p)
    if(p <= 0.) return 4. + fabs(p - 1) // fabs() = absolute value
    return 1./p + sqrt(p)
```

Now the optimizer says the minimum is 4 at  $p = -10^6$ . It found the local “minimum” just to the left of zero. Your function is still ill-behaved. Since only  $p > 1$  is sensible, you make yet another fix (above right):

```
float fom(p)
    if(p <= 1.) return 4.
    return 1./p + sqrt(p)
```

Finally, the optimizer returns the minimum FOM of 1.89 at  $p = 1.59$ . After 5 tries, you have made your FOM function bullet-proof:

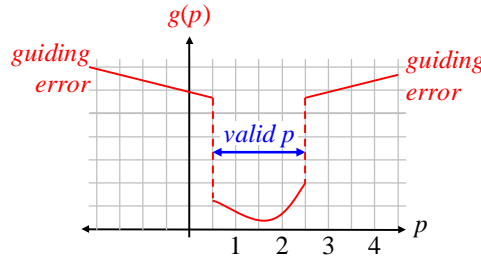
A bullet-proof FOM has only one minimum, which it monotonically approaches from both sides, even with invalid parameters, and it *never* crashes on any parameter set.

In this example, the FOM is naturally bullet-proof from the right. However, if it weren't, the absolute value of  $(p - 1)$  on the error return value provides a V-shape which guides the optimizer into the valid range from either side. Such “guiding errors” are analogous to so-called penalty functions, but better, because they take effect only for invalid parameter choices, thus leaving the valid parameter space completely free for full optimization.

**Multi-parameter FOMs:** Most fit models use several parameters,  $p_i$ , and the optimizer searches over all of them iteratively to find a minimum. Your FOM function must be bullet-proof over *all* parameters: it must check each parameter for validity, and *must* return a large (guaranteed unoptimal) result for invalid inputs. It must also *slope the function toward valid values*, i.e. provide a “restoring force” to the invalid parameters toward the valid region. Typically, with multiple parameters  $p_i$ , one uses:

$$\text{guiding\_bad\_FOM} = \text{big \#} + \sum_{i=1}^M |p_i - \text{valid}_i| \quad \text{where } \text{valid}_i \equiv \text{a valid value for } p_i .$$

This guides the minimization search when any parameter is outside its valid range.



“Guiding errors” lead naturally to a valid solution, and are better than traditional penalty functions.

A final note:

The “big #” for invalid parameters may need to be much bigger than you think.

In my dissertation research, I used reduced  $\chi^2$  as my FOM, and the true minimum FOM is near 1. I started with 1,000,000 as my “big #”, but it wasn't big enough! I was fitting to histograms with nearly a thousand counts in several bins. When the trial model bin count was small, the error was about 1,000, and the sum-squared-error over several bins was  $> 1,000,000$ . This caused the optimizer to settle on an invalid set of parameter values as the minimum! I had to raise “big #” to  $10^9$ .

## 8 Numerical Analysis

### Round-Off Error, And How to Reduce It

Floating point numbers are stored in a form of scientific notation, with a **mantissa** and **exponent**. E.g.,

$$1.23 \times 10^{45} \quad \text{has mantissa} \quad m = 1.23 \quad \text{and exponent} \quad e = 45.$$

Computer floating point stores only a finite number of digits. ‘float’ (aka single-precision) typically stores at least 6 digits; ‘double’ typically stores at least 15 digits. We’ll work out some examples in 6-digit decimal scientific notation; actual floating point numbers are stored in a binary form, but the concepts remain the same. (See “IEEE Floating Point” in this document.)

Precision loss due to summation: Adding floating point numbers with different exponents results in **round-off error**:

$$\begin{aligned} 1.234\,56 \times 10^2 &\rightarrow 1.234\,56 \times 10^2 \\ + 6.111\,11 \times 10^0 &\quad + 0.061\,111\,1 \times 10^2 \\ &= 1.295\,67 \times 10^2 \quad \text{where } 0.000\,001\,1 \text{ of the result is lost,} \end{aligned}$$

because the computer can only stored 6 digits. (Similar round-off error occurs if the exponent of the result is bigger than both of the addend exponents.) When adding many numbers of similar magnitude (as is common in statistical calculations), the round-off error can be quite significant:

```
float sum = 1.23456789;           // Demonstrate precision loss in sums
printf("%.9f\n", sum);           // show # significant digits
for(i = 2; i < 10000; i++)
    sum += 1.23456789;
printf("Sum of 10,000 = %.9f\n", sum);

1.234567881           8 significant digits
Sum of 10,000 = 12343.28 only 4 significant digits
```

You lose about 1 digit of accuracy for each power of 10 in  $n$ , the number of terms summed. I.e.

$$\text{digit-loss} \approx \log_{10} n$$

When summing numbers of different magnitudes, you get a better answer by adding the small numbers first, and the larger ones later. This minimizes the round-off error on each addition.

E.g., consider summing  $1/n$  for 1,000,000 integers. We do it in both single- and double-precision, so you can see the error:

```
float sum = 0.;
double dsum = 0.;
// sum the inverses of the first 1 million integers, in order
for(i = 1; i <= 1000000; i++)
    sum += 1./i, dsum += 1./i;
printf("sum: %f\ndsum: %f. Relative error = %.2f %%\n",
       sum, dsum, (dsum-sum)/dsum);

sum: 14.357358
dsum: 14.392727. Relative error = 0.002457
```

This was summed in the worst possible order: largest to smallest, and (in single-precision) we lose about 5 digits of accuracy, leaving only 3 digits. Now sum in reverse (smallest to largest):

```
float sumb = 0.;
double dsumb = 0.;
for(i = 1000000; i >= 1; i--)
    sumb += 1./i, dsumb += 1./i;
printf(" sumb: %f\ndsumb: %f. Relative error = %.6f\n",
```

```

sumb, dsumb, (dsumb-sumb)/dsumb);
sumb: 14.392652
dsumb: 14.392727. Relative error = 0.000005

```

The single-precision sum is now good to 5 digits, losing only 1 or 2.

[In my research, I needed to fit a polynomial to 6000 data points, which involves many sums of 6000 terms, and then solving linear equations. I needed 13 digits of accuracy, which easily fits in double-precision ('double', 15-17 decimal digits). However, the precision loss due to summing was over 3 digits, and my results failed. Simply changing the sums to 'long double', then converting the sums back to 'double', and doing all other calculations in 'double' solved the problem. The dominant loss was in the sums, not in solving the equations.]

Summing from smallest to largest is very important for evaluating polynomials, which are widely used for transcendental functions. Suppose we have a 5<sup>th</sup> order polynomial,  $f(t)$ :

$$f(t) = a_0 + a_1x + a_2x^2 + a_3x^3 + a_4x^4 + a_5x^5$$

which might suggest a computer implementation as :

```
f = a0 + a1*t + a2*t*t + a3*t*t*t + a4*t*t*t*t + a5*t*t*t*t*t
```

Typically, the terms get progressively smaller with higher order. Then the above sequence is in the *worst* order: biggest to smallest. (It also takes 15 multiplies.) It is more accurate (and faster) to evaluate the polynomial as:

```
f = (((((a5*t + a4)*t + a3)*t + a2)*t + a1)*t + a0
```

This form adds small terms of comparable size first, progressing to larger ones, and requires only 5 multiplies.

## How To Extend Precision In Sums Without Using Higher Precision Variables

(Handy for statistical calculations): You can avoid round-off error in sums without using higher precision variables with a simple trick. For example, let's sum an array of  $n$  numbers:

```
sum = 0.;
for(i = 0; i < n; i++) sum += a[i];
```

This suffers from precision loss, as described above. The trick is to actually measure the round-off error of each addition, and save that error for the next iteration:

```
sum = 0.;
error = 0.; // the carry-in from the last add
for(i = 0; i < n; i++)
{
    newsum = sum + (a[i] + error); // include the lost part of prev add
    diff = newsum - sum; // what was really added
    error = (a[i] + error) - diff; // the round-off error
    sum = newsum;
}
```

The 'error' variable is always small compared to the 'sum', because it is the round-off error. Keeping track of it effectively doubles the number of accurate digits in the sum, until it is lost in the final addition. Even then, 'error' still tells you how far off your sum is. For all practical purposes, this eliminates any precision loss due to sums. Let's try summing the inverses of integers again, in the "bad" order, but with this trick:

```
float newsum, diff, sum = 0., error = 0.;
for(i = 1; i <= 1000000; i++)
{
    newsum = sum + (1./i + error);
    diff = newsum - sum; // what was really added
    error = (1./i + error) - diff; // the round-off error
    sum = newsum;
}
printf(" sum: %f\ndsumb: %f. Relative error = %.6f, error = %g\n",
       sum, dsumb, (dsumb-sum)/dsumb, error);
```

```
sum: 14.392727
dsumb: 14.392727. Relative error = -0.000000, error = -1.75335e-07
```

As claimed, the sum is essentially perfect.

## Numerical Integration

The above method of sums is extremely valuable in numerical integration. Typically, for accurate numerical integration, one must carefully choose an integration step size: the increment by which you change the variable of integration. E.g., in time-step integration, it is the time step-size. If you make the step size too big, accuracy suffers because the “rectangles” (or other approximations) under the curve don’t follow the curve well. If you make the step size too small, accuracy suffers because you’re adding tiny increments to large numbers, and the round-off error is large. You must “thread the needle” of step-size, getting it “just right” for best accuracy. This fact is independent of the integration interpolation method: trapezoidal, quadratic, (Runge-Kutta??).

By virtually eliminating round-off error in the sums (using the method above), you eliminate the lower-bound on step size. You can then choose a small step-size, and be confident your answer is right. It might take more computer time, but integrating 5 times slower and getting the right answer is vastly better than integrating 5 times faster and getting the wrong answer.

## Sequences of Real Numbers

Suppose we want to generate the sequence 2.01, 2.02, ... 2.99, 3.00. A simple approach is this:

```
real s;
for(s = 2.01; s <= 3.; s += 0.01) ...
```

The problem with this is round-off error: 0.01 is inexact in binary (has round-off error). This error accumulates 100 times in the above loop, making the last value 100 times more wrong than the first. In fact, the loop might run 101 times instead of 100. The fix is to use integers where possible, because they are exact:

```
real s;
int i;
for(i = 201; i <= 300; i++) s = i/100.;
```

When the increment is itself a variable, note that multiplying a real by an integer incurs only a single round-off error:

```
real s, base, incr;
int i;
for(i = 1; i <= max; i++) s = base + i*incr;
```

Hence, every number in the sequence has only one round-off error.

## Root Finding

In general, a **root** of a function  $f(x)$  is a value of  $x$  for which  $f(x) = 0$ . It is often not possible to find the roots analytically, and it must be done numerically. [TBS: binary search]

### Simple Iteration Equation

Some forms of  $f()$  make root finding easy and fast; if you can rewrite the equation in this form:

$$f(x) = 0 \quad \rightarrow \quad x = g(x)$$

then you may be able to iterate, using each value of  $g()$  as the new estimate of the root,  $r$ .

This is the simplest method of root finding, and generally the slowest to converge.

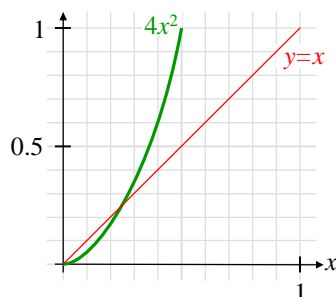
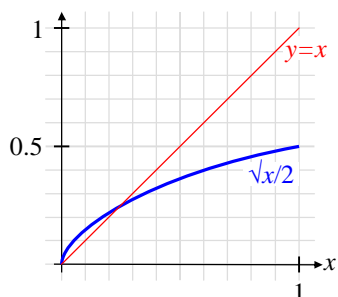
It may be suitable if you have only a few thousand solutions to compute, but may be too slow for millions of calculations.

You start with a guess that is close to the root, call it  $r_0$ . Then

$$r_1 = g(r_0), \quad r_2 = g(r_1), \quad \dots \quad r_{n+1} = g(r_n)$$

If  $g(x)$  has the right property (specifically,  $|g'(x)| < 1$  near the root) this sequence will converge to the solution. We describe this necessary property through some examples. Suppose we wish to solve

$\sqrt{x}/2 - x = 0$  numerically. First, we re-arrange it to isolate  $x$  on the left side:  $x = \frac{\sqrt{x}}{2}$  (below left).



Two iteration equations for the same problem. The left converges; the right fails.

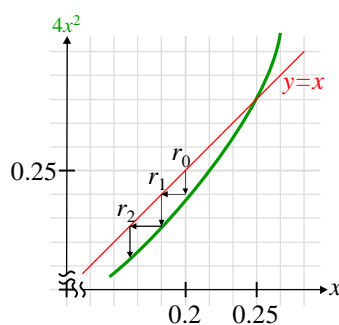
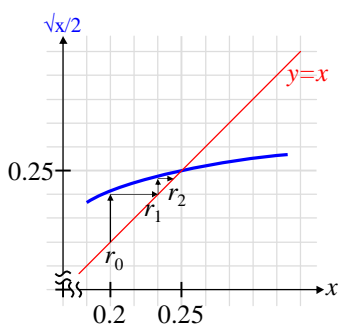
From the graph, we might guess  $r_0 \approx 0.2$ . Then we would find,

$$r_1 = \sqrt{0.2}/2 = 0.2236, \quad r_2 = \sqrt{r_1}/2 = 0.2364, \quad r_3 = 0.2431, \quad r_4 = 0.2465, \quad r_5 = 0.2483, \\ R_6 = 0.2491, \quad r_7 = 0.2496$$

We see that the iterations approach the exact answer of 0.25. But we could have re-arranged the equation differently:  $2x = \sqrt{x}$ ,  $x = 4x^2$  (above right). Starting with the same guess  $x = 0.2$ , we get this sequence:

$$r_1 = \sqrt{0.2}/2 = 0.16, \quad r_2 = \sqrt{r_1}/2 = 0.1024, \quad r_3 = 0.0419, \quad r_4 = 0.0070$$

But they are not converging on the nearby root; the sequence diverges away from it. So what's the difference? Look at a graph of what's happening, magnified around the equality:



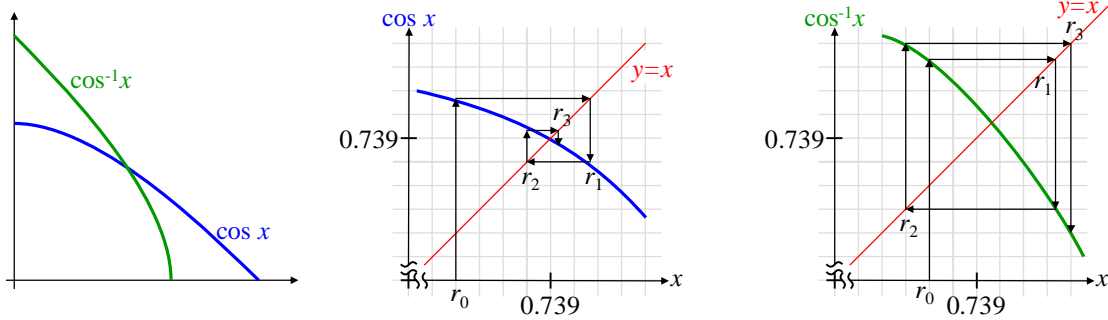
The left converges; the right fails.

When the curve is flatter than  $y = x$  (above left), then trial roots that are too small get bigger, and trial roots that are too big get smaller. So iteration approaches the root. When the curve is steeper than  $y = x$  (above right), trial roots that are too small get even smaller, too big get even bigger; the opposite of what we want. So for positive slope curves, the condition for convergence is

$$\frac{\Delta y}{\Delta r} = \frac{y(s) - y(r)}{s - r} < 1, \quad \text{in the region} \quad r - |r_0 - r| < s < r + |r_0 - r| \quad |r_i - r| < |r_0 - r|,$$

where  $r$  is the exact root;  $r_0$  is the first guess.

Consider another case, where the curve has negative slope. Suppose we wish to solve  $\cos^{-1} x - x = 0$ , ( $x$  in radians). We re-write it as  $x = \cos^{-1} x$ . On the other hand, we could take the cosine of both sides and get an equivalent equation:  $x = \cos x$ . Which will converge? Again look at the graphs:



**Figure 8.1** (Left)  $\cos$  and  $\cos^{-1}$  are superficially similar. (Middle)  $\cos$  converges everywhere. (Right)  $\cos^{-1}$  fails everywhere.

So long as the magnitude of the slope  $< 1$  in the neighborhood of the solution, the iterations converge. When the magnitude of the slope  $> 1$ , they diverge. We can now generalize to all curves of any slope:

The general condition for convergence is

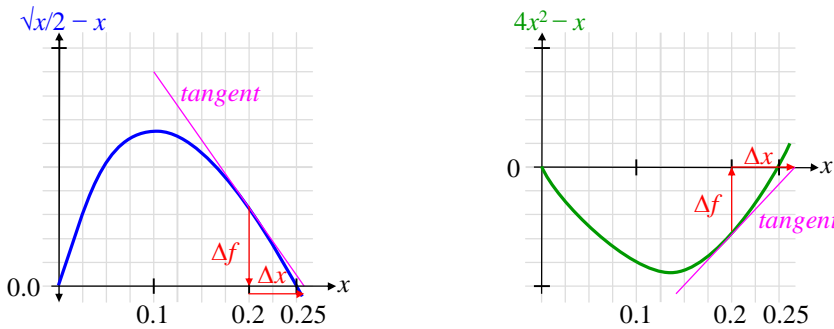
$$\left| \frac{\Delta y}{\Delta r} \right| = \left| \frac{y(s) - y(r)}{s - r} \right| < 1, \quad \text{in the region} \quad r - |r_0 - r| < s < r + |r_0 - r|.$$

The flatter the curve, the faster the convergence.

Given this, we could have easily predicted that the converging form of our iteration equation is  $x = \cos x$ , because the slope of  $\cos x$  is always  $< 1$ , and  $\cos^{-1} x$  is always  $> 1$ . Note, however, that if the derivative (slope) is  $> 1/2$ , then the binary search will be faster than iteration.

### Newton-Raphson Iteration

The above method of variable iteration is kind of “blind,” in that it doesn’t use any property of the given functions to advantage. **Newton-Raphson iteration** is a method of finding roots that uses the derivative of the given function to provide more reliable and faster convergence. Newton-Raphson uses the original form of the equation:  $f(x) = \sqrt{x}/2 - x = 0$ . The idea is to use the derivative of the function to approximate its slope to the root (below left). We start with the same guess,  $r_0 = 0.2$ .



$$\frac{\Delta f}{\Delta x} \approx f'(r_0) \quad \Rightarrow \quad \Delta r \approx -\frac{f(r_i)}{f'(r_i)} \quad (\text{Note } f'(r_0) < 0)$$

$$f'(x) = \frac{1}{4}x^{-1/2} - 1 \quad \Rightarrow \quad \Delta r = -\frac{r_i^{1/2}/2 - r_i}{r_i^{-1/2}/4 - 1} \cdot \frac{4r_i^{1/2}}{4r_i^{1/2}} = -\frac{2r_i - 4r_i^{3/2}}{1 - 4r_i^{1/2}}$$

Here's a sample computer program fragment, and its output:

```
// Newton-Raphson iteration
r = 0.2;
for(i = 1; i < 10; i++)
{
    r -= (2.*r - 4.*r*sqrt(r)) / (1. - 4.*sqrt(r));
    printf("r%d %.16f\n", i, r);
}
```

```
r1 0.2535322165454392
r2 0.2500122171752588
r3 0.2500000001492484
r4 0.2500000000000000
```

In 4 iterations, we get essentially the exact answer, to double precision accuracy of 16 digits. This is much faster than the variable isolation method above. In fact, it illustrates a property of some iterative numerical methods called **quadratic convergence**:

Quadratic convergence is when the fractional error (aka relative error) gets squared on each iteration, which doubles the number of significant digits on each iteration.

You can see this clearly above, where  $r_1$  has 2 accurate digits,  $r_2$  has 4,  $r_3$  has 9, and  $r_4$  has at least 16 (maybe more). Derivation of quadratic convergence??

Also, Newton-Raphson does not have the restriction on the slope of any function, as does variable isolation. We can use it just as well on the reverse formula (previous diagram, right):

$$f(x) = 4x^2 - x, \quad f'(x) = 8x - 1, \quad \Delta r = -\frac{f(r_i)}{f'(r_i)} = -\frac{4x^2 - x}{8x - 1}, \text{ with these computer results:}$$

```
r1 0.2666666666666667
r2 0.2509803921568627
r3 0.2500038147554742
r4 0.2500000000582077
r5 0.2500000000000000
```

This converges essentially just as fast, and clearly shows quadratic convergence.

If you are an old geek like me, you may remember the iterative method of finding square roots on an old 4-function calculator: to find  $\sqrt{a}$ : divide  $a$  by  $r$ , then average the result with  $r$ . Repeat as needed:

$$r_{n+1} = \frac{a/r_n + r_n}{2}$$

You may now recognize that as Newton-Raphson iteration:

$$f(r) = r^2 - a = 0, \quad f'(r) = 2r,$$

$$r_{n+1} = r_n + \Delta r = r_n - \frac{f(r)}{f'(r)} = r_n - \frac{r_n^2 - a}{2r_n} = r_n - \frac{r_n}{2} + \frac{a}{2r_n} = \frac{1}{2} \left( r_n + \frac{a}{r_n} \right)$$



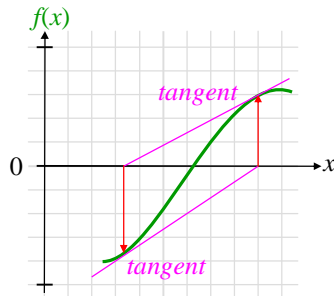
If you are truly a geek, you tried the averaging method for cube roots:  $r_{n+1} = \frac{a/r_n^2 + r_n}{2}$ . While you found that it converged, it was very slow; cube-root(16) with  $r_0 = 2$  gives only 2 digits after 10 iterations. Now you know that the proper Newton-Raphson iteration for cube roots is:

$$f(r) = r^3 - a = 0, \quad f'(r) = 3r^2, \quad r_{n+1} = r_n - \frac{r_n^3 - a}{3r_n^2} = r_n - \frac{r_n}{3} + \frac{a}{3r_n^2} = \frac{1}{3} \left( 2r_n + \frac{a}{r_n^2} \right)$$

which gives a full 17 digits in 5 iterations for  $r_0 = 2$ , and shows (of course) quadratic convergence:

```
r1 2.6666666666666665
r2 2.5277777777777777
r3 2.5198669868999541
r4 2.5198421000355395
r5 2.5198420997897464
```

It is possible for Newton-Raphson to cycle endlessly, if the initial estimate of the root is too far off, and the function has an inflection point between two successive iterations:



### Failure of Newton-Raphson iteration.

It is fairly easy to detect this failure in code, and pull in the root estimate before iterating again.

## Pseudo-Random Numbers

We use the term “random number” to mean “pseudo-random number,” for brevity. **Uniformly distributed** random numbers are equally likely to be anywhere in a range, typically (0, 1).

Uniformly distributed random numbers are the starting point  
for many other statistical applications.

Computers can easily generate *uniformly distributed* random numbers. The best generators today are based on linear feedback shift registers (LFSR) [*Numerical Recipes*, 3<sup>rd</sup> ed.]. The old linear-congruential generator is:

```
// Uniform random value, 0 < v < 1, i.e. on (0,1) exclusive.
// Numerical Recipes in C, 2nd ed., p284
static uint32 seed=1; // starting point
vflt rand_uniform(void)
{
    do seed = 1664525L*seed + 1013904223L; // period 2^32-1
    while(seed == 0);
    rand_calls++; // count calls for repetition check
    return seed / 4294967296.;
} // rand_uniform()
```

Many algorithms that use such random numbers fail on 0 or 1,  
so this generator never returns them.

After a long simulation with a large number of calls, it's a good idea to check 'rand\_calls' to be sure it's < ~400,000,000 = 10% period. This confirms that the generated numbers are essentially random, and not predictable.

**Arbitrary distribution random numbers:** To generate any distribution from a uniform random number:

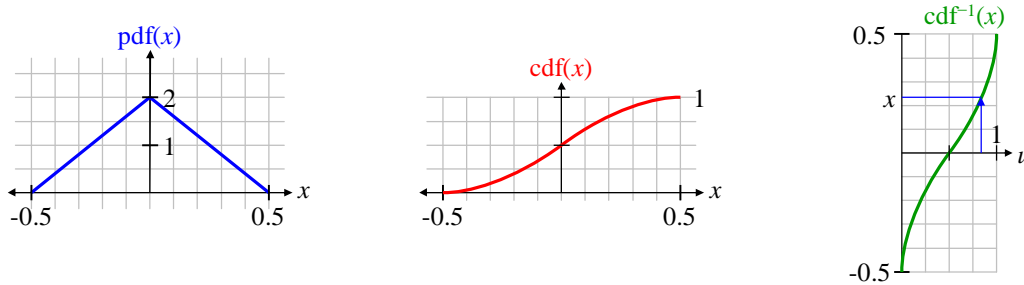
$$R = \text{cdf}_R^{-1}(U) \quad \text{where } R \text{ is the random variable of the desired distribution}$$

$$\text{cdf}_R^{-1} = \text{inverse of the desired cumulative distribution function of } R$$

$$U \text{ is a uniform random number on } (0,1)$$

Figure 8.2 illustrates the process graphically. We can derive it mathematically as follows: recall that the **cumulative distribution function** gives the probability of a random variable being less than or equal to its argument:

$$\text{cdf}_X(a) \equiv \Pr(X \leq a) = \int_{-\infty}^a dx \text{pdf}_X(x) \quad \text{where } X \text{ is a random variable.}$$



**Figure 8.2** Steps to generating the probability distribution function (pdf) on the left.

Also recall that the pdf of a function of a random variable, say  $F = f(u)$ , is (see *Probability and Statistics* elsewhere in this document):

$$\text{pdf}_F(x) = \frac{\text{pdf}_U(u)}{f'(x)}, \quad \text{where } f'(x) \equiv \text{derivative of } f(x).$$

Let  $Q \equiv \text{cdf}_R^{-1}(U)$ . Using  $\text{pdf}_U(u) = 1$  on  $[0, 1]$

$$\text{pdf}_Q(r) = \frac{\text{pdf}_U(u)}{\frac{d}{du} \text{cdf}_R^{-1}(u)} = \frac{1}{\left(\frac{d}{dr} \text{cdf}_R(r)\right)^{-1}} \quad \text{using } \frac{d}{du} g^{-1}(u) = \left(\frac{d}{du} g(u)\right)^{-1}, \text{ and } u \rightarrow r$$

$$= \text{pdf}_R(r), \quad \text{as desired.}$$

**Generating Gaussian Random Numbers**

The inverse CDF method is a problem for gaussian random numbers (any many others), because there is no closed-form expression for the CDF of a gaussian (or for the CDF<sup>-1</sup>):

$$\text{CDF}(a) = \int_{-\infty}^a dx \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \quad \text{(gaussian).}$$

But [Knu] describes a clever way based on polar coordinates to use two uniform random numbers to generate a gaussian. He gives the details, but the final result is this:

$$\text{gaussian} = (\sqrt{-2 \ln u}) \cos \theta \quad \text{where } \theta \text{ is uniform on } (0, 2\pi)$$

$u$  is uniform on  $(0,1)$

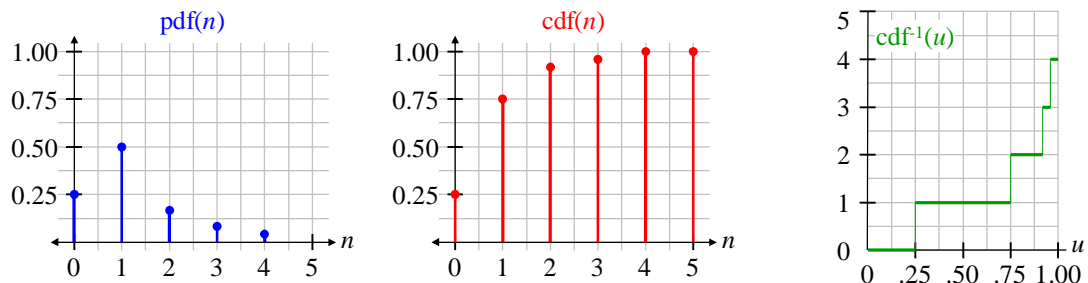
```

/* Gaussian random value, 0 mean, unit variance. From Knuth, "The Art of
Computer Programming, Vol. 2: Seminumerical Algorithms," 2nd Ed., p. 117.
It is exactly normal if rand_uniform() is uniform. */
PUBLIC double rand_gauss(void)
{ double theta = (2.*M_PI) * rand_uniform();
  return sqrt( -2. * log(rand_uniform()) ) * cos(theta);
} // rand_gauss()

```

## Generating Poisson Random Numbers

Poisson random numbers are integers; we say the Poisson distribution is **discrete**:



### Example of generating the (discrete) Poisson distribution.

We can still use the inverse-cdf method to generate them, but in an iterative way. The code starts with a helper function, `poisson()`, that compute the probability of exactly  $n$  events in a Poisson distribution with an average of  $avg$  events:

```

// -----
PUBLIC vflt poisson( // Pr(exactly n events in interval)
  vflt avg, // average events in interval
  int n) // n to compute Pr() of
{ vflt factorial;
  int i;

  if(n <= 20) factorial = fact[n];
  else
  { factorial = fact[20];
    for(i = 21; i <= n; i++) factorial *= i;
  }
  return exp(-avg) * pow(avg, n) / factorial;
} // poisson()

/*-----
Generates a Poisson random value (an integer), which must be <= 200.
Prefix 'irand_...' emphasizes the discreteness of the Poisson distribution.
-----*/
PUBLIC int irand_poisson( // Poisson random integer <= 200
  double avg) // avg # "events"
{
  int i;
  double cpr; // uniform probability

  // Use inverse-cdf(uniform) for Poisson distribution, where
  // inverse-cdf() consists of flat, discontinuous steps
  cpr = rand_uniform();
  for(i = 0; i <= 200; i++) // safety limit of 200
  { cpr -= poisson(avg, i);

```

```

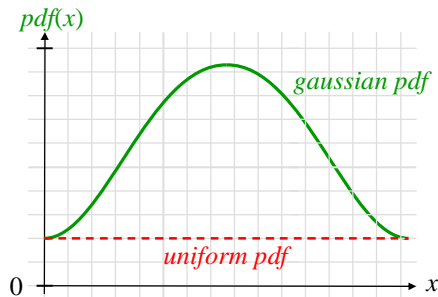
        if(cpr <= 0) break;
    }
    return i; // 201 indicates an error
} // irand_poisson()

```

Other example random number generators: TBS.

## Generating Weirder Random Numbers

Sometimes you need to generate more complex distributions, such as a combination of a gaussian with a uniform background of noise. This is a raised gaussian:



### Construction of a raised gaussian PDF random variable from a uniform and a gaussian.

Since this distribution has a uniform “component,” it is only meaningful if it’s limited to some finite “width.” To generate distributions like this, you can compose two different distributions, and use the principle:

The PDF of a random choice of two random variables is the weighted sum of the individual PDFs.

For example, the PDF for an RV (random variable) which is taken from  $X$  20% of the time, and  $Y$  the remaining 80% of the time is:

$$\text{pdf}(z) = 0.2 \text{pdf}_X(z) + 0.8 \text{pdf}_Y(z).$$

In this example, the two component distributions are uniform and gaussian. Suppose the uniform part of the pdf has amplitude 0.1 over the interval (0, 2). Then it accounts for 0.2 of all the random values. The remainder are gaussian, which we assume to be mean of 1.0, and  $\sigma = 1$ . Then the random value can be generated from three more-fundamental random values:

```

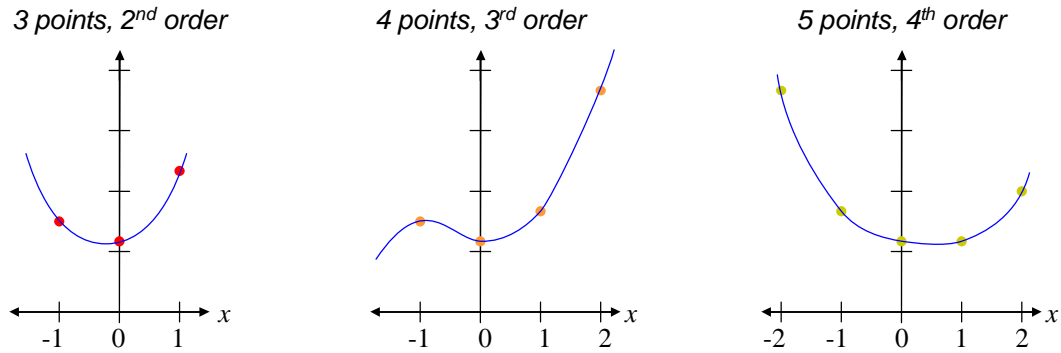
// Raised Gaussian random value: gaussian part: mean=1, sigma=1
// Uniform part (20% chance): interval (0, 2)
if(rand_uniform() <= 0.2)
    random_variable = rand_uniform()*2.0;
else
    random_variable = rand_gauss() + 1.0; // mean = 1, sigma = 1

```

---

## Exact Polynomial Fits

It’s sometimes handy to make an exact fit of a quadratic, cubic, or quartic polynomial to 3, 4, or 5 data points, respectively.



The quadratic case illustrates the principle simply. We seek a quadratic function

$$y(x) = a_2x^2 + a_1x + a_0$$

which exactly fits 3 equally spaced points, at  $x = -1$ ,  $x = 0$ , and  $x = 1$ , with value  $y_{-1}$ ,  $y_0$ , and  $y_1$ , respectively (shown above). So long as your actual data are equally spaced, you can simply scale and offset to the  $x$  values  $-1$ ,  $0$ , and  $1$ . We can directly solve for the coefficients  $a_2$ ,  $a_1$ , and  $a_0$ :

$$\left. \begin{aligned} a_2(-1)^2 + a_1(-1) + a_0 &= y_{-1} \\ a_2(0)^2 + a_1(0) + a_0 &= y_0 \\ a_2(1)^2 + a_1(1) + a_0 &= y_1 \end{aligned} \right\} \Rightarrow \left. \begin{aligned} a_2 - a_1 + a_0 &= y_{-1} \\ a_0 &= y_0 \\ a_2 + a_1 + a_0 &= y_1 \end{aligned} \right\}$$

$$\Rightarrow \quad a_2 = (y_{-1} + y_1)/2 - y_0, \quad a_1 = (y_1 - y_{-1})/2, \quad a_0 = y_0$$

Similar formulas for the 3<sup>rd</sup> and 4<sup>th</sup> order fits yield this code:

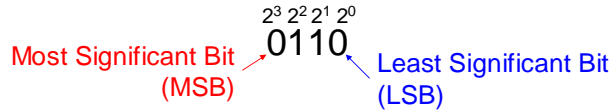
```
// -----
// fit3rd() computes 3rd order fit coefficients.  4 mult/div, 8 adds
PUBLIC void fit3rd(
    double ym1, double y0, double y1, double y2)
{
    a0 = y0;
    a2 = (ym1 + y1)/2. - y0;
    a3 = (2.*ym1 + y2 - 3.*y0)/6. - a2;
    a1 = y1 - y0 - a2 - a3;
} // fit3rd()

// -----
// fit4th() computes 4th order fit coefficients.  6 mult/div, 13 add
PUBLIC void fit4th(
    double ym2, double ym1, double y0, double y1, double y2)
{
    b0 = y0;
    b4 = (y2 + ym2 - 4*(ym1 + y1) + 6*y0)/24.;
    b2 = (ym1 + y1)/2. - y0 - b4;
    b3 = (y2 - ym2 - 2.*(y1 - ym1))/12.;
    b1 = (y1 - ym1)/2. - b3;
} // fit4th()
```

TBS: Alternative 3<sup>rd</sup> order (4 point) symmetric fit, with  $x \in \{-3, -1, 1, 3\}$ .

## Two's Complement Arithmetic

Two's complement is a way of representing negative numbers in binary. It is universally used for integers, and rarely used for floating point. This section assumes the reader is familiar with positive binary numbers and simple binary arithmetic.



Two's complement uses the most significant bit (MSB) of an integer as a sign bit: zero means the number is  $\geq 0$ ; 1 means the number is negative. Two's complement represents non-negative numbers as ordinary binary, with the sign bit = 0. Negative numbers have the sign bit = 1, but are stored in a special way: for a  $b$ -bit word, a negative number  $n$  ( $n < 0$ ) is stored as if it were unsigned with a value of  $2^b + n$ . This is shown below, using a 4-bit "word" as a simple example:

	bits	unsigned	signed
	0000	0	0
	0001	1	1
	0010	2	2
	0011	3	3
sign bit (MSB)	0100	4	4
	0101	5	5
	0110	6	6
	0111	7	7
	1000	8	-8
	1001	9	-7
	1010	10	-6
	1011	11	-5
	1100	12	-4
	1101	13	-3
	1110	14	-2
	1111	15	-1

With two's complement, a 4-bit word can store integers from  $-8$  to  $+7$ . E.g.,  $-1$  is stored as  $16 - 1 = 15$ . This rule is usually defined as follows (which completely obscures the purpose):

Let $n = -a$ $n < 0, a > 0$	Example: $n = -4, a = 4$
Start with the bit pattern for $a$	0100
complement it (change all 0s to 1s and 1s to 0s).	1011
add 1	1100

Let's see how two's complement works in practice. There are 4 possible addition cases:

(1) **Adding two positive numbers:** so long as the result doesn't overflow, we simply add normally (in binary).

(2) **Adding two negative numbers:** Recall that when adding unsigned integers, if we overflow our 4 bits, the "carries" out of the MSB are simply discarded. This means that the result of adding  $a + c$  is actually  $(a + c) \bmod 16$ . Now, let  $n$  and  $m$  be negative numbers in twos complement, so their bit patterns are  $16 + n$ , and  $16 + m$ . If we add their bit patterns as unsigned integers, we get

$$(16+n)+(16+m) = [32+(n+m)] \bmod 16 = 16+(n+m), \quad n+m < 0$$

which is the 2's complement representation of  $(n+m) < 0$ .

E.g.,

$$\begin{array}{r} -2 \quad 1110 \quad 16+(-2) \\ + \quad -3 \quad + 1101 \quad + 16+(-3) \\ \hline -5 \quad 1011 \quad 16+(-5) \end{array}$$

So with two's complement, adding negative numbers uses the same algorithm as adding unsigned integers! That's why we use two's complement.

**(3) Adding a negative and a positive number, with positive result:**

$$(16+n)+a = [16+(n+a)] \bmod 16 = n+a, \quad n+a > 0$$

E.g.,

$$\begin{array}{r} -2 \quad 1110 \quad 16+(-2) \\ + \quad 5 \quad 0101 \quad + 5 \\ \hline 3 \quad 0011 \quad 3 \end{array}$$

**(4) Adding a negative and a positive number, with negative result:**

$$(16+n)+a = 16+(n+a), \quad n+a < 0$$

E.g.,

$$\begin{array}{r} -6 \quad 1010 \quad 16+(-6) \\ + \quad 3 \quad 0011 \quad + 3 \\ \hline -3 \quad 1101 \quad 16+(-3) \end{array}$$

In all cases,

With two's complement arithmetic, adding signed integers uses the same algorithm as adding unsigned integers! That's why we use two's complement.

The computer hardware need not know which numbers are signed, and which are unsigned: it adds the same way no matter what.

It works the same with subtraction: subtracting two's complement numbers is the same as subtracting unsigned numbers. It even works multiplying to the same word size:

$$-+ : (16+n)a = [16a+(na)] \bmod 16 = 16+na, \quad n < 0, a > 0, na < 0$$

$$-- : (16+n)(16+m) = [256+16(n+m)+nm] \bmod 16 = nm, \quad n < 0, m < 0, nm > 0$$

In reality, word sizes are usually 32 (or maybe 16) bits. Then in general, we store  $b$ -bit negative numbers ( $n < 0$ ) as  $2^b + n$ . E.g., for 16 bits,  $(n < 0) \rightarrow 65536 + n$ .

## How Many Digits Do I Get, 6 or 9?

How many decimal digits of accuracy do I get with a binary floating point number? You often see a range: 6 to 9 digits. Huh? We jump ahead, and assume here that you understand binary floating point (see below for explanation).

**Wobble, but don't fall down:** The idea of "number of digits of accuracy" is somewhat flawed. Six digits of accuracy near 100,000 is ~10 times worse than 6 digits of accuracy near 999,999. The smallest increment is 1 in the least-significant digit. One in 100,000 is accuracy of  $10^{-5}$ ; 1 in 999,999 is almost  $10^{-6}$ , or 10 times more accurate.

Aside: The **wobble** of a floating point number is the ratio of the lowest accuracy to the highest accuracy for a fixed number of digits. It is always equal to the base in which the floating point number is expressed, which is 10 in this example. The wobble of binary floating point is 2. The wobble of hexadecimal floating point (mostly obsolete now) is 16.

We assume IEEE-754 compliant numbers (see later section). To insure, say, 6 decimal digits of accuracy, the worst-case binary accuracy must exceed the best-case decimal accuracy. For IEEE single-precision, there are 23 fraction bits (and one implied-1 bit), so the worst case accuracy is  $2^{-23} = 1.2 \times 10^{-7}$ . The best 6-digit accuracy is  $10^{-6}$ ; the best 7 digit accuracy is  $10^{-7}$ . Thus we see that single-precision guarantees 6 decimal digits, but almost gets 7, i.e. most of the time, it actually achieves 7 digits. The table in the next section summarizes 4 common floating point formats.

### How many digits do I need?

Often, we need to convert a binary number to decimal, write it to a file, and then read it back in, converting it back to binary. An important question is, how many decimal digits do we need to write to insure that we get back *exactly* the same binary floating point number we started with? In other words, how many binary digits do I get with a given number of decimal digits? (This is essentially the reverse of the preceding section.) We choose our number of decimal digits to insure full binary accuracy (assuming our conversion software is good, which is not always the case).

Our worst-case decimal accuracy has to exceed our best-case binary accuracy. For single precision, the best accuracy is  $2^{-24} = 6.0 \times 10^{-8}$ . The worst case accuracy of 9 decimal digits is  $10^{-8}$ , so we need 9 decimal digits to fully represents IEEE single precision. Here's a table of precisions for 4 common formats:

Format	Fraction bits	Minimum decimal digits accuracy	Decimal digits for exact replication	Decimal digits range
IEEE single	23	$2^{-23} = 1.2 \times 10^{-7} \Rightarrow 6$	$2^{-24} = 6.0 \times 10^{-8} \Rightarrow 9$	6 – 9
IEEE double	52	$2^{-52} = 2.2 \times 10^{-16} \Rightarrow 15$	$2^{-53} = 1.1 \times 10^{-16} \Rightarrow 17$	15 – 17
x86 long double	63	$2^{-63} = 1.1 \times 10^{-19} \Rightarrow 18$	$2^{-64} = 5.4 \times 10^{-20} \Rightarrow 21$	18 – 21
SPARC REAL*16	112	$2^{-112} = 1.9 \times 10^{-34} \Rightarrow 33$	$2^{-113} = 9.6 \times 10^{-35} \Rightarrow 36$	33 – 36

These number of digits agree exactly with the quoted ranges in the “IEEE Floating Point” section, and the ULP table in the underflow section. In C, then, to insure exact binary accuracy when writing, and then reading, in decimal, for double precision, use

```
printf(dec, "%.17g", x);
```

### How Far Can I Go?

A natural question is: What is the range, in decimal, of numbers that can be represented by the IEEE formats? The answer is dominated by the number of bits in the binary exponent. This table shows it:

Range and Precision of Storage Formats				
Format	Significant Bits	Smallest Normal Number	Largest Number	Decimal Digits
IEEE single	24	$1.175... \times 10^{-38}$	$3.402... \times 10^{+38}$	6-9
IEEE double	53	$2.225... \times 10^{-308}$	$1.797... \times 10^{+308}$	15-17
x86 long double	64	$3.362... \times 10^{-4932}$	$1.189... \times 10^{+4932}$	18-21
SPARC REAL*16	113	$3.362... \times 10^{-4932}$	$1.189... \times 10^{+4932}$	33-36

---

## Software Engineering

Software Engineering is much more than computer programming: it is the art and science of designing and implementing programs efficiently, over the long term, across multiple developers. Software engineering maximizes productivity and fun, and minimizes annoyance and roadblocks.



Engineers first design, then implement, systems that are useful, fun, and efficient.

Hackers just write code. Software engineering includes:

- Documentation: lots of it in the code as comments.
- Documentation: design documents that give an overview and conceptual view that is infeasible to achieve in source code comments.
- Coding guidelines: for consistency among developers. Efficiency can only be achieved by cooperation among the developers, including a consistent coding style that allows others to quickly understand the code. E.g., [physics.ucsd.edu/~emichels/Coding%20Guidelines.pdf](http://physics.ucsd.edu/~emichels/Coding%20Guidelines.pdf).
- Clean code: it is easy to read and follow.
- Maintainable code: it functions in a straightforward and comprehensible way, so that it can be changed easily and still work.

Notice that all of the above are subjective assessments. That's the nature of all engineering:

Engineering is lots of tradeoffs, with subjective approximations of the costs and benefits.

Don't get me wrong: sometimes I hack out code. The judgment comes in knowing when to hack and when to design.

#### Fun quotes:

“Whenever possible, ignore the coding standards currently in use by thousands of developers in your project's target language and environment.”

- **Roedy Green**, *How To Write Unmaintainable Code*, [www.strauss.za.com/sla/code\\_std.html](http://www.strauss.za.com/sla/code_std.html)

“Debugging is twice as hard as writing the code in the first place. Therefore, if you write the code as cleverly as possible, you are, by definition, not smart enough to debug it.” - **Brian W. Kernighan**

Coding guidelines make everyone's life easier, even yours. - **Eric L. Michelsen**

## Object Oriented Programming

This is a much used and abused term, with no definitive definition. The goal of Object Oriented Programming (OOP) is to allow reusable code that is clean and maintainable. The best definition I've seen of OOP is that it uses a language and approach with these properties:

- **User defined data types**, called **classes**, which (1) allow a single object (data entity) to have multiple data items, and (2) provide user-defined **methods** (functions and operators) for manipulating objects of that class.
- **Information hiding**: a class can define a public interface which hides the implementation details from the (client) code which uses the class.
- **Overloading**: the same named function or operator can be invoked on multiple data types, including both built-in and user-defined types. The language chooses which of the same-named functions to invoke based on the data types of its arguments.
- **Inheritance**: new data types can be created by extending existing data types. The **derived class** inherits all the data and methods of the **base class**, but can add data, and override (overload) any methods it chooses with its own, more specialized versions.
- **Polymorphism**: this is more than overloading. Polymorphism allows derived-class objects to be handled by (often older) code which only knows about the base class (i.e., which does not even know of the existence of the derived class.) Even though the application code knows nothing of the derived class, the data object itself insures calling proper specialized methods for itself.

In C++, polymorphism is implemented with virtual functions.

OOP does not have to be a new “paradigm.” It is usually more effective to make it an improvement on the good software engineering practices you already use.

## The Best of Times, the Worst of Times

We give here some ways to speed up common computations, using matrices as examples. The principles are applicable to almost any computation performed over a large amount of data.

For the vast majority of programs, execution time is so short that it doesn’t matter how efficient it is; clarity and simplicity are more important than speed.

In rare cases, time is a concern. For some simple examples, we show how to easily cut your execution times to 1/3 of original. We also show that things are not always so simple as they seem. This section assumes knowledge of computer programming with simple classes (the beginning of object oriented programming).

This topic is potentially huge, so we can only touch on some basics. The main point here is:

Computer memory management is the key to fast performance.

We proceed along these lines:

- We start with a simple C++ class for matrix addition. We give run times for this implementation (the worst of times).
- A simple improvement greatly improves execution times (the best of times).
- We try another expected improvement, but things are not as expected.
- We describe the general operation of “memory cache” (pronounced “cash”) in simple terms.
- Moving on to matrix multiplication, we find that our previous tricks don’t work well.
- However, due to the cache, adding *more* operations greatly improves the execution times.

### Matrix Addition

The basic concept in improving matrix addition is to avoid C++’s hidden copy operations. However:

Computer memory access is tricky, so things aren’t always what you’d expect. Nonetheless, we can be efficient, even without details of the computer hardware.

The tricks are due to computer hardware called RAM “cache,” whose general principles we describe later, but whose details are beyond our scope.

First, here is a simple C++ class for matrix creation, destruction, and addition. (For simplicity, our sample code has no error checking; real code, of course, does. In this case, we literally don’t want reality to interfere with science.) The class data for a matrix are the number of rows, the number of columns, and a pointer to the matrix elements (data block).

```
typedef    double T;                // matrix elements are double precision

class IMatrix    // 2D matrix
{ public:
    int    nr, nc;                // # rows & columns
    T*db;                // pointer to data

    IMatrix(int r, int c);        // create matrix of given size
    IMatrix(const IMatrix &b);    // copy constructor
    ~IMatrix();                // destructor

    T * operator [] (int r) const {return db + r*nc;};    // subscripting
```

```

    ILMatrix & operator =(const ILMatrix& b);        // assignment
    ILMatrix operator +(const ILMatrix& b) const;   // matrix add
};

```

The matrix elements are indexed starting from 0, i.e. the top-left corner of matrix 'a' is referenced as 'a[0][0]'. Following the data are the minimum set of methods (procedures) for matrix addition. Internally, the pointer 'db' points to the matrix elements (data block). The subscripting operator finds a linear array element as (row)(#columns) + column. Here is the code to create, copy, and destroy matrices:

```

// create matrix of given size (constructor)
ILMatrix::ILMatrix(int r, int c) : nr(r), nc(c)    // set nr & nc here
{
    db = new T[nr*nc];                            // allocate data block
} // ILMatrix(r, c)

// copy a matrix (copy constructor)
ILMatrix::ILMatrix(const ILMatrix & b)
{
    int    r,c;

    nr = b.nr, nc = b.nc;                        // matrix dimensions
    if(b.db)
    {
        db = new T[nr*nc];                      // allocate data block
        for(r = 0; r < nr; r++)                // copy the data
            for(c = 0; c < nc; c++)
                (*this)[r][c] = b[r][c];
    }
} // copy constructor

// destructor
ILMatrix::~ILMatrix()
{
    if(db) {delete[] db;}                       // free existing data
    nr = nc = 0, db = 0;                        // mark it empty
}

// assignment operator
ILMatrix & ILMatrix::operator =(const ILMatrix& b)
{
    int    r, c;

    for(r = 0; r < nr; r++) // copy the data
        for(c = 0; c < nc; c++)
            (*this)[r][c] = b[r][c];
    return *this;
} // operator =()

```

**The good stuff:** With the tedious preliminaries done, we now implement the simplest matrix addition method. It adds two matrices element by element, and returns the result as a new matrix:

```

// matrix addition to temporary
ILMatrix ILMatrix::operator +(const ILMatrix& b) const
{
    int    r, c;
    ILMatrix    result(nr, nc);

    for (r=0; r < nr; r++)
        for (c=0; c < nc; c++)
            result[r][c] = (*this)[r][c] + b[r][c];
    return result; // invokes copy constructor!
} // operator +()

```

How long does this simple code take? To test it, we standardize on  $300 \times 300$  and  $400 \times 400$  matrix sizes, each on two different computers: computer 1 is a circa 2000 Compaq Workstation W6000 with a 1.7 GHz Xeon. Computer 2 is a circa 2000 Gateway Solo 200 ARC laptop with a 2.4 GHz CPU. We time 100 matrix additions, e.g.:

```

int    n = 300; // matrix dimension

```

```

    ILMatrix      a(n,n), b(n,n), d(n,n);

// Addition test
d = a + b;           // prime memory caches
cpustamp("start matrix addition\n");
for(i = 0; i < 100; i++)
    d = a + b;
cpustamp("end matrix addition\n");

```

With modern operating systems, you may have to run your code several times before the execution times stabilize.

[This may be due to internal operations of allocating memory, and flushing data to disk.] We find that, on computer 1, it takes  $\sim 1.36 \pm 0.10$  s to execute 100 simple matrix additions (see table at end of this section). Wow, that seems like a long time. Each addition is 90,000 floating point adds; 100 additions is 9 million operations. Our 2.4 GHz machine should execute 2.4 additions per ns. Where's all the time going? C++ has a major flaw. Though it was pretty easy to create our matrix class:

C++ copies your data twice in a simple class operation on two values.

So besides our actual matrix addition, C++ copies the result *twice* before it reaches the matrix 'd'. The first copy happens at the 'return result' statement in our matrix addition function. Since the variable 'result' will be destroyed (go out of scope) when the function returns, C++ must copy it to a temporary variable in the main program. Notice that the C++ language has no way to tell the addition function that the result is headed for the matrix 'd'. So the addition function has no choice but to copy it into a temporary matrix, created by the compiler and *hidden* from programmers. The second copy is when the temporary matrix is assigned to the matrix 'd'. Each copy operation copies 90,000 8-byte double-precision numbers,  $\sim 720$ k bytes. That's a lot of copying.

What can we do about this? The simplest improvement is to make our copies more efficient. Instead of writing our own loops to copy data, we can call the library function `memcpy()`, which is specifically optimized for copying blocks of data. Our copy constructor is now:

```

ILMatrix::ILMatrix(const ILMatrix & b)
{   int      r,c;

    nr = b.nr, nc = b.nc;           // matrix dimensions
    if(b.db)
    {   db = new T[nr*nc];         // allocate data block
        memcpy(db, b.db, sizeof(T)*nr*nc); // copy the data
    }
} // copy constructor

```

Similarly for the assignment operator. This code takes  $0.98 \pm 0.10$  s, 28 % better than the old code. Not bad for such a simple change, but still bad: we still have two needless copies going on.

For the next improvement, we note that C++ can pass *two* matrix operands to an operator function, but not three. Therefore, if we do one copy ourselves, we can then perform the addition "in place," and avoid the second copy. For example:

```

// Faster code to implement d = a + b:
d = a;           // the one and only copy operation
d += b;          // '+=' adds 'b' to the current value of 'd'

```

The expression in parentheses copies 'a' to 'd', and evaluates as the matrix 'd', which we can then act on with the '+=' operator. We can simplify this main code to a single line as:

```
(d = a) += b;
```

To implement this code, we need to add a "+=" operator function to our class:

```

// matrix addition in-place
ILMatrix & ILMatrix::operator +=(const ILMatrix & b)
{
    int      r, c;

```

```

    for (r = 0; r < nr; r++)
        for (c = 0; c < b.nc; c++)
            (*this)[r][c] += b[r][c];

    return *this;          // returns by reference, NO copy!
}

```

This code runs in  $0.45 \pm 0.02$  s, or 1/3 the original time! The price, though, is somewhat uglier code.

Perhaps we can do even better. Instead of using operator functions, which are limited to only two matrix arguments, we can write our own addition function, *with any arguments we want*. The main code is now:

```
mat_add(d, a, b);          // add a + b, putting result in 'd'
```

Requiring the new function “mat\_add()”:

```

// matrix addition to new matrix: d = a + b
ILmatrix & mat_add(ILmatrix & d, const ILmatrix & a, const ILmatrix & b)
{
    int          r, c;

    for (r = 0; r < d.nr; r++)
        for (c = 0; c < d.nc; c++)
            d[r][c] = a[r][c] + b[r][c];

    return d;          // returned by reference, NO copy constructor
} // mat_add()

```

This runs in  $0.49 \pm 0.02$  s, slightly *worse* than the one-copy version. It’s also even uglier than the previous version. How can this be?

Memory access, including data copying, is dominated by the effects of a complex piece of hardware called “memory cache.”

There are hundreds of different variations of cache designs, and even if you know the exact design, you can rarely predict its exact effect on real code. We will describe cache shortly, but even then, there is no feasible way to know exactly why the zero-copy code is slower than one-copy. This result also held true for the  $400 \times 400$  matrix on computer 1, and the  $300 \times 300$  matrix on computer 2, but not the  $400 \times 400$  matrix on computer 2. All we can do is try a few likely cases, and go with the general trend. More on this later.

**Beware** Leaving out a single character from your code can produce code that works, but runs over 2 times slower than it should. For example, in the function definition of `mat_add`, if we leave out the “&” before argument ‘a’:

```
ILmatrix & mat_add(ILmatrix & d, const ILmatrix a, const ILmatrix & b)
```

then the compiler passes ‘a’ to the function by copying it! This completely defeats our goal of zero copy. [Guess how I found this out.]

Also notice that the ‘`memcpy()`’ optimization doesn’t apply to this last method, since it has no copies at all.

Below is a summary of matrix addition. The best code choice was a single copy, with in-place addition. It is medium ugly. While there was a small discrepancy with this on computer 2,  $400 \times 400$ , it’s not worth the required additional ugliness.

Algorithm	Computer 1 times (ms, $\pm \sim 100$ ms)		Computer 2 times (ms, $\pm \sim 100$ ms)	
	300 × 300	400 × 400	300 × 300	400 × 400
d = a + b, loop copy	1360 $\equiv$ 100 %	5900 $\equiv$ 100 %	1130 $\equiv$ 100 %	2180 $\equiv$ 100 %
d = a + b, memcpy( )	985 = 72 %	4960 = 84 %	950 = 84 %	1793 = 82 %

(d = a) += b	445 = 33 %	3850 = 65 %	330 = 29 %	791 = 36 %
mat_add(d, a, b)	490 = 36 %	4400 = 75 %	371 = 33 %	721 = 33 %

Run times for matrix addition with various algorithms. Uncertainties are very rough  $\pm 1\sigma$ . Best performing algorithms are highlighted

### Memory Consumption vs. Run Time

In the old days, the claim was clear (but not the reality): the less memory you use, the slower your algorithm, and speeding your algorithm requires more memory.

The fallacy there is that most code is not well written. When you go in to clean up your code, you often create implementations that are generally more efficient in *both* memory and time. I have personally done this many times, even when revising my own code.

However, given reasonably efficient implementations, then with no memory cache, one can usually speed the function by using an algorithm with more memory. Conversely, an algorithm that uses less memory is usually slower.

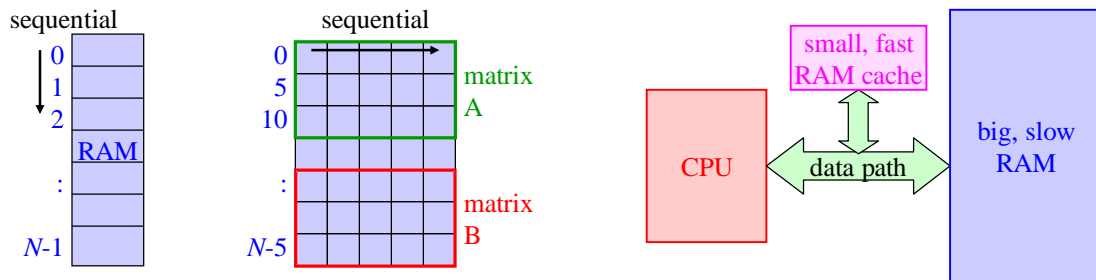
Since all modern computers use cache, there is a new factor to consider. If you “blow the cache”, i.e. your algorithm repeatedly works through more memory than the cache can hold, you will suffer many cache misses, and a dramatic slow-down in speed. In such a case, an algorithm that uses less memory may be faster: the algorithmic performance loss may be offset by the cache performance increase, possibly many times over.

### Cache Value

Before about 1990, computations were slower than memory accesses. Therefore, we optimized by increasing memory use, and decreasing computations. Today, things are exactly reversed:

Modern CPUs (c. 2009) can compute about 50 times faster than they can access main memory. Therefore, the biggest factor in overall speed is efficient use of memory.

To help reduce the speed degradation of slow memory, computers use a **memory cache**: a small memory that is very fast. A typical main memory is 1 Gb, while a typical cache is 1 Mb, or 1000x smaller. The CPU can access cache memory as fast as it can compute, so cache is  $\sim 50x$  faster than main memory. The cache is invisible to program *function*, but is critical to program *speed*. The programmer usually does not have access to details about the cache, but she can use general cache knowledge to greatly reduce run time.



(Left) Computer memory (RAM) is a linear array of bytes. (Middle) For convenience, we draw it as a 2D array, of arbitrary width. We show sample matrix storage. (Right) A very fast memory cache keeps a copy of recently used memory locations, so they can be quickly used again.

The cache does two things (diagram above):

1. Cache remembers recently used memory values, so that if the CPU requests any of them again, the cache provides the value instantly, and the slow main memory access does not happen.

- Cache “looks ahead” to fetch memory values immediately following the one just used, before the CPU might request it. If the CPU in fact later requests the next sequential memory location, the cache provides the value instantly, having already fetched it from slow main memory.

The cache is small, and eventually fills up. Then, when the CPU requests new data, the cache must discard old data, and replace it with the new. Therefore, if the program jumps around memory a lot, the benefits of the cache are reduced. If a program works repeatedly over a small region of memory (say, a few hundred k bytes), the benefits of cache increase. Typically, cache can follow four separate regions of memory concurrently. This means you can interleave accesses to four different regions of memory, and still retain the benefits of cache. Therefore, we have three simple rules for efficient memory use:

For efficient memory use: (1) access memory sequentially, or at most in small steps, (2) reuse values as much as possible in the shortest time, and (3) access few memory regions concurrently, preferably no more than four.

There is huge variety in computer memory designs, so these rules are general, and behavior varies from machine to machine, sometimes greatly. Our data below demonstrate this.

### Your Cache at Work

We can now understand some of our timing data given above. We see that the one-copy algorithm unexpectedly takes less time than the zero-copy algorithm. The one-copy algorithm accesses only two memory regions at a time: first matrix ‘a’ and ‘d’ for the copy, then matrix ‘b’ and ‘d’ for the add. The zero-copy algorithm accesses *three* regions at a time: ‘a’, ‘b’, and ‘d’. This is probably reducing cache efficiency. Recall that the CPU is also fetching instructions (the program) concurrently with the data, which is at least a fourth region. Exact program layout in memory is virtually impossible to know. Also, the cache on this old computer may not support 4-region concurrent access. The newer machine, computer 2, probably has a better cache, and the one- and zero-copy algorithms perform very similarly.

Here’s a new question for matrix addition: the code given earlier loops over rows in the outer loop, and columns in the inner loop. What if we reversed them, and looped over columns on the outside, and rows on the inside? The result is 65% longer run time, on both machines. Here’s why: the matrices are stored by rows, i.e. each row is consecutive memory locations. Looping over columns on the inside accesses memory sequentially, taking advantage of cache look-ahead. When reversed, the program jumps from row to row on the inside, giving up any benefit from look-ahead. The cost is quite substantial. This concept works on almost every machine.

**Caution** FORTRAN stores arrays in the opposite order from C and C++. In FORTRAN, the first index is cycled most rapidly, so you should code with the outer loop on the second index, and the inner loop on the first index. E.g.,

```
DO C = 1, N
  DO R = 1, N
    A(R, C) = blah blah ...
  ENDDO
ENDDO
```

**Scaling behavior:** Matrix addition is an  $O(N^2)$  operation, so increasing from  $300 \times 300$  to  $400 \times 400$  increase the computations by a factor of 1.8. On the older computer 1, the runtime penalty is much larger, between 4.5x and 9x slower. On the newer computer 2, the difference is much closer, between 1.8x and 2.2x slower. This is likely due to cache size. A  $300 \times 300$  double precision matrix takes 720 k bytes, or under a MB. A  $400 \times 400$  matrix takes 1280 k bytes, just *over* one MB. It could be that on computer 1, with the smaller matrix, a whole matrix or two fits in cache, but with the large matrix, cache is overflowed, and more (slow) main memory accesses are needed. The newer computer probably has bigger caches, and may fit both sized matrices fully in cache.

### Cache Withdrawal: Matrix Multiplication

We now show that the above tricks don’t work well for large-matrix multiplication, but a different trick cuts multiplication run time dramatically. To start, we use a simple matrix multiply in the main code:

```
d = a * b;
```

The straightforward matrix multiply operator is this:

```
// matrix multiply to temporary
ILmatrix ILmatrix::operator *(const ILmatrix & b) const
{
    int          r, c, k;
    ILmatrix     result(nr, b.nc);          // temporary for result
    T            sum;

    for(r = 0; r < nr; r++)
    { for(c = 0; c < b.nc; c++)
      { sum = 0.;
        for(k = 0; k < nc; k++) sum += (*this)[r][k] * b[k][c];
        result[r][c] = sum;
      }
    } return result;          // invokes copy constructor!
} // operator *()
```

While matrix addition is an  $O(N^2)$  operation, matrix multiplication is an  $O(N^3)$  operation. Multiplying two 300 x 300 matrices is about 54,000,000 floating point operations, which is *much* slower than addition. Timing the simple multiply routine, similarly to timing matrix addition, but with only 5 multiplies, we find it takes  $7.8 \pm 0.1$  s on computer 1.

First we try the tricks we already know to improve and avoid data copies: we started already with `memcpy()`. We compare the two-copy, one-copy, and zero-copy algorithms as with addition, but this time, 5 of the 6 trials show no measurable difference. Matrix multiply is so slow that the copy times are insignificant. The one exception is the one-copy algorithm on computer 2, which shows a significant reduction of ~35%. This is almost certainly due to some quirk of memory layout and the cache, but we can't identify it precisely. However, if we have to choose from these 3 algorithms, we choose the one-copy (which coincidentally agrees with the matrix addition favorite). And certainly, we drop the ugly 3-argument `mat_mult()` function, which gives no benefit.

Now we'll improve our matrix multiply greatly, by *adding more work* to be done. The extra work will result in more efficient memory use, that pays off handsomely in reduced runtime. Notice that in matrix multiplication, for each element of the results, we access a row of the first matrix *a*, and a column of the second matrix *b*. But we learned from matrix addition that accessing a column is much slower than accessing a row. And in matrix multiplication, we have to access the same column *N* times. Extra bad. If only we could access *both* matrices by rows!

Well, we can. We first make a temporary copy of matrix *b*, and transpose it. Now the columns of *b* become the *rows* of  $b^T$ . We perform the multiply as rows of *a* with *rows* of  $b^T$ . We've already seen that copy time is insignificant for multiplication, so the cost of one copy and one transpose (similar to a copy) is negligible. But the benefit of cache look-ahead is large. The transpose method reduces runtime by 30% to 50%.

Further thought reveals that we only need *one* column of *b* at a time. We can use it *N* times, and discard it. Then move on to the next column of *b*. This reduces memory usage, because we only need extra storage for one column of *b*, not for the whole transpose of *b*. It costs us nothing in operations, and reduces memory. That can only help our cache performance. In fact, it cuts runtime by about another factor of two, to about one third of the original runtime, on both machines. (It does require us to loop over columns of *b* on the outer loop, and rows of *a* on the inner loop, but that's no burden.)

Note that optimizations that at first were insignificant, say reducing runtime by 10%, may become significant after the runtime is cut by a factor of 3. That original 10% is now 30%, and may be worth doing.

Algorithm	Computer-1 times (ms, ± ~ 100 ms)		Computer-2 times (ms, ± ~ 100 ms)	
	300 × 300	400 × 400	300 × 300	400 × 400
d = a * b	7760 ≡ 100 %	18,260 ≡ 100 %	5348 ≡ 100 %	16,300 ≡ 100 %
(d = a) *= b	7890 = 102 %	18,210 = 100 %	3485 = 65 %	11,000 = 67 %



mat_mult(d, a, b)	7720 = 99 %	18,170 = 100 %	5227 = 98 %	16,200 = 99 %
d = a * b, transpose 'b'	4580 = 59 %	12,700 = 70 %	2900 = 54 %	7800 = 48%
(d = a) *= b, transpose 'b'	4930 = 64 %	12,630 = 69 %	4250 = 79 %	11,000 = 67 %
d = a * b, copy 'b' column	<b>2710 = 35 %</b>	<b>7875 = 43 %</b>	3100 = 58 %	8000 = 49 %
(d = a) *= b, copy 'b' column	2945 = 38 %	7835 = 43 %	<b>2100 = 39 %</b>	<b>5400 = 33 %</b>

Run times for matrix multiplication with various algorithms. Uncertainties are very rough  $\pm 1\sigma$ . Best performing algorithms are highlighted

### Cache Summary

In the end, exact performance is nearly impossible to predict. However, general knowledge of cache, and following the three rules for efficient cache use (given above), will greatly improve your runtimes.

Conflicts in memory between pieces of data and instruction cannot be precisely controlled. Sometimes even tiny changes in code will cross a threshold of cache, and cause huge changes in performance.

### IEEE Floating Point Formats And Concepts

Much of this section is taken from [http://docs.sun.com/source/806-3568/ncg\\_math.html](http://docs.sun.com/source/806-3568/ncg_math.html) , an excellent article introducing IEEE floating point. However, many clarifications are made here.

#### What Is IEEE Arithmetic?

In brief, IEEE 754 specifies exactly how floating point operations are to occur, and to what precision. It does *not* specify how the floating point numbers are stored in memory. Each computer makes its own choice for how to store floating point numbers. We give some popular formats later.

In particular, IEEE 754 specifies a binary floating point standard, with:

- Two basic floating-point formats: **single** and **double**.
- The IEEE single format has a significand (aka **mantissa**) precision of 24 bits, and is 32 bits overall. The IEEE double format has a significand precision of 53 bits, and is 64 bits overall.
- Two classes of *extended* floating-point formats: **single extended** and **double extended**. The standard specifies only the *minimum* precision and size. For example, an IEEE double extended format must have a significand precision of at least 64 bits and occupy at least 79 bits overall.
- Accuracy requirements on floating-point operations: add, subtract, multiply, divide, square root, remainder, round numbers in floating-point format to integer values, convert between different floating-point formats, convert between floating-point and integer formats, and compare. The remainder and compare operations must be exact. Other operations must minimally modify the exact result according to prescribed rounding modes.
- Accuracy requirements for conversions between decimal strings and binary floating-point numbers. Within specified ranges, these conversions must be exact, if possible, or minimally modify such exact results according to prescribed rounding modes. Outside the specified ranges, these conversions must meet a specified tolerance that depends on the rounding mode.
- Five types of floating-point exceptions, and the conditions for the occurrence of these exceptions. The five exceptions are *invalid operation*, *division by zero*, *overflow*, *underflow*, and *inexact*.

- Four rounding directions: *toward the nearest representable value*, with "even" values preferred whenever there are two nearest representable values; *toward negative infinity* (down); *toward positive infinity* (up); and *toward 0* (chop).
- Rounding precision; for example, if a system delivers results in double extended format, the user should be able to specify that such results be rounded to either single or double precision.

The IEEE standard also *recommends* support for user handling of exceptions.

IEEE 754 floating-point arithmetic offers users great control over computation. It simplifies the task of writing numerically sophisticated, portable programs not only by imposing rigorous requirements, but also by allowing implementations to provide refinements and enhancements to the standard.

**Storage Formats**

The IEEE floating-point formats define the fields that compose a floating-point number, the bits in those fields, and their arithmetic interpretation, but not how those formats are stored in memory. A storage format specifies how a number is stored in memory.

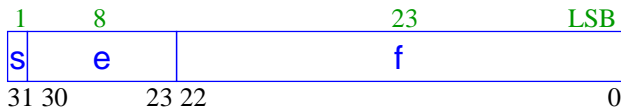
Each computer defines its own storage formats, though they are obviously all related.

High level languages have different names for floating point data types, which usually correspond to the IEEE formats as shown:

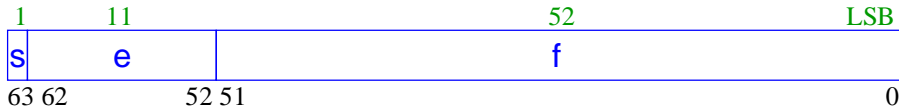
IEEE Formats and Language Types		
IEEE Precision	C, C++	Fortran
single	float	REAL <i>or</i> REAL*4
double	double	DOUBLE PRECISION <i>or</i> REAL*8
double extended	long double	
double extended		REAL*16 [e.g., SPARC]. Note that in many implementations, REAL*16 is <i>different</i> than 'long double'

IEEE 754 specifies exactly the single and double floating-point formats, and it defines ways to extend each of these two basic formats. The long double and REAL\*16 types shown above are two double extended formats compliant with the IEEE standard.

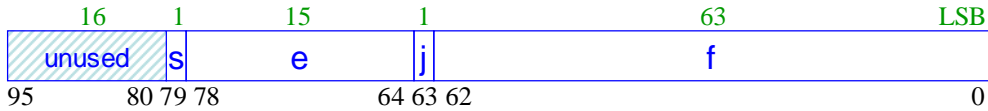
**Single (6-9 decimal digits)**



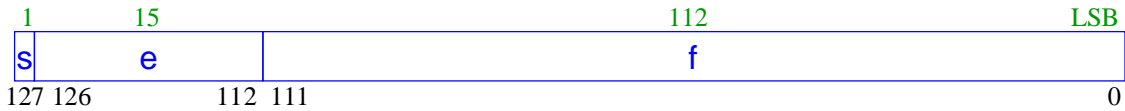
**Double (15-17 decimal digits)**



**Double-Extended (long double) (x86) (18-21 decimal digits)**



**Double-Extended (SPARC) (33-36 decimal digits)**



The following sections describe each of the floating-point *storage* formats on SPARC and x86 platforms.

**When a Bias Is a Good Thing**

IEEE floating point uses **biased exponents**, where the actual exponent is the unsigned value of the ‘e’ field minus a constant, called a bias:

$$\text{exponent} = e - \text{bias}$$

The bias makes the ‘e’ field an unsigned integer, and smallest numbers have the smallest ‘e’ field (as well as the smallest exponent). This format allows (1) floating point numbers sort in the same order as if their bit patterns were integers; and (2) true floating point zero is naturally represented by an all-zero bit pattern. These might seem insignificant, but they are quite useful, and so biased exponents are nearly universal.

**Single Format**

The IEEE single format consists of three fields: a 23-bit fraction, *f*; an 8-bit biased exponent, *e*; and a 1-bit sign, *s*. These fields are stored contiguously in one 32-bit word, as shown above.

The table below shows the three constituent fields *s*, *e*, and *f*, and the value represented in single-format:

Single-Format Fields	Value
$1 \leq e \leq 254$	$(-1)^s \times 2^{-127} \times 1.f$ (normal numbers)
$e = 0; f \neq 0$ (at least one bit in <i>f</i> is nonzero)	$(-1)^s \times 2^{-126} \times 0.f$ (denormalized numbers)
$e = 0; f = 0$ (all bits in <i>f</i> are zero)	$(-1)^s \times 0.0$ (signed zero)
$s = 0/1; e = 255; f = 0$ (all bits in <i>f</i> are zero)	$+/- \infty$ (infinity)

$s = \text{either}; e = 255; f \neq 0$ (at least one bit in $f$ is nonzero)	NaN (Not-a-Number)
---	--------------------

Notice that when  $1 \leq e \leq 254$ , the value is formed by inserting the binary radix point to the left of the fraction's most significant bit, and inserting an implicit 1-bit to the left of the binary point, thus representing a whole number plus fraction, called the *significand*, where  $1 \leq \text{significand} < 2$ . The implicit bit's value is not explicitly given in the single-format bit pattern, but is implied by the biased exponent field.

A **denormalized** number (aka **subnormal** number) is one which is too small to be represented by an exponent in the range  $1 \leq e \leq 254$ . The difference between a **normal** number and a denormalized number is that the bit to left of the binary point of a normal number is 1, but that of a denormalized number is 0.

The 23-bit fraction combined with the implicit leading significand bit provides 24 bits of precision in single-format normal numbers.

Examples of important bit patterns in the single-storage format are shown below. The maximum positive normal number is the largest finite number representable in IEEE single format. The minimum positive denormalized number is the smallest positive number representable in IEEE single format. The minimum positive normal number is often referred to as the underflow threshold. (The decimal values are rounded to the number of figures shown.)

Important Bit Patterns in IEEE Single Format		
Common Name	Bit Pattern (Hex)	Approximate Value
+0	0000 0000	0.0
-0	8000 0000	-0.0
1	3f80 0000	1.0
2	4000 0000	2.0
maximum normal number	7f7f ffff	3.40282347e+38
minimum positive normal number	0080 0000	1.17549435e-38
maximum subnormal number	007f ffff	1.17549421e-38
minimum positive subnormal number	0000 0001	1.40129846e-45
$+\infty$	7f80 0000	$+\infty$ (positive infinity)
$-\infty$	ff80 0000	$-\infty$ (negative infinity)
Not-a-Number (NaN)	7fc0 0000 (e.g.)	NaN

A NaN (Not a Number) can be represented with many bit patterns that satisfy the definition of a NaN; the value of the NaN above is just one example.

**Double Format**

The IEEE double format is the obvious extension of the single format, and also consists of three fields: a 52-bit fraction,  $f$ ; an 11-bit biased exponent,  $e$ ; and a 1-bit sign,  $s$ . These fields are stored in two consecutive 32-bit words. In the SPARC architecture, the higher address 32-bit word contains the least

significant 32 bits of the fraction, while in the x86 architecture the lower address 32-bit word contains the least significant 32 bits of the fraction.

The table below shows the three constituent fields *s*, *e*, and *f*, and the value represented in double-format:

Double-Format Fields	Value
$1 \leq e \leq 2046$	$(-1)^s \times 2^{-1023} \times 1.f$ (normal numbers)
$e = 0; f \neq 0$ (at least one bit in <i>f</i> is nonzero)	$(-1)^s \times 2^{-1022} \times 0.f$ (denormalized numbers)
$e = 0; f = 0$ (all bits in <i>f</i> are zero)	$(-1)^s \times 0.0$ (signed zero)
$s = 0/1; e = 2047; f = 0$ (all bits in <i>f</i> are zero)	$\pm \infty$ (infinity)
$s = \text{either}; e = 2047; f \neq 0$ (at least one bit in <i>f</i> is 1)	NaN (Not-a-Number)

This is the obvious analog of the single format, and retains the implied 1-bit in the significand. The 52-bit fraction combined with the implicit leading significand bit provides 53 bits of precision in double-format normal numbers.

Below, the 2<sup>nd</sup> column has two hexadecimal numbers. For the SPARC architecture, the left one is the lower addressed 32-bit word; for the x86 architecture, the left one is the higher addressed word. The decimal values are rounded to the number of figures shown.

Important Bit Patterns in IEEE Double Format		
Common Name	Bit Pattern (Hex)	Approximate Value
+ 0	00000000 00000000	0.0
- 0	80000000 00000000	-0.0
1	3ff00000 00000000	1.0
2	40000000 00000000	2.0
max normal number	7fefffff ffffffff	1.797 693 134 862 3157e+308
min positive normal number	00100000 00000000	2.225 073 858 507 2014e-308
max denormalized number	000fffff ffffffff	2.225 073 858 507 2009e-308
min positive denormalized number	00000000 00000001	4.940 656 458 412 4654e-324
+ ∞	7ff00000 00000000	+ ∞ (positive infinity)
- ∞	fff00000 00000000	- ∞ (negative infinity)
Not-a-Number	7ff80000 00000000 (e.g.)	NaN

A NaN (Not a Number) can be represented with many bit patterns that satisfy the definition of a NaN; the value of the NaN above is just one example.

**Double-Extended Format (SPARC)**

The SPARC floating-point quadruple-precision format conforms to the IEEE definition of double-extended format. The quadruple-precision format occupies four 32-bit words and consists of three fields: a 112-bit fraction, *f*; a 15-bit biased exponent, *e*; and a 1-bit sign, *s*. These fields are stored contiguously. The lowest addressed word has the sign, exponent, and the 16 most significant bits of the fraction. The highest addressed 32-bit word contains the least significant 32-bits of the fraction.

Below shows the three constituent fields and the value represented in quadruple-precision format.

Double-Extended Fields (SPARC)	Value
$1 \leq e \leq 32766$	$(-1)^s \times 2^{-16383} \times 1.f$ (normal numbers)
$e = 0, f \neq 0$ (at least one bit in <i>f</i> is nonzero)	$(-1)^s \times 2^{-16382} \times 0.f$ (denormalized numbers)
$e = 0, f = 0$ (all bits in <i>f</i> are zero)	$(-1)^s \times 0.0$ (signed zero)
$s = 0/1, e = 32767, f = 0$ (all bits in <i>f</i> are zero)	$\pm \infty$ (infinity)
$s = \text{either}, e = 32767, f \neq 0$ (at least one bit in <i>f</i> is 1)	NaN (Not-a-Number)

In the hex digits below, the left-most number is the lowest addressed 32-bit word.

Important Bit Patterns in IEEE Double-Extended Format (SPARC)					
Name	Bit Pattern (SPARC, hex)				Approximate Value
+0	00000000	00000000	00000000	00000000	0.0
-0	80000000	00000000	00000000	00000000	-0.0
1	3fff0000	00000000	00000000	00000000	1.0
2	40000000	00000000	00000000	00000000	2.0
max normal	7ffeffff	ffffffff	ffffffff	ffffffff	1.189 731 495 357 231 765 085 759 326 628 0070 e+4932
min normal	00010000	00000000	00000000	00000000	3.362 103 143 112 093 506 262 677 817 321 7526 e-4932
max subnormal	0000ffff	ffffffff	ffffffff	ffffffff	3.362 103 143 112 093 506 262 677 817 321 7520 e-4932
min pos subnormal	00000000	00000000	00000000	00000001	6.475 175 119 438 025 110 924 438 958 227 6466 e-4966
+ ∞	7fff0000	00000000	00000000	00000000	+ ∞

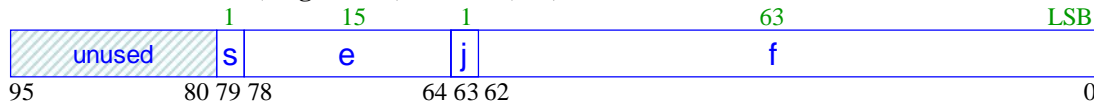
- ∞	ffff0000 00000000 00000000 00000000	- ∞
Not-a-Number	7fff8000 00000000 00000000 00000000 (e.g.)	NaN

**Double-Extended Format (x86)**

The important difference in the x86 long-double format is the lack of an implicit leading 1-bit in the significand. Instead, the 1-bit is explicit, and always present in normalized numbers. This clearly violates the spirit of the IEEE standard. However, big companies carry a lot of clout with standards bodies, so Intel claims this double-extended format conforms to the IEEE definition of double-extended formats, because IEEE 754 does not specify how (or if) the leading 1-bit is stored. X86 long-double consists of four fields: a 63-bit fraction, *f*; a 1-bit explicit leading significand bit, *j*; a 15-bit biased exponent, *e*; and a 1-bit sign, *s* (note the additional *j* field as the explicit leading bit).

In the x86 architectures, these fields are stored contiguously in ten successively addressed 8-bit bytes. However, the UNIX System V Application Binary Interface Intel 386 Processor Supplement (Intel ABI) requires that double-extended parameters and results occupy three consecutive 32-bit words in the stack, with the most significant 16 bits of the highest addressed word being unused, as shown below.

**Double-Extended (long double) Format (x86)**



The lowest addressed 32-bit word contains the least significant 32 bits of the fraction, *f*[31:0], with bit 0 being the least significant bit of the entire fraction. Though the upper 16 bits of the highest addressed 32-bit word are unused by x86, they are essential for conformity to the Intel ABI, as indicated above.

Below shows the four constituent fields and the value represented by the bit pattern. *x* = don't care.

Double-Extended Fields (x86)	Value
<i>j</i> = 0, 1 <= <i>e</i> <= 32766	Unsupported
<i>j</i> = 1, 1 <= <i>e</i> <= 32766	$(-1)^s \times 2^{e-16383} \times 1.f$ (normal numbers)
<i>j</i> = 0, <i>e</i> = 0; <i>f</i> ≠ 0 (at least one bit in <i>f</i> is nonzero)	$(-1)^s \times 2^{-16382} \times 0.f$ (denormalized numbers)
<i>j</i> = 1, <i>e</i> = 0	$(-1)^s \times 2^{-16382} \times 1.f$ (pseudo-denormal numbers)
<i>j</i> = 0, <i>e</i> = 0, <i>f</i> = 0 (all bits in <i>f</i> are zero)	$(-1)^s \times 0.0$ (signed zero)
<i>j</i> = 1; <i>s</i> = 0/1; <i>e</i> = 32767; <i>f</i> = 0 (all bits in <i>f</i> are zero)	+/- ∞ (infinity)
<i>j</i> = 1; <i>s</i> = <i>x</i> ; <i>e</i> = 32767; <i>f</i> = .1xxx...xx	QNaN (quiet NaNs)
<i>j</i> = 1; <i>s</i> = <i>x</i> ; <i>e</i> = 32767; <i>f</i> = .0xxx...xx ≠ 0 (at least one of the <i>x</i> in <i>f</i> is 1)	SNaN (signaling NaNs)

Notice that bit patterns in x86 double-extended format do *not* have an *implicit* leading significand bit. The leading significand bit is given *explicitly* as a separate field, *j*. However, when *e* ≠ 0, any bit pattern

with  $j = 0$  is unsupported: such a bit pattern as an operand in floating-point operations provokes an invalid operation exception.

The union of the fields  $j$  and  $f$  in the double extended format is called the *significand*. The significand is formed by inserting the binary radix point between the leading bit,  $j$ , and the fraction's most significant bit.

In the x86 double-extended format, a bit pattern whose leading significand bit  $j$  is 0 and whose biased exponent field  $e$  is also 0 represents a denormalized number, whereas a bit pattern whose leading significand bit  $j$  is 1 and whose biased exponent field  $e$  is nonzero represents a normal number. Because the leading significand bit is represented explicitly rather than being inferred from the exponent, this format also admits bit patterns whose biased exponent is 0, like the subnormal numbers, but whose leading significand bit is 1. Each such bit pattern actually represents the same value as the corresponding bit pattern whose biased exponent field is 1, i.e., a normal number, so these bit patterns are called *pseudo-denormals*. Pseudo-denormals are merely an artifact of the x86 double-extended storage format; they are implicitly converted to the corresponding normal numbers when they appear as operands, and they are never generated as results.

Below are some important bit patterns in the double-extended storage format. The 2<sup>nd</sup> column has three hex numbers. The first number is the 16 least significant bits of the highest addressed 32-bit word (recall that the upper 16 bits of this 32-bit word are unused), and the right one is the lowest addressed 32-bit word.



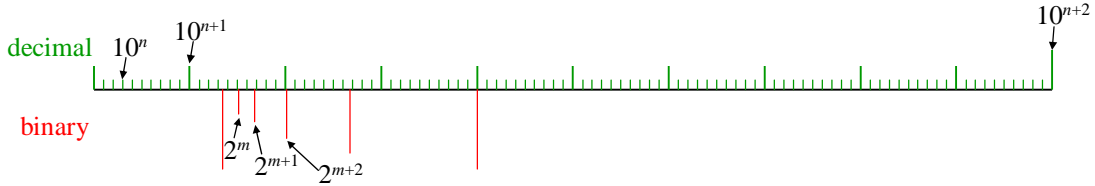
Important Bit Patterns in Double-Extended (x86) Format and their Values		
Common Name	Bit Pattern (x86)	Approximate Value
+0	0000 00000000 00000000	0.0
-0	8000 00000000 00000000	-0.0
1	3fff 80000000 00000000	1.0
2	4000 80000000 00000000	2.0
max normal	7ffe ffffffff ffffffff	1.189 731 495 357 231 765 05 e+4932
min positive normal	0001 80000000 00000000	3.362 103 143 112 093 506 26 e-4932
max subnormal	0000 7fffffff ffffffff	3.362 103 143 112 093 506 08 e-4932
min positive subnormal	0000 00000000 00000001	3.645 199 531 882 474 602 53 e-4951
+ ∞	7fff 80000000 00000000	+ ∞
- ∞	ffff 80000000 00000000	- ∞
quiet NaN with greatest fraction	7fff ffffffff ffffffff	QNaN
quiet NaN with least fraction	7fff c0000000 00000000	QNaN
signaling NaN with greatest fraction	7fff bfffffff ffffffff	SNaN
signaling NaN with least fraction	7fff 80000000 00000001	SNaN

A NaN (Not a Number) can be represented by any of the bit patterns that satisfy the definition of NaN. The most significant bit of the fraction field determines whether a NaN is quiet (bit = 1) or signaling (bit = 0).

### Precision in Decimal Representation

This section covers the precisions of the IEEE single and double formats, and the double-extended formats on SPARC and x86. See the earlier section on *How Many Digits Do I Get?* for more information.

The IEEE standard specifies the set of numerical values representable in a binary format. Each format has some number of bits of precision (e.g., single has 24 bits). But the decimal numbers of roughly the same precision do not match exactly the binary numbers, as you can see on the number line:



Comparison of a Set of Numbers Defined by Decimal and Binary Representation

Because the decimal numbers are different than the binary numbers, estimating the number of significant decimal digits corresponding to *b* significant binary bits requires some definition.

Reformulate the problem in terms of converting floating-point numbers between binary and decimal. You might convert from decimal to binary and back to decimal, or from binary to decimal and back to binary. It is important to notice that because the sets of numbers are different, conversions are in general **inexact**. If done correctly, converting a number from one set to a number in the other set results in choosing one of the two neighboring numbers from the second set (which one depends on rounding).

All binary numbers can be represented exactly in decimal, but usually this requires unreasonably many digits to do so. What really matters is how many decimal digits are *needed*, to insure no loss in converting from binary to decimal and back to binary.

Most decimal numbers cannot be represented exactly in binary (because decimal fractions include a factor of 5, which requires infinitely repeating binary digits). For example, run the following Fortran program:

```

REAL Y, Z
Y = 838861.2
Z = 1.3
WRITE(*,40) Y
40  FORMAT("y: ",1PE18.11)
WRITE(*,50) Z
50  FORMAT("z: ",1PE18.11)
    
```

The output should resemble:

```

y: 8.38861187500E+05
z: 1.29999995232E+00
    
```

The difference between the value  $8.388612 \times 10^5$  assigned to *y* and the value printed out is 0.0125, which is seven decimal orders of magnitude smaller than *y*. So the accuracy of representing *y* in IEEE single format is about 6 to 7 significant digits, or *y* has about 6 **significant digits**.

Similarly, the difference between the value 1.3 assigned to *z* and the value printed out is 0.00000004768, which is eight decimal orders of magnitude smaller than *z*. The accuracy of representing *z* in IEEE single format is about 7 to 8 significant digits, or *z* has about 7 significant digits.

See [Appendix F](http://docs.sun.com/source/806-3568/ncg_references.html) of [http://docs.sun.com/source/806-3568/ncg\\_references.html](http://docs.sun.com/source/806-3568/ncg_references.html) for references on base conversion. They say that particularly good references are Coonen's thesis and Sterbenz's book.

**Underflow**

**Underflow** occurs, roughly speaking, when the result of an arithmetic operation is so small that it cannot be stored in its intended destination format without suffering a rounding error that is larger than usual; in other words, when the result is smaller than the smallest normal number.

Underflow Thresholds in Each Precision		
single	smallest normal number	1.175 494 35e-38
	largest subnormal number	1.175 494 21e-38

double	smallest normal number largest subnormal number	2.225 073 858 507 201 4e-308 2.225 073 858 507 200 9e-308
double extended (x86)	smallest normal number largest subnormal number	3.362 103 143 112 093 506 26e-4932 3.362 103 143 112 093 505 90e-4932
double extended (SPARC)	smallest normal number largest subnormal number	3.362 103 143 112 093 506 262 677 817 321 752 6e-4932 3.362 103 143 112 093 506 262 677 817 321 752 0e-4932

The positive subnormal numbers are those numbers between the smallest normal number and zero. Subtracting two (positive) tiny numbers that are near the smallest normal number might produce a subnormal number. Or, dividing the smallest positive normal number by two produces a subnormal result.

The presence of subnormal numbers provides greater precision to floating-point calculations that involve small numbers, although the subnormal numbers themselves have fewer bits of precision than normal numbers. **Gradual underflow** produces subnormal numbers (rather than returning the answer zero) when the mathematically correct result has magnitude less than the smallest positive normal number.

There are several other ways to deal with such **underflow**. One way, common in the past, was to flush those results to zero. This method is known as **Store 0** and was the default on most mainframes before the advent of the IEEE Standard.

The mathematicians and computer designers who drafted IEEE Standard 754 considered several alternatives, while balancing the desire for a mathematically robust solution with the need to create a standard that could be implemented efficiently.

### How Does IEEE Arithmetic Treat Underflow?

[IEEE Standard 754 requires gradual underflow](#). This method requires defining two representations for stored values, normal and subnormal.

Recall that the IEEE value for a normal floating-point number is:  $(-1)^s \times 2^{e-bias} \times 1.f$

where  $s$  is the sign bit,  $e$  is the biased exponent, and  $f$  is the fraction. Only  $s$ ,  $e$ , and  $f$  need to be stored to fully specify the number. Because the leading bit of the significand is 1 for normal numbers, it need not be stored (but may be).

The smallest positive normal number that can be stored, then, has the negative exponent of greatest magnitude and a fraction of all zeros. Even smaller numbers can be accommodated by considering the leading bit to be zero rather than one. In the double-precision format, this effectively extends the minimum exponent from  $10^{-308}$  to  $10^{-324}$ , because the fraction part is 52 bits long (roughly 16 decimal digits.) These are the **subnormal** numbers; returning a subnormal number (rather than flushing an underflowed result to zero) is **gradual underflow**.

Clearly, the smaller a subnormal number, the fewer nonzero bits in its fraction; computations producing subnormal results do not enjoy the same bounds on *relative* roundoff error as computations on normal operands. However, the key fact is:

Gradual underflow implies that underflowed results never suffer a loss of accuracy any greater than that which results from ordinary roundoff error.

Addition, subtraction, comparison, and remainder are always exact when the result is very small.

Recall that the IEEE value for a subnormal floating-point number is:  $(-1)^s \times 2^{-bias+1} \times 0.f$

where  $s$  is the sign bit, the biased exponent  $e$  is zero, and  $f$  is the fraction. Note that the implicit power-of-two bias is one greater than the bias in the normal format, and the leading bit of the fraction is zero.

Gradual underflow allows you to extend the lower range of representable numbers. It is not *smallness* that renders a value questionable, but its associated error. Algorithms exploiting subnormal numbers have

smaller error bounds than other systems. The next section provides some mathematical justification for gradual underflow.

**Why Gradual Underflow?**

The purpose of subnormal numbers is not to avoid underflow/overflow entirely, as some other arithmetic models do. Rather, subnormal numbers eliminate underflow as a cause for concern for a variety of computations (typically, multiply followed by add). For a more detailed discussion, see *Underflow and the Reliability of Numerical Software* by James Demmel, and *Combatting the Effects of Underflow and Overflow in Determining Real Roots of Polynomials* by S. Linnainmaa.

The presence of subnormal numbers in the arithmetic means that untrapped underflow (which implies loss of accuracy) cannot occur on addition or subtraction. If  $x$  and  $y$  are within a factor of two, then  $x - y$  is error-free. This is critical to a number of algorithms that effectively increase the working precision at critical places in algorithms.

In addition, gradual underflow means that errors due to underflow are no worse than usual roundoff error. This is a much stronger statement than can be made about any other method of handling underflow, and this fact is one of the best justifications for gradual underflow.

**Error Properties of Gradual Underflow**

Most of the time, floating-point results are rounded:

$$computed\ result = true\ result + roundoff$$

How large can the roundoff be? One convenient measure of its size is called a *unit in the last place*, abbreviated **ulp**. The least significant bit of the fraction of a floating-point number is its *last place*. The value represented by this bit (e.g., the absolute difference between the two numbers whose representations are identical except for this bit) is a *unit in the last place* of that number. If the true result is rounded to the nearest representable number, then clearly the roundoff error is no larger than half a unit in the last place of the computed result. In other words, in IEEE arithmetic with rounding mode to nearest,

$$0 \leq |roundoff| \leq 1/2\ ulp$$

of the computed result.

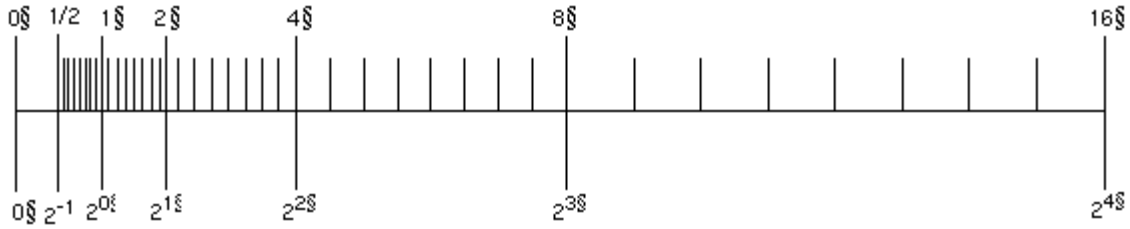
Note that an ulp is a relative quantity. An ulp of a very large number is itself very large, while an ulp of a tiny number is itself tiny. This relationship can be made explicit by expressing an ulp as a function:  $ulp(x)$  denotes a unit in the last place of the floating-point number  $x$ .

Moreover, an ulp of a floating-point number depends on the floating point precision. For example, this shows the values of  $ulp(1)$  in each of the four floating-point formats described above:

<b>ulp(1) in Four Different Precisions</b>	
single	$ulp(1) = 2^{-23} \sim 1.192093e-07$
double	$ulp(1) = 2^{-52} \sim 2.220446e-16$
double extended (x86)	$ulp(1) = 2^{-63} \sim 1.084202e-19$
quadruple (SPARC)	$ulp(1) = 2^{-112} \sim 1.925930e-34$

Recall that only a finite set of numbers can be exactly represented in any computer arithmetic. As the magnitudes of numbers get smaller and approach zero, the gap between neighboring representable numbers narrows. Conversely, as the magnitude of numbers gets larger, the gap between neighboring representable numbers widens.

For example, imagine you are using a binary arithmetic that has only 3 bits of precision. Then, between any two powers of 2, there are  $2^3 = 8$  representable numbers, as shown here:



The number line shows how the gap between numbers doubles from one exponent to the next.

In the IEEE single format, the difference in magnitude between the two smallest positive subnormal numbers is approximately  $10^{-45}$ , whereas the difference in magnitude between the two largest finite numbers is approximately  $10^{31}$ !

Below,  $\text{nextafter}(x, +\infty)$  denotes the next representable number after  $x$  as you move towards  $+\infty$ .

Gaps Between Representable Single-Format Floating-Point Numbers		
$x$	$\text{nextafter}(x, +\infty)$	Gap
0.0	1.4012985e-45	1.4012985e-45
1.1754944e-38	1.1754945e-38	1.4012985e-45
1.0	1.0000001	1.1920929e-07
2.0	2.0000002	2.3841858e-07
16.000000	16.000002	1.9073486e-06
128.00000	128.00002	1.5258789e-05
1.0000000e+20	1.0000001e+20	8.7960930e+12
9.9999997e+37	1.0000001e+38	1.0141205e+31

Any conventional set of representable floating-point numbers has the property that the worst effect of one inexact result is to introduce an error no worse than the distance to one of the representable neighbors of the computed result. When subnormal numbers are added to the representable set and gradual underflow is implemented, the worst effect of one inexact or *underflowed* result is to introduce an error no greater than the distance to one of the representable neighbors of the computed result.

In particular, in the region between zero and the smallest *normal* number, the distance between any two neighboring numbers equals the distance between zero and the smallest *subnormal* number. Subnormal numbers eliminate the possibility of introducing a roundoff error that is greater than the distance to the nearest representable number.

Because roundoff error is less than the distance to any of the representable neighbors of the true result, many important properties of a robust arithmetic environment hold, including these:

- $x \neq y \iff x - y \neq 0$
- $(x - y) + y \approx x$ , to within a rounding error in the larger of  $x$  and  $y$
- $1/(1/x) \approx x$ , when  $x$  is a normalized number, implying  $1/x \neq 0$

An old-fashioned underflow scheme is **Store 0**, which flushes underflow results to zero. Store 0 violates the first and second properties whenever  $x - y$  underflows. Also, Store 0 violates the third property whenever  $1/x$  underflows.

Let  $\lambda$  represent the smallest positive normalized number, which is also known as the underflow threshold. Then the error properties of gradual underflow and Store 0 can be compared in terms of  $\lambda$ .

gradual underflow:  $|\text{error}| < \frac{1}{2} \text{ulp}$  in  $\lambda$

Store 0:  $|\text{error}| \approx \lambda$

Even in single precision, the round-off error is **millions of times worse with Store 0** than gradual underflow.

## Two Examples of Gradual Underflow Versus Store 0

The following are two well-known mathematical examples. The first example computes an inner product.

```
sum = 0;
for (i = 0; i < n; i++)
{
    sum = sum + a[i] * y[i];
}
```

With gradual underflow, the result is as accurate as roundoff allows. In Store 0, a small but nonzero sum could be delivered that looks plausible but is wrong in nearly every digit. To avoid these sorts of problems, clever programmers must scale their calculations, which is only possible *if they can anticipate where minuteness might degrade accuracy*.

The second example, deriving a complex quotient, is not amenable to scaling:

$$a + ib = \frac{p + iq}{r + is}, \quad \text{assuming } |r/s| \leq 1, \quad = \frac{(p(r/s) + q) + i(q(r/s) - p)}{s + r(r/s)}$$

It can be shown that, despite roundoff, (1) the computed complex result differs from the exact result by no more than what would have been the exact result if  $p + iq$  and  $r + is$  each had been perturbed by no more than a few *ulps*, and (2) this error analysis holds in the face of underflows, except that when both  $a$  and  $b$  underflow, the error is bounded by a few *ulps* of  $|a + ib|$ . *Neither conclusion is true when underflows are flushed to zero.*

This algorithm for computing a complex quotient is robust, and amenable to error analysis, in the presence of gradual underflow. A similarly robust, easily analyzed, and efficient algorithm for computing the complex quotient in the face of Store 0 *does not exist*. In Store 0, the burden of worrying about low-level, complicated details shifts from the implementer of the floating-point environment to its users.

The class of problems that succeed in the presence of gradual underflow, but fail with Store 0, is larger than the fans of Store 0 may realize. Many frequently used numerical techniques fall in this class:

- Linear equation solving
- Polynomial equation solving
- Numerical integration
- Convergence acceleration
- Complex division

## Does Underflow Matter?

In the absence of gradual underflow, user programs need to be sensitive to the implicit inaccuracy threshold. For example, in single precision, if underflow occurs in some parts of a calculation, and Store 0 is used to replace underflowed results with 0, then accuracy can be guaranteed only to around  $10^{-31}$ , not  $10^{-38}$ , the usual lower range for single-precision exponents. This means that programmers need to

implement their own method of detecting when they are approaching this inaccuracy threshold, or else abandon the quest for a robust, stable implementation of their algorithm.

Some algorithms can be scaled so that computations don't take place in the constricted area near zero. However, scaling the algorithm and detecting the inaccuracy threshold can be difficult and time-consuming for each numerical program.

## 9 Fourier Transforms and Digital Signal Processing

Signals, noise, and Fourier Transforms are an essential part of much data analysis. It is a deep and broad subject, in which we can here establish only some foundational principles. The subject is, however, rife with misunderstandings and folklore. Therefore, we here also dispel some myths. For more specialized information, one must consult more specialized texts.

This section assumes you are familiar with complex arithmetic and exponentials, and with basic sampling and Fourier Transform principles. In particular, you must be familiar with decomposing a function into an orthonormal basis of functions. Understanding that a Fourier Transform is a phasor-valued function of frequency is very helpful, but not essential (see *Funky Electromagnetic Concepts* for a discussion of phasors).

We start with the most general (and simplest) case, then proceed through more specialized cases. We include some important (often overlooked) properties of Discrete Fourier Transforms. Topics:

- Complex sequences, and complex Fourier Transform (it's actually easier to start with the complex case, and specialize to real numbers later)
- Sampling and the Model of Digitization
- Even number of points vs. odd number of points
- Basis Functions and Orthogonality
- Real sequences: even and odd # points
- Normalization and Parseval's Theorem
- Continuous vs. discrete time and frequency; finite vs. infinite time and frequency
- Non-uniformly spaced samples

### Brief Definitions

**Fourier Series** represents a *periodic continuous* function as an infinite sum of sinusoids at discrete frequencies:

$$s(t) = \sum_{k=0}^{\infty} S_k e^{i2\pi k f_1 t}, \quad \text{where } S_k \text{ are complex (phasors), } t \equiv \text{time}$$

$$f_1 = 1/\text{period (in cycle/s or Hz)}, \quad \omega_1 \equiv 2\pi f_1 \text{ (in rad/s)}$$

$f_1 = 1/\text{period}$ , the lowest nonzero frequency, is called the **fundamental frequency**.  $f_0 = 0$ , always.

### Fourier Transform (FT)

represents a continuous function as an integral of sinusoids over continuous frequencies:

$$s(t) = \int_{-\infty}^{\infty} S(2\pi f) e^{i2\pi f t} df = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) e^{i\omega t} d\omega, \quad \text{where } S(\cdot) \text{ is complex}$$

We do not discuss this here. The function  $s(t)$  is not periodic, so there is no fundamental frequency.  $S(\omega)$  is a phasor-valued function of angular frequency.

### Discrete Fourier Transform (DFT)

represents a finite sequence of numbers as a finite sum of sinusoids:

$$s_j = \sum_{k=0}^{n-1} S_k e^{i2\pi(k/n)j}, \quad \text{where } S_k \text{ are complex (phasors), } j = 0, \dots, n-1 \equiv \text{the sample index,}$$

$$f_1 = 1/\text{period (in cycle/s)}, \quad \omega_1 = 2\pi f_1 \text{ (in rad/s)}$$

The sequence  $s_j$  may be thought of as either periodic, or undefined outside the sampling interval. As in the Fourier Series, the fundamental frequency is  $1/\text{period}$ , or equivalently  $1/(\text{sampled interval})$ , and  $f_0 = 0$ , always.



[Since a DFT essentially treats the input as periodic, it might be better called a Discrete Fourier Series (rather than Transform), but Discrete Fourier Transform is completely standard.]

**Fast Fourier Transform (FFT)**

an algorithm for implementing special cases of DFT.

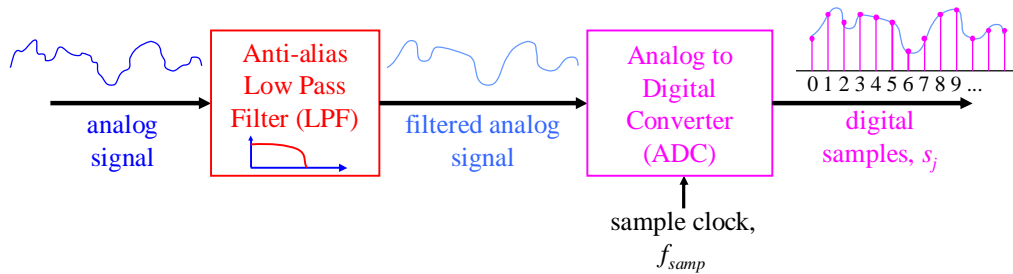
**Inverse Discrete Fourier Transform (IDFT)**

gives the sequence of numbers  $s_j$  from the DFT components.

The general digital Fourier Transform is a Discrete Fourier Transform (DFT).  
An FFT is an algorithm for special cases of DFT.

**Model of Digitization and Sampling**

All realistic systems which digitize analog signals must comprise at least the components in Figure 9.1.



**Figure 9.1** Minimum components of a Digital Signal Processing system, with uniformly spaced samples.

In this example, the output of the digitizer is a sequence of real numbers,  $s_j$ . Other systems (such as coherent quadrature downconverters) produce complex numbers.

**Sampling Does Not Produce Impulses**

It is often said that sampling a signal is like setting it to zero everywhere except at the sample times, or like creating a series of impulses. It is not.

These notions are not true, and can be misleading [O&S p8b]. Note that a single impulse (in time) has infinite power. Therefore, a sum (sequence) of such impulses also has infinite power. In contrast, the original signal, and the sequence of samples, has finite power. This suggests immediately that samples are *not* equivalent to a series of impulses.

Nonetheless, there *is* an identity that involves impulse functions, which we discuss after introducing the DFT.

**Complex Sequences and Complex Fourier Transform**

It's actually easier to start with the complex case, and specialize to real numbers later. Given a sequence of  $n$  complex numbers  $s_j$ , we can write the sequence as a sum of sinusoids, i.e. complex exponentials:

**Inverse Discrete Fourier Transform:**

$$s_j = \sum_{k=0}^{n-1} S_k e^{i2\pi(k/n)j}, \quad \text{where } j = 0, \dots, n-1 \text{ is the sample index}$$

$\frac{k}{n} \equiv$  the frequency of the  $k^{\text{th}}$  component, in cycle/sample

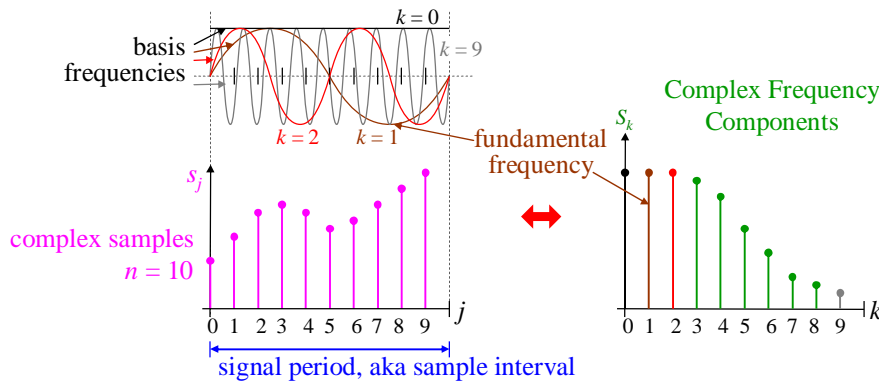
$S_k \equiv$  the complex frequency component (phasor)

Note that there are  $n$  original complex numbers, and  $n$  complex frequency components, so no information is lost. The transform is exact, unique, and reversible. (In other words, this is *not* a "fit.")

The above equation forces all normalization conventions. We use the simple scheme wherein a function equals the sum of its components (with no factors of  $2\pi$  or anything else).

Often, the index  $j$  is a measure of time or distance, and the sequence comprises samples of a signal taken at equal intervals. Without loss of generality, we will refer to  $j$  as a measure of “time,” but it could be anything. Note that the equation above actually defines the **Inverse Discrete Fourier Transform (IDFT)**, because it gives the original sequence from the Fourier components. [Mathematicians often reverse the definitions of DFT and IDFT, by putting a minus sign in the exponent of the IDFT equation above. Engineers and physicists usually use the convention given here.]

Each number in the sequence is called a **sample**, because such sequences are often generated by sampling a continuous signal  $s(t)$ . For  $n$  samples, there are  $n$  frequency components,  $S_k$ , each at normalized frequency  $k/n$  (defined soon); see Figure 9.2.



**Figure 9.2** Samples in time, and their frequencies. For simplicity, the samples, sinusoids, and component amplitudes are shown as real, but in general, they are all complex valued.

Note that there are a full  $n$  sample times in the sample interval (aka *signal period*), not  $(n - 1)$ .

The above representation is used by many DFT functions in computer libraries.

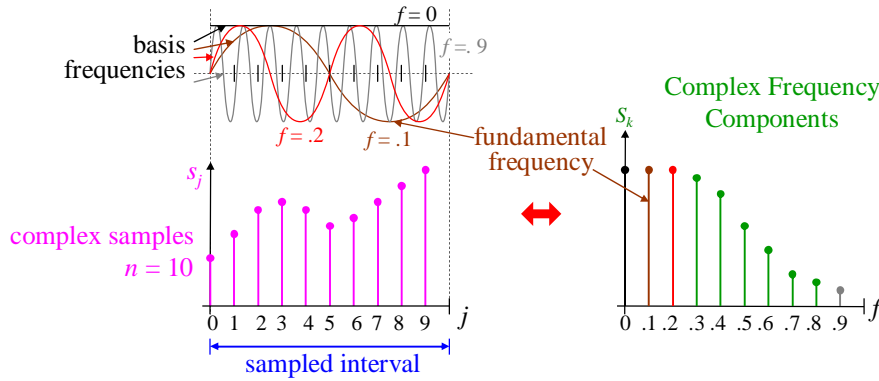
Also, there is no need for any other frequencies, because  $k = 10$  has exactly the same values at all the sample points as  $k = 0$ . If the samples are from a continuous signal that had a frequency component at  $k = 10$ , then that component will be **aliased** down to  $k = 0$ , and added to the actual  $k = 0$  component. It is forever lost, and cannot be recovered from the samples, nor distinguished from the  $k = 0$  (DC) component. The same aliasing occurs for any two frequencies  $k$  and  $k + n$ .

The above definition is the *only* correct meaning for “aliasing.” Many (most?) people misuse this word to mean other things (e.g., “harmonics” or “sidebands”).

To avoid a dependence on  $n$ , we usually label the frequencies as fractions. For  $n$  samples, there are  $n$  frequencies, measured in units of cycles/sample, and running from  $f = 0$  to  $f = (1 - 1/n)$  cycles/sample. The  $n$  **normalized frequencies** are

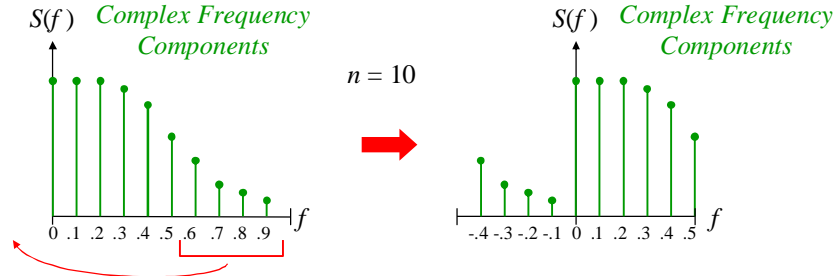
$$f_k = \frac{k}{n}, \quad k = 0, 1, \dots, n-1, \quad \text{that is} \quad \{f_k\} = \left\{0, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \dots, \frac{n-1}{n}\right\}.$$

There is no  $f = 1$ , just as there is no  $k = n$ , because  $f = 1$  is an alias of  $f = 0$ . Continuous Fourier components are written as  $S(f)$ , a function of  $f$ , so we re-label the above diagram with normalized frequencies:

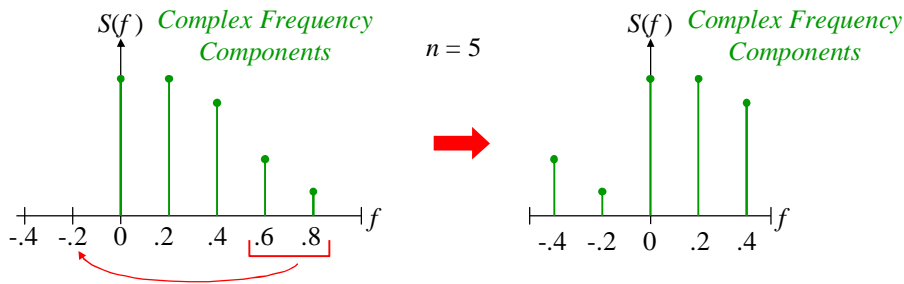


Normalized frequencies are equivalent to measuring time in units of the sample time, and frequencies in cycles/sample.

For theoretical analysis, it is often more convenient to have the frequency range be  $-0.5 < f \leq 0.5$ , instead of  $0 \leq f < 1$ . Since any frequency  $f$  is equivalent to (an alias of)  $f - 1$ , we can simply move the frequencies in the range  $0.5 < f < 1$  down to  $-0.5 < f < 0$ :



For an even number of samples (and frequencies, diagram above), the resulting frequency set is necessarily asymmetric, because there is no  $f = -0.5$ , but there is an  $f = +0.5$ . For an odd number of points (below), the frequency set is symmetric, and there is neither  $f = -0.5$  nor  $f = +0.5$ :



**Non-Equivalence of DFT and FT of Series of Time-Domain Impulses (Again)**

As noted earlier, it is often said that sampling is like setting the function to zero between samples, or creating a series of impulse functions. This is a common misconception. It is well refuted by [Openheim and Schaffer p??, and dozens of other signal processing experts]. It is easy to show that that claim is not true, in several ways. One simple way is this: For a band-limited signal, I can reconstruct the signal between the sample times from just the samples alone. That makes no sense if sampling amounted to zeroing the signal between samples, because that would be a new function, which would destroy information about the original. There would then be no way to recreate the original function from its DFT.

Furthermore, it is often said that the FT of a series of time-domain impulses is identical to the DFT of samples at those times. From the previous paragraph, this cannot be true, either. However, the following is true only at the (integer)  $k$  defined DFT frequencies of  $k\omega_1$ :

$$S(\omega_k) = S(k\omega_1) = \frac{1}{n} \sum_{j=0}^{n-1} s_j e^{-ik\omega_1 t_j} = \frac{1}{n} \int_{-\infty}^{\infty} \left[ \sum_{j=0}^{n-1} s_j \delta(t - t_j) \right] e^{-ik\omega_1 t} dt$$

$\omega_1 \equiv$  fundamental frequency.

For frequencies in between the  $k\omega_1$ , the DFT is formally undefined, but can be taken as zero for the purpose of reconstructing the original samples. However, at those in-between frequencies the FT has some non-zero values which are usually of little interest. So we see that the FT of a series of weighted impulses (representing the sample values), *evaluated at the DFT frequencies*  $k\omega_1$ , is proportional to the DFT, but the full-spectrum of the FT is *different* from the spectrum of the DFT. Hence, the two transformations are *not* equivalent.

### Basis Functions and Orthogonality

The basis functions of the DFT are the discrete-time exponentials, which are equivalent to sines and cosines:

$$b_k(j) = e^{i(2\pi k/n)j}$$

where  $j \equiv$  sample index = 0,1, ... n-1,

$$k \equiv \text{frequency index} = 0,1, \dots n-1 \quad \text{or} \quad \begin{cases} n \text{ even: } -n/2+1, \dots -1, 0, 1 \dots n/2 \\ n \text{ odd: } -\text{int}(n/2), \dots -1, 0, 1, \dots \text{int}(n/2) \end{cases}$$

Note that:

The DFT and FT are simply decompositions of functions into basis functions, just like in ordinary quantum mechanics. The transform equations are just the inner products of the given functions with the basis functions.

The basis functions are orthogonal, normalized (in our convention) such that  $\langle b_k | b_m \rangle = n \delta_{km}$ . Proof:

$$\langle b_k | b_m \rangle \equiv \sum_{j=0}^{n-1} b_k^*(j) b_m(j) = \sum_{j=0}^{n-1} e^{-i(2\pi k/n)j} e^{i(2\pi m/n)j} = \sum_{j=0}^{n-1} e^{i(2\pi/n)(m-k)j} = \sum_{j=0}^{n-1} \left[ e^{i(2\pi/n)(m-k)} \right]^j$$

For  $k = m$ , we have  $\langle b_k | b_m \rangle = \sum_{j=0}^{n-1} \left[ e^0 \right]^j = n$ .

For  $k \neq m$ , use  $\sum_{j=0}^{n-1} r^j = \frac{1-r^n}{1-r}$  where  $r = \left[ e^{i(2\pi/n)(m-k)} \right]$ . Then:

$$\langle b_k | b_m \rangle = \frac{1 - \left[ e^{i(2\pi/n)(m-k)} \right]^n}{1 - \left[ e^{i(2\pi/n)(m-k)} \right]} = \frac{1 - e^{i(2\pi)(m-k)}}{1 - \left[ e^{i(2\pi/n)(m-k)} \right]} = \frac{1-1}{1 - \left[ e^{i(2\pi/n)(m-k)} \right]} = 0 \Rightarrow \langle b_k | b_m \rangle = n \delta_{km}$$

The orthogonality condition allows us to immediately write the DFT from the definition of the IDFT above:

**Discrete Fourier Transform:**

$$S_k = \frac{1}{n} \sum_{j=0}^{n-1} s_j e^{-i2\pi(k/n)j}, \quad \text{where } S_k \equiv \text{the } k^{\text{th}} \text{ complex frequency component}$$

$$\frac{k}{n} = \text{the normalized frequency of the } k^{\text{th}} \text{ component.}$$

Note that there are  $2n$  independent real numbers in the complex sequence  $s_j$ , and there are also  $2n$  independent real numbers in the complex spectrum  $S_k$ , as there must be (same number of degrees of freedom).

**Real Sequences**

An important case of sequence  $s_j$  is a real-valued sequence, which is a special case of a complex-valued sequence. In this section, we use the positive and negative frequency DFT form, where  $k$  takes on both negative and positive integer values. Then for  $s_j = \sum_{k \approx -n/2}^{\approx n/2} S_k e^{i2\pi(k/n)j}$  to be real, the  $S_k$  must occur in complex conjugate pairs, i.e., the spectrum  $S_k$  must be **conjugate symmetric**:

$$S_k = S_{-k}^* \quad \text{for } s_j \text{ real, and } k \leq \text{int}(n/2) + 1.$$

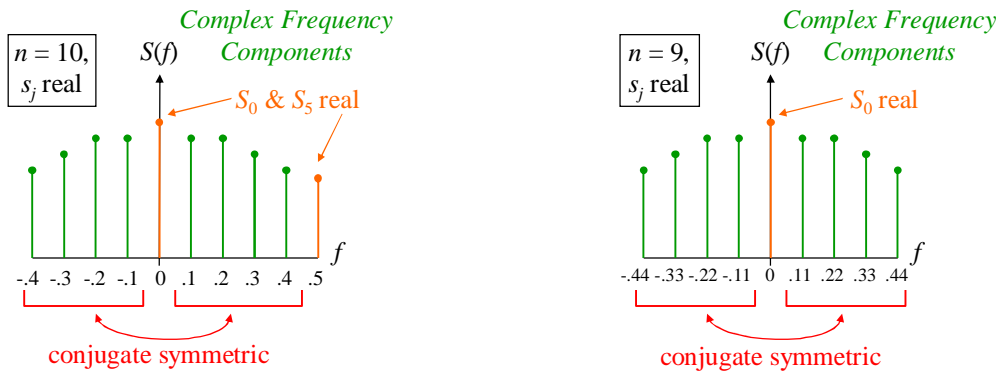
This implies that  $S_0$  is always real, which is also clear since  $S_0$  is just the average of the real sequence.

We now discuss the lower limit for  $k$ . (As discussed earlier, there is no  $k = -n/2$ ). There are  $n$  independent real numbers in the real sequence  $s_j$ . We now consider two subcases:  $n$  is even, and  $n$  is odd.

For  $n$  even,

$$s_j = \sum_{k=-n/2+1}^{n/2} S_k e^{i2\pi(k/n)j} \quad n \text{ even, } s_j \text{ real,}$$

and we use the asymmetric frequency range  $-0.5 < f \leq 0.5$ , which corresponds to  $-n/2 + 1 \leq k \leq n/2$  (Figure 9.3, left). For an even number of sample points, since there are no conjugates to  $k = 0$  or  $k = n/2$ , we must have that  $S_0$  and  $S_{n/2}$  are real (actually,  $S_0$  being real still satisfies conjugate symmetry). All other  $S_k$  may be complex, and are conjugate symmetric:  $S_{-k} = S_k^*$ .



**Figure 9.3** (Left) Sequence with even number of samples,  $n = 10$ . (Right) Sequence with odd number,  $n = 9$ .

Therefore, in the spectrum, there are  $(n/2 - 1)$  independent complex frequency components, plus two real components, totaling  $n$  independent real numbers in the spectrum, matching the  $n$  independent real numbers in the sequence  $s_j$ .

In terms of sine and cosine components (rather than the complex components), there are  $(n/2 + 1)$  independent cosine components, and  $(n/2 - 1)$  independent sine components. All frequencies are nonnegative:

$$s_j = A_0 + \sum_{k=0}^{n/2-1} \{A_k \cos[2\pi(k/n)j] + B_k \sin[2\pi(k/n)j]\} + A_{n/2} \cos \pi j \tag{9.1}$$

$n$  even,  $s_j$  real.

Note that in the last term,  $\cos \pi j$  is just an alternating sequence of  $+1, -1, +1, \dots$ .

For an odd number of points (Figure 9.3, right),

$$s_j = \sum_{k=-(n-1)/2}^{(n-1)/2} S_k e^{i2\pi(k/n)j}, \quad n \text{ odd},$$

there is no  $k = n/2$  component, and again there is no conjugate to  $k = 0$ . Therefore, we must have that  $S_0$  is real. All other  $S_k$  are conjugate symmetric. Therefore, in the spectrum, there are  $(n - 1)/2$  independent complex frequency components, plus one real component ( $S_0$ ), totaling  $n$  independent real numbers in the spectrum, matching the  $n$  independent real numbers in the sequence  $s_j$ .

In terms of sine and cosine components (rather than the complex components), there are  $(n + 1)/2$  independent cosine components, and  $(n - 1)/2$  independent sine components. All frequencies are nonnegative:

$$s_j = A_0 + \sum_{k=0}^{(n-1)/2} \{A_k \cos[2\pi(k/n)j] + B_k \sin[2\pi(k/n)j]\}, \quad n \text{ odd}, s_j \text{ real} . \tag{9.2}$$

Note that there is no final lone-cosine term, and no alternating sequence.

These examples illustrate how the notation is slightly more involved for the cosine/sine form than for the complex exponential form.

### Normalization and Parseval's Theorem

When the original sequence represents something akin to samples of voltage over time, we sometimes speak of "energy" in the signal. The energy of the signal is the sum of the energies of each sample:

$$E_j = G s_j^2 = s_j^2, \quad \text{where } G = \text{"conductance"}, \text{ chosen to be } 1.$$

$$E = \sum_{j=0}^{n-1} E_j = \sum_{j=0}^{n-1} s_j^2.$$

When the "conductance" is chosen to be 1, or some other reference value, the "energy" in the signal does *not* correspond to any physical energy (say, in joules).

The energies of the sinusoidal components in the DFT add as well, because the sinusoids are orthogonal (show why??):

$$E \propto \sum_{k=0}^{n-1} |S_k|^2.$$

Parseval's Theorem equates the energy of the original sequence to the energy of the sinusoidal components, by providing the constant of proportionality. First, we evaluate the energy of a single sinusoid:

$$E_k = |S_k|^2 \sum_{j=0}^{n-1} |e^{i(2\pi k/n)j}|^2 = n |S_k|^2 \quad \text{where } k \equiv \text{frequency index} = 0, 1, \dots, n-1.$$

Then summing over all frequencies yields:

$$E = \sum_{k=0}^{n-1} E_k = n \sum_{k=0}^{n-1} |S_k|^2 \quad \Rightarrow \quad \boxed{E = n \sum_{k=0}^{n-1} |S_k|^2 = \sum_{j=0}^{n-1} s_j^2} \quad (9.3)$$

**Energy For Real Sequences:** We derive Parseval’s Theorem for real sequences in two ways. Since the  $s_j$  are real, the interesting form is the cosine/sine form of DFT, (9.1) and (9.2). We again consider separately the cases of  $n$  even, and  $n$  odd.

First, for  $n$  even,  $k$  runs from 0 to  $(n/2)$ . We can deduce the equation for Parseval’s Theorem by exploiting the conjugate symmetry of the  $S_k$ . Recall that  $S_0$  has no conjugate term, nor does  $S_{n/2+1}$ . Therefore:

$$E_0 = nA_0^2, \quad E_{n/2+1} = n(A_{n/2})^2.$$

For  $k = 1, \dots, n/2$ , we have:

$$A_k = 2 \operatorname{Re}\{S_k\}, \quad B_k = 2 \operatorname{Im}\{S_k\}, \quad |S_k|^2 + |S_{-k}|^2 = 2 \operatorname{Re}\{S_k\}^2 + 2 \operatorname{Im}\{S_k\}^2 \quad \Rightarrow$$

$$E_k = \frac{n}{2} (A_k^2 + B_k^2), \quad k = 1, \dots, n/2.$$

We can derive this another way directly from (9.1). Since  $A_0$  is a constant added to each  $s_j$ , the energy contributed from this term is  $E_0 = nA_0$ . Since  $\cos \pi j$  is just alternating  $+1, -1, \dots$ , it’s energy at each sample is 1, and  $E_{n/2+1} = n(A_{n/2})^2$ . Finally, the average value of  $\cos^2$  over a full period is  $1/2$ , as is the average of  $\sin^2$ . Therefore, for  $k = 1, \dots, n/2$ ,  $E_k = (n/2)(A_k^2 + B_k^2)$ .

Second, for  $n$  odd,  $k$  runs from 0 to  $(n - 1)/2$ . The above arguments still apply, but there is no lone-cosine term at the end. Therefore the result is the same, without the lone-cosine term.

Summarizing:

$$n \text{ even: } E = nA_0^2 + n(A_{n/2})^2 + \frac{n}{2} \sum_{k=1}^{n/2-1} (A_k^2 + B_k^2)$$

$$n \text{ odd: } E = nA_0^2 + \frac{n}{2} \sum_{k=1}^{(n-1)/2} (A_k^2 + B_k^2)$$

**Other normalizations:** Besides our normalization choice above, there are several other choices in common use. In general, between the DFT, IDFT, and Parseval’s Theorem, you can choose a normalization for one, which then fixes the normalization for the other two. For example, some people choose to make the DFT and IDFT more symmetric by defining:

$$\left. \begin{array}{l} \text{IDFT: } s_j = \frac{1}{\sqrt{n}} \sum_{k=0}^{n-1} S_k e^{(i2\pi k/n)j} \\ \text{DFT: } S_k = \frac{1}{\sqrt{n}} \sum_{j=0}^{n-1} s_j e^{-i(2\pi k/n)j} \end{array} \right\} \Rightarrow \sum_{k=0}^{n-1} |S_k|^2 = \sum_{j=0}^{n-1} s_j^2 \quad (\text{alternate normalizations}).$$

**Continuous and Discrete, Finite and Infinite**

TBS: Finite length implies discrete frequencies; infinite length implies continuous frequencies. Discrete time implies finite frequencies; continuous time implies infinite frequencies. Finite length is equivalent to periodic.

## White Noise and Correlation

**White noise** has, on average, all frequency components equal (named by incorrect analogy with white light); samples of white noise are uncorrelated. Non-white noise has unequal frequency components (on average); samples of non-white noise are necessarily correlated. (Show this??).

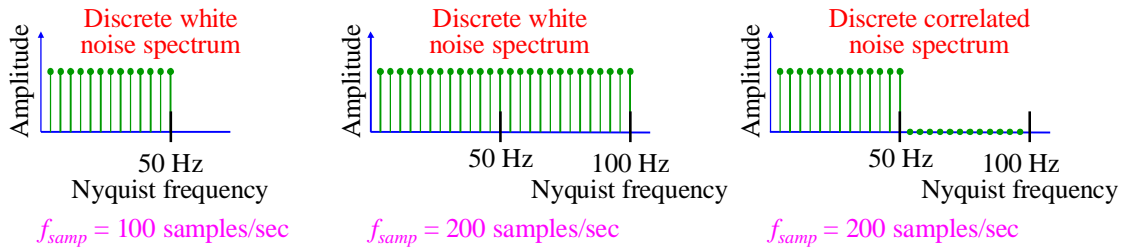
### Why Oversampling Does Not Improve Signal-to-Noise Ratio

Sometimes it might seem that if I oversample a signal (i.e., sample above the Nyquist rate), the noise power stays constant (= noise variance is constant), but I have more samples of the signal which I can average. Therefore, by oversampling, I should be able to improve my SNR by averaging out more noise, but keeping all the signal.

This reasoning is wrong, of course, because it implies that by sampling arbitrarily fast, I can filter out arbitrarily large amounts of noise, and ultimately recover anything from almost nothing. So what's wrong with this reasoning? Let's take an example.

Suppose I sample a signal at 100 samples/sec, with white noise. Then my Nyquist frequency is 50 Hz, and I must use a 50 Hz Low Pass Filter (LPF) for anti-aliasing before sampling. This LPF leaves me with 50 Hz worth of noise power (= variance).

Now suppose I double the sampling frequency to 200 samples/sec. To maintain white noise, I must open my anti-alias filter cutoff to the new Nyquist frequency, 100 Hz. This doubles my noise power. Now I have twice as many samples of the signal, with twice as much noise power. I can run a LPF to reduce the noise (say, averaging adjacent samples). At best, I cut the noise by half, reducing it back to its 100 sample/sec value, and reducing my sample rate by 2. Hence, I'm right back where I was when I just sampled at 100 samples/sec in the first place.



But wait! Why open my anti-alias LPF? Let's try keeping the LPF at 50 Hz, and sampling at 200 samples/sec. But then, my noise occupies only 1/2 of the sampling bandwidth: it occupies only 50 Hz of the 100 Hz Nyquist band. Hence, the noise is *not* white, which means adjacent noise samples are *correlated*! Hence, when I average adjacent samples, the noise variance does *not* decrease by a factor of 2. The factor of 2 *gain only* occurs with uncorrelated noise. In the end, oversampling buys me nothing.

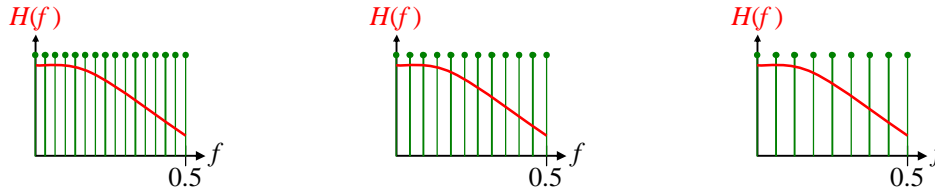
### Filters TBS??

FIR vs. IIR. Because the data set can be any size, and arbitrarily large:

The transfer function of an FIR or IIR is continuous.

Consider some filter. We must carefully distinguish between the filter in general, which can be applied to *any* data set (with any  $n$ ), and the filter as applied to one particular data set. Any given data set has only discrete frequencies; if we apply the filter to the data set, the data set's frequencies will be multiplied by the filter's transfer function at those frequencies. But we can apply *any* size data set to the filter, with frequency components,  $f = k/n$ , anywhere in the Nyquist interval. For every data set, the filter has a transfer function at all its frequencies. Therefore, the filter in general has a *continuous* transfer function.





Data sets with different  $n$  sample the transfer function  $H(f)$  at different points.  $H(f)$ , in general, is a continuous curve, defined at *all* points in the Nyquist interval  $f \in [0, 1)$  or  $(-0.5, +0.5]$ .

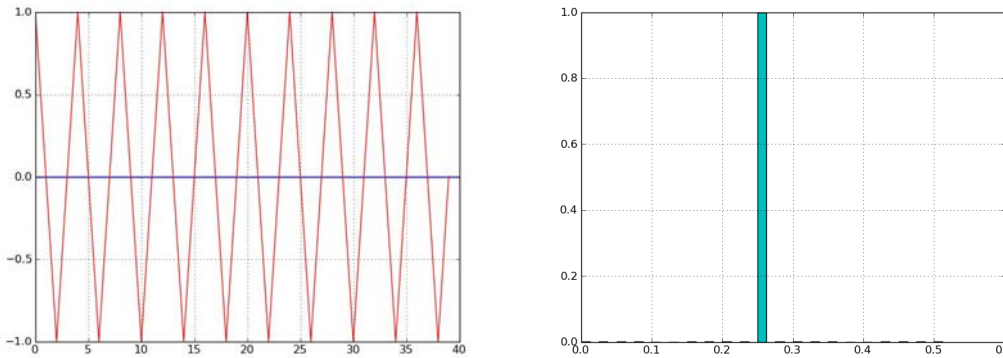
### What Happens to a Sine Wave Deferred?

“... Maybe it just sags, like a heavy load. Or does it explode?” [Sincere apologies to Langston Hughes.] You may ask, “The DFT has only a finite set of basis frequencies. Can I use a DFT to estimate the frequency of an unknown signal? What happens if I sample a sinusoid of a frequency *in between* the chosen DFT basis frequencies? What is its spectrum?” Good questions. We now demonstrate. The results here are important for generalizing the DFT, and spectral analysis in general, to non-uniformly sampled signals.

We choose  $n = 40$  samples, which means the basis frequencies are  $k(1/n)$ ,  $k = -19, \dots, 0, \dots, 20$ , measured in cycles per sample (or equivalently, in units of the sampling rate,  $f_{\text{samp}}$ ). The frequency spacing is  $1/n = 0.025$  cycle/sample. No other frequencies exist in the DFT.

First, we show the result of sampling an *existing*-frequency sinusoid of  $f = 10/n = 0.25$  cycle/sample ( $k = 10$ ). Since the signal is real, the spectrum is conjugate symmetric ( $S_{-k} = S_k^*$ ); therefore, I show only the positive frequencies, and double their magnitudes:

$$s_j = \cos\left(\frac{2\pi k}{n} j\right), \quad f = \frac{k}{n} \text{ cycle/sample .}$$

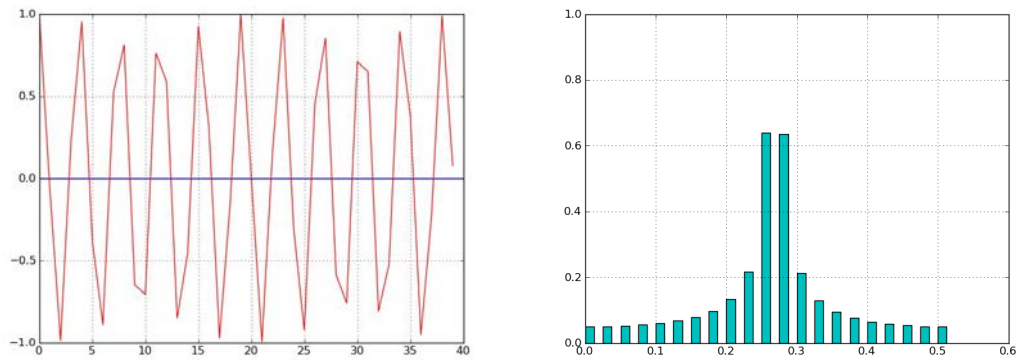


(Left) A sampled sinusoid of  $f = 0.25$ ,  $n = 40$ . (Right) As expected, its magnitude spectrum (DFT) has exactly one component at  $f = 0.25$ , with magnitude 1.0.

[Notice that when the sample points are connected by straight segments, the sinusoid doesn’t “look” sinusoidal, but recall that connecting with straight segments is *not* the proper way to interpolate between samples.]

The “energy” of the sample set is exactly  $(1/2)40 = 20$ , because there is an integral number of cycles in the sample set, and the average energy of a sinusoid is  $1/2$ .

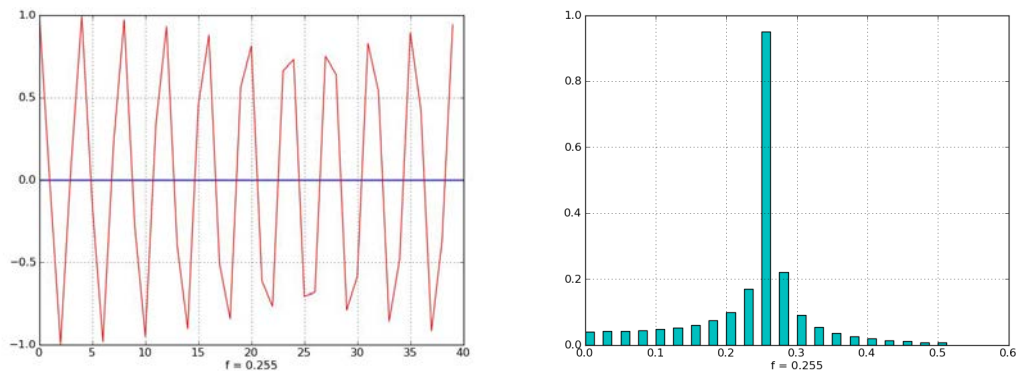
Now we take our signal off-frequency by half the frequency spacing:  $f = 10.5/n = 0.2625$  cycle/sample, halfway between two chosen basis frequencies:



(Left) A sampled sinusoid of  $f = 0.2625$ ,  $n = 40$ . (Right) Its magnitude spectrum (DFT) has components everywhere, but is peaked around  $f = 0.2625$ .

Not too surprisingly, the components are peaked at the two basis frequencies closest to the sinusoid frequency, but there are also components at *all other* frequencies. This is an artifact of sampling a pure sinusoid of a non-basis frequency for a finite time. Note also that the total energy in the sampled signal is slightly *larger* than that of the  $f = 0.25$  signal, even though they are both the *same* amplitude. This is due to a few more of the samples being near a peak of the signal. This shift in total energy is another artifact of sampling a non-basis frequency. For other signal frequencies, or other time shifts, the energy could just as well be *lower* in the sampled signal. This energy shift also explains why the two largest components of the spectrum are not exactly equal, even though they are equally distant from the true signal frequency of  $f = 0.2625$ .

Finally, instead of being half-way between allowed frequencies, suppose we're only 0.2 of the way,  $f = 10.2/n = 0.255$  cycle/sample:



(Left) A sampled sinusoid of  $f = 0.255$ ,  $n = 40$ . (Right) Its magnitude spectrum (DFT) has components everywhere, is asymmetric, and peaked at  $f = 0.25$ .

The two largest components are still those surrounding the signal frequency, with the larger of the two being the one closer to the signal frequency.

These examples show that a DFT, with its fixed basis frequencies, can give only a rough estimate of an unknown sinusoid's frequency. The estimate gets worse if the unknown signal is not exactly a sinusoid, because that means it has an even smaller spectral peak, with more components spread around the spectrum.

Other methods exist for estimating the frequency of an unknown signal, even one that is non-uniformly sampled in time. If the signal is fairly sinusoidal, one can correlate with a sinusoidal basis frequency, and numerically search for the frequency with maximum correlation. This avoids the discrete-frequency

limitation of a DFT. Other methods usually require many periods of data, e.g. epoch folding [Leahy, Ap J, 1983, vol 266, p160??].

## Nonuniform Sampling and Arbitrary Basis Functions

So far, we have used a signal sampled uniformly in time. We now show that one can find a Fourier transform of a signal with *any* set of  $n$  samples, uniform or not. This has many applications: some experiments (such as lunar laser ranging) cannot sample the signal uniformly for practical, economic, or political reasons. Magnetic Resonance Imaging (MRI) often uses non-uniform sampling to reduce imaging time, which can be an hour or more for a patient.

We write the required transform as a set of simultaneous equations, with  $t_j$  as the arbitrary sample times, and keeping (for now) the uniformly spaced frequencies:

$$\begin{aligned}
 s(t_0) &= \sum_{k=0}^{n-1} S_k \exp(i(2\pi k/n)t_0) \\
 s(t_1) &= \sum_{k=0}^{n-1} S_k \exp(i(2\pi k/n)t_1) && \text{OR} \\
 &\dots \\
 s(t_{n-1}) &= \sum_{k=0}^{n-1} S_k \exp(i(2\pi k/n)t_{n-1})
 \end{aligned}$$

$$\begin{bmatrix} s(t_0) \\ s(t_1) \\ \vdots \\ s(t_{n-1}) \end{bmatrix} = \begin{bmatrix} \exp(2\pi f_0 t_0) & \exp(2\pi f_1 t_0) & \dots & \exp(2\pi f_{n-1} t_0) \\ \exp(2\pi f_0 t_1) & \exp(2\pi f_1 t_1) & \dots & \exp(2\pi f_{n-1} t_1) \\ \vdots & \vdots & \ddots & \vdots \\ \exp(2\pi f_0 t_{n-1}) & \exp(2\pi f_1 t_{n-1}) & \dots & \exp(2\pi f_{n-1} t_{n-1}) \end{bmatrix} \begin{bmatrix} S_0 \\ S_1 \\ \vdots \\ S_{n-1} \end{bmatrix}$$

How can we find the required coefficients,  $S_k$ ?

The exponential functions are no longer orthogonal over the sample times; they are only orthogonal over uniformly spaced samples.

Nonetheless, we have  $n$  unknowns ( $S_0, \dots, S_{n-1}$ ), and  $n$  equations. So long as the basis functions are *linearly independent* over the sample times, we can (in principle) solve for the needed coefficients,  $S_k$ . We have now greatly expanded our ability to decompose arbitrary samples into basis functions:

We can decompose a signal over any set of sample times into any set of linearly independent (not necessarily orthogonal) basis functions.

Note that Parseval's theorem does *not* apply to the coefficients, since the basis functions (evaluated at the non-uniform sample points) are no longer orthogonal. Also,  $S_0$  is no longer the average of the signal values, since the sinusoids may have nonzero average over the sample points.

There is one more subtlety: what is the fundamental frequency  $f_0$ ? Equivalently, what is the signal period? The two are related, because  $f_0 = 1/\text{period}$ . There is no unique answer to this. However, since a finite signal transforms as if it is periodic, the period cannot be  $(t_{n-1} - t_0)$ , since the first and last samples would then have to be identical. The period must be longer than that. A convenient choice is to simply mimic what happens when the samples *are* uniform. In that case,

$$\text{period} = (t_{n-1} - t_0) \frac{n}{n-1}, \quad f_0 = 1/\text{period}$$

This choice for period reproduces the traditional DFT when the samples are uniform, and is usually adequate for non-uniform samples, as well.

**Example: DFT of a real, non-uniformly sampled sequence:** We can set up the matrix equation to be solved by recalling the frequency layout for even and odd  $n$ , and applying the above. We set  $t_0 = 0$ , and define  $n/2$  as floor( $n/2$ ). For illustration of the last two columns, we take  $n$  odd:

$$\begin{aligned}
 s(t_0) &= \sum_{k=0}^{n/2} S_k [\cos(k\omega t_0) + \sin(k\omega t_0)] && \text{where } \omega \equiv 2\pi/n \\
 s(t_1) &= \sum_{k=0}^{n/2} S_k [\cos(k\omega t_1) + \sin(k\omega t_1)] && \text{OR} \\
 &\vdots \\
 s(t_{n-1}) &= \sum_{k=0}^{n/2} S_k [\cos(kt_{n-1}) + \sin(k\omega t_{n-1})]
 \end{aligned}$$

$$\begin{bmatrix} s(t_0) \\ s(t_1) \\ \vdots \\ s(t_{n-1}) \end{bmatrix} = \begin{bmatrix} 1.0 & 1.0 & 0.0 & \dots & 1.0 & 0.0 \\ 1.0 & \cos(\omega t_1) & \sin(\omega t_1) & \dots & \cos((n/2)\omega t_1) & \sin((n/2)\omega t_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1.0 & \cos(\omega t_{n-1}) & \sin(\omega t_{n-1}) & \dots & \cos((n/2)\omega t_{n-1}) & \sin((n/2)\omega t_{n-1}) \end{bmatrix} \begin{bmatrix} S_0 \\ S_1 \\ \vdots \\ S_{n-1} \end{bmatrix}$$

This gives us the sine and cosine components separately. For  $n$  even, the highest frequency component is  $k = n/2$ , or  $\omega = 2\pi k/n = 2\pi(1/2) = \pi$  rad/sample, and the final column of  $\sin(\cdot)$  is not present.

Note that this is not a fit; it is an exact, reversible transformation. The matrix is the set of all the basis functions (across each row), evaluated at the sample points (down each column). The matrix has no summations in it, and depends on the sample points, but not on the sample values.

**Example: basis functions as powers of  $x$ :** In the continuous world, a Taylor series is a decomposition of a function into powers of  $(x - a)$ , which are a set of linearly independent (but not orthogonal) basis functions. Despite this lack of orthogonality, Taylor devised a clever way to evaluate the basis-function coefficients without solving simultaneous equations.

**Example: sampled standard basis functions:** We could choose a standard (continuous) mathematical basis set, such as Bessel functions,  $J_n(t)$ . For  $n$  sample points,  $t_1, \dots, t_n$ , the Bessel functions are linearly independent, and we can solve for the coefficients,  $A_k$ . We need a scale factor  $\alpha$  for the time (equivalent to  $2\pi k/n$  in the Fourier transform). For example, we might use  $\alpha =$  the  $(n - 1)^{th}$  zero of  $J_{n-1}(t)$ . Then:

$$\begin{aligned}
 s(t_0) &= \sum_{k=0}^{n-1} A_k J_k \left( t_0 \frac{\alpha}{t_{n-1}} \right) \\
 s(t_1) &= \sum_{k=0}^{n-1} A_k J_k \left( t_1 \frac{\alpha}{t_{n-1}} \right) \\
 &\dots \\
 s(t_{n-1}) &= \sum_{k=0}^{n-1} A_k J_k \left( t_{n-1} \frac{\alpha}{t_{n-1}} \right)
 \end{aligned}$$

We have  $n$  equations and  $n$  unknowns,  $A_0, \dots, A_{n-1}$ , so we can solve for the  $A_k$ .

---

### Don't Pad Your Data, Even for FFTs

Old fashioned FFT implementations required you to have  $N =$  a power of 2 number of samples (64, 1024, etc.). Modern FFT implementations are general to any number of samples, and use the prime decomposition of  $N$  to provide the fastest and most accurate DFT known. The worst case is when  $N$  is

prime, and no FFT optimization is possible: the DFT is evaluated directly from the defining summations. But with modern computers, this is usually so fast that we don't care.

In the old days, if people had a non-power-of-2 number of data points, they used to "pad" their data, typically (and horribly) by just adding zeros to the end until they reached a power of 2. This introduced artifacts into the spectrum, which often obscured or destroyed the very information they sought [Ham p??].

Don't pad your data. It screws up the spectrum.  
With a modern FFT implementation, there is no need for it, anyway.

If for some reason, you absolutely must constrain  $N$  to some preferred values, it is much better to throw away some data points than to add fake ones.

## Two Dimensional Fourier Transforms

One dimensional Fourier transforms often have time or space as the independent variable. Two dimensional transforms almost always have space, say  $(x, y)$ , as the independent variables. The most common 2D transform is of pictures.

In the continuous world of light, lenses can physically project a Fourier transform of an image based on optics, with no computations. This allows for filtering the image with opaque masks, and re-transforming back to the original-but-filtered image, all at the speed of light with no computer. But digitized images store the image as pixels, each with some light intensity. These are computationally processed by computer.

### Basis Functions

TBS. Not sines and cosines, or products of sines and cosines. Products of complex exponentials. Wave fronts at various angles, discrete  $k_x$  and  $k_y$ .

## Note on Continuous Fourier Series and Uniform Convergence

The continuous Fourier Series is defined for a periodic signal  $s(t)$  over a continuous range of times,  $t \in [0, T)$ :

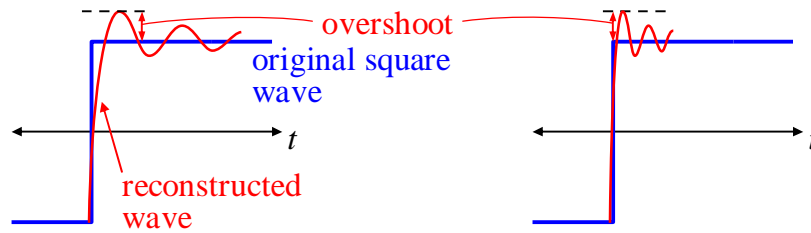
$$s(t) = \sum_{k=0}^{\infty} S_k e^{i2\pi k \omega_0 t}, \quad \text{where } k \omega_0 \text{ is the frequency of the } k^{\text{th}} \text{ component}$$

$S_k$  is the complex frequency component

Note that the time interval is continuous, but the frequency components are discrete. In general, periodic signals lead to discrete frequency components.

The continuous Fourier Series is *not* always uniformly convergent.  
Therefore, the order of integrations and summations cannot always be interchanged.

Non-uniform convergence is illustrated by the famous **Gibbs phenomenon**: when we transform a square wave to the frequency domain (aka Fourier space), retain only a finite number of frequency components, and then transform back to the time domain, the square wave comes back with **overshoot**: wiggles that are large near the discontinuities:



Gibbs phenomenon: (Left) After losing high frequencies, the reconstructed square wave has overshoot and wiggles. (Right) Retaining more frequencies reduces wiggle time, but not amplitude.

As we include more and more frequency components, the wiggles get narrower (damp faster), but do not get lower in amplitude. This means that there are always some time points for which the inverse transform does not converge to the original square wave. Such wiggles are commonly observed in many electronic systems, which must necessarily drop high frequency components above some cut-off frequency.

However:

Continuous signals have Fourier Series that converge uniformly. This applies to most physical phenomena, so interchanging integration and summation is valid [F&W p217+].

This is true even if the derivative of the signal is discontinuous.

## Fourier Transforms, Periodograms, and Lomb-Scargle

In some circles, one hears the terms “Fourier Transform,” “periodogram,” and “Lomb-Scargle” a lot. Each of these is distinct, but they are related. Understanding the differences can help you analyze your data. We provide here an overview of some common signal processing algorithms. Be warned:

Because spectral analysis can be tricky, its practice is rife with misunderstanding and mythology.

Throughout the text, I will occasionally note common misunderstandings, but there are too many for me to correct them all. We address the Lomb-Scargle algorithm in particular, since it is widely misunderstood.

Correspondingly, the terminology is also highly confused and abused. We define here some common terms in ways that are consistent with the majority of our (limited) experience in the literature. However, there appears to be little universal agreement on precise definitions, especially across different disciplines. (Words are the tools of communication; it is impossible to make fine points with dull tools.) In this work, we adhere to the following definitions:

- **Spectral analysis** is the examination of periodic components of a data sequence.
- The **energy** of a single data point is its squared magnitude, and is always  $\geq 0$  (the term “energy” derives from early applications where the squared magnitude was proportional to physical energy). The “energy” of a sample-set is the sum of the squares of the samples. The “energy” of a frequency component is the sum of the “energies” of that frequency taken over the sample times.
- The **power** in a frequency component is its squared magnitude, and is often normalized in some specified way. In some references, the term “energy” is used for “power.” (As with “energy,” the “power” in a component might be unrelated to physical power.) In this work, we occasionally write “energy” and “power” in quotes, as a reminder that they are not physical energy and power. For non-uniform sampling, the “energy” of a frequency component is *not* proportional to its “power.”
- The **statistical significance** is the false alarm probability, often called alpha  $\alpha$ . Experimenters usually choose  $\alpha$  before analyzing the data. It is the probability that a pure noise signal will, by chance, suggest the presence of a signal. It is essentially the same as the **p-value**. Note that a *lower* significance means a result is *more* significant, i.e. more likely real, and less likely random. Nonetheless, authors often speak loosely of “higher” significance meaning “more significant” or a

lower significance value. It is more clear to say “more significant” instead of “higher significance.”

- A **detection parameter** is a statistic calculated for a trial signal that (roughly) tells how likely the trial is to be a real phenomenon, rather than a result of random chance. A higher significance means a result is less likely to be random. In Lomb-Scargle, the significance of a frequency tells how likely that frequency is to be a significant component of the signal. Note that “significance” is different than “power.”
- A **DFT** (Discrete Fourier Transform) is a precisely defined decomposition of a sequence (uniformly spaced or not) into a set of sinusoidal components. (An “FFT” is just an efficient way to perform a DFT in some limited cases. We have limited use for “FFT” here.) DFTs use uniformly spaced frequencies, but are easily extended to non-uniformly spaced frequencies.
- A **periodogram** is some kind of graph of periodic components of a sample set. There are many methods for producing a periodogram, which produce different results [2]. Therefore, a “periodogram” is less well-defined than a DFT. Usually, the frequencies in a periodogram are uniformly spaced, but the periodogram frequency spacing may be tighter than the DFT.
- **Lomb-Scargle** is a formula for finding the significance of a *given* sinusoidal frequency in data.
- A **Lomb-Scargle periodogram** is a graph of detection parameter vs. frequency, where each parameter is computed with the Lomb-Scargle algorithm. The “LS” in periodogram can stand for either “Lomb-Scargle” or “Least Squares”, since the Lomb-Scargle algorithm produces the detection parameter for a least squares sinusoidal fit. Note that the LS algorithm produces a *detection parameter*, not power, despite common belief to the contrary.

Be careful to distinguish between uniformly spaced *samples* of data,  
and uniformly spaced *frequencies* in the periodogram.

**Caution:** For orthogonal basis functions (as in a uniformly sampled DFT), the energy and power of every frequency are proportional, and therefore the terms are often interchanged. However, for non-uniform sampling, the “energy” of a frequency component is *not* proportional to its “power.” This is the crux of the confusion about the LS “periodogram.” The LS result is essentially the “energy” of a given sinusoidal frequency in the data, used to help find significant sinusoids in the data.

## The Discrete Fourier Transform vs. the Periodogram

The single biggest distinction between a DFT and a Lomb-Scargle periodogram is that the DFT *simultaneously* optimizes *all* the components to form an *exact* transformation. A Lomb-Scargle periodogram examines each frequency *by itself*, regardless of other frequencies.

The DFT is exact, and invertible, with no loss of information. At times, this can be a plus, but in many cases, this “exactness” results in anomalies. In particular, any set of physical measurements is only a subset of the *exact* representation of the physical phenomenon. In other words, a *sample set* is incomplete, and so the information contained in it is limited. In addition, all measurements contain some noise. If we put such a sample set into a DFT, it gives us frequency components which *exactly* match the given incomplete samples, noise and all. To achieve this exactness, the DFT must sometimes contort the spectrum in unphysical ways. In particular, highly non-uniformly sampled signals often result in large DFT artifacts. By definition, a DFT produces a spectrum of precisely defined, uniformly spaced frequencies. [However, one can easily compute an exact decomposition onto an arbitrary set of frequencies, and furthermore, onto an arbitrary set of basis functions that need not be sinusoidal.]

As scientists, we often would rather see something less mathematically exact, and more physically meaningful. We combine all our knowledge of the system (and science in general) with our limited, noisy data, to reach new conclusions. A periodogram provides a way to look at frequency content of a signal, without some of the unphysical anomalies of an exact DFT. Also, a periodogram can plot results at an arbitrary set of frequencies, not just those defined in a DFT. In fact, periodograms usually choose a larger, and more densely packed, set of frequencies than a DFT produces. However, periodograms suffer from anomalies, as well.

In a DFT, the frequency components don't "overlap," in the sense that none of the "information" of one component appears in any other component. This is true even though the basis sinusoids are not orthogonal over the given sample times. There is no "extra" information in the DFT: e.g., a sample set of  $n = 40$  points transforms into exactly 21 cosines and 19 sines, having exactly the same 40 degrees of freedom as the original data set.

In contrast, in a periodogram, the component powers are themselves correlated, and the information from one component also shows up in some of the other components (especially adjacent components). Furthermore, especially in small data sets [ref??], the non-orthogonality of the periodogram's sinusoids may cause a single component of the data to produce spikes at multiple widely-spaced frequencies in the periodogram. This may mislead the user into believing there are multiple causes, one for each peak. Finally, for any sample size  $n$ , we can make a periodogram with any number of frequencies, even far more than  $n$ . This again shows that the periodogram contains redundant information.

Despite common belief, a Lomb-Scargle periodogram is not a periodogram of sinusoidal "power." It is a graph of *detection parameter* vs. frequency, where each parameter is computed by a minimum least-squares residual fit of a single sinusoid at that frequency [3]. For large data sets, or well-randomized sample times, the parameter value approaches the power, so people often "get away with" confusing the two. However, for small data sets, or those where the sample times are clustered around a periodic event (say, nighttime), the significance of a frequency can be very different than its "power" estimate. Note that when the sample times are clustered around a frequency, say 1 cpd (cycle per day), it can affect many frequencies in the sample, especially near harmonics or sub-harmonics (e.g., 2 cpd, 3 cpd, 0.5 cpd, etc.).

When fitting a sinusoid of *given* frequency to data, there are two fit parameters. These may be taken as cosine and sine amplitudes, or equivalently as magnitude and phase of a single sinusoid. The true "power" at that frequency (considered by itself) is the sum of the squares of the cosine and sine amplitudes, or equivalently, the square of the magnitude.

## Practical Considerations

Here are a few possible issues with spectral analysis. Again, it is a highly involved topic, and these issues are only a tiny introduction to it.

**Removing trends:** Before using spectral analysis, it is common to remove simple trends in the data, such as a constant offset, or straight line trends [ref??]. A straight-line introduces a complicated spectral structure which often obscures the phenomena of interest. Thus, removing it before spectral analysis usually helps. A constant offset introduces spurious frequency detections, especially for bunched samples, as are typical astronomical data. Also, constant offsets may lead to worse round-off error. Furthermore, even though you should never pad your data (see below), padding with zeros when your data has a non-zero average only compounds your error.

**Stepwise regression:** Sometimes we have in our data a frequency component which is obscuring the phenomenon of interest. We may model (fit) that frequency, and subtract it from the data, in hopes of revealing the interesting data. Note that finding frequencies in our data, and subtracting them, one at a time, is simply the standard statistical method of *stepwise* multiple regression (not simultaneous multiple regression). We are "regressing" one frequency component at a time. Therefore, stepwise frequency subtraction has all the usual pitfalls of stepwise regression. In particular, the *single* biggest component may be completely subsumed by two (or more) smaller components. Therefore, when performing such stepwise frequency modeling, it may help to use the standard method of backward elimination to delete from the model any previously found component that is no longer useful in the presence of newer components.

**Computational burden:** Many decomposition algorithms rely on some form of orthogonality, e.g., this is the basis (wink) of Discrete Fourier Transforms. Orthogonality allows a basis decomposition to be done by correlation (aka using inner-products). Recall that such a correlation decomposition, including Lomb-Scargle periodograms, requires  $O(n^2)$  operations. In contrast, a non-orthogonal decomposition, such as a DFT over non-uniform sample times, solves simultaneous equations requiring  $O(n^3)$  operations, so can be *much* slower. For  $n = 1,000$  samples, the non-orthogonal decomposition is about 1,000 times slower, and requires billions of operations. This may be a noticeable burden, even on modern computers (perhaps requiring many minutes).



**Smoothed DFT:** One surprisingly common approach to making a periodogram (not Lomb-Scargle) is to make a DFT, with its possible anomalies, and then try to disperse those anomalies by “smoothing” the resulting graph of power vs. frequency. I believe smoothing a DFT is like trying to invest wisely after you’ve lost all your money gambling. It’s too late: you can’t get back what’s already lost. Likely much better is to make some other kind of periodogram in the first place, and don’t use a DFT, or use it only as guidance for more appropriate analyses. In particular, with highly non-uniformly spaced samples, the DFT anomalies include large (but unphysical) amplitudes, which are not removed by smoothing. Furthermore, smoothing a DFT of nonuniformly spaced samples requires  $O(n^3)$  operations, so it not only likely produces poor results, it does so slowly.

One possible advantage of the “smoothed DFT” approach is that for very large data sets ( $n > \sim 10,000$ ), if  $n$  is amenable to a Fast Fourier Transform *and* your samples are uniformly spaced, then the DFT can be done in  $O(n \log n)$  operations. A typical Lomb-Scargle periodogram requires  $O(n^2)$  operations. However, Press and Rybicki [1] provide a way to use FFT-like methods to create a Lomb-Scargle periodogram, thus using  $O(n \log n)$  operations. While still slower than a true FFT, this makes Lomb-Scargle periodograms of millions of data points feasible.

**Bad information:** As mentioned earlier, many references (seemingly most references, especially on the web), are wrong in important (but sometimes subtle) ways. E.g., some references actually *recommend* padding your data (almost always a terrible idea, discussed elsewhere in *Funky Mathematical Physics Concepts*). Many references incorrectly describe the differences between uniform and non-uniform sampling, the meaning of FFT, aliasing, and countless other concepts. In particular,

Some references say that sampling a signal is like setting it to zero everywhere except the sample times. It is not.

This is a common misconception, which is discussed earlier in the section “Sampling.”

**The Lomb-Scargle Algorithm**

We here describe the Lomb-Scargle (L-S) algorithm; the next section explains how it works. We start with  $n$  discrete measurements (samples),  $s_j$ , taken at times  $t_j, j = 0, \dots, n-1$ . The algorithm first finds the time offset that makes the cosine and sine orthogonal *over the given sample times*:

$$\exists \tau \text{ such that } \sum_{j=0}^{n-1} \cos \omega t_j \sin \omega t_j = 0. \quad \tau \text{ satisfies } \tan(2\omega\tau) = \frac{\sum_{j=0}^{n-1} \sin 2\omega t_j}{\sum_{j=0}^{n-1} \cos 2\omega t_j}$$

Note that  $\tau$  depends on  $\omega$ ; so each  $\omega$  has its own  $\tau$ . Also,  $\tau$  depends on the sample times, but not on the measurements,  $s_j$ .

Next, L-S subtracts out the average signal, giving samples

$$h_j \equiv s_j - \langle s_j \rangle \quad \text{where } \langle s_j \rangle \equiv \text{average of } s_j.$$

Then the Lomb-Scargle *normalized periodogram* is, in inner product notation:

$$D(\omega) = \frac{1}{2s^2} \left[ \frac{\langle \cos | h \rangle^2}{\langle \cos | \cos \rangle} + \frac{\langle \sin | h \rangle^2}{\langle \sin | \sin \rangle} \right] \quad [\text{from [1] eq. 3 p277}]$$

where  $s^2 \equiv$  unbiased weighted sample variance.

We deliberately use the non-standard notation  $D(\omega)$ , rather than  $P(\omega)$ , to emphasize that the L-S parameter is a detection statistic, *not* a power (despite widespread belief). Expanded in more conventional notation, the L-S normalized periodogram is [1]:

$$D(\omega) = \frac{1}{2s^2} \left[ \frac{\left( \sum_{j=0}^{n-1} h_j \cos \omega(t_j - \tau) \right)^2}{\sum_{j=0}^{n-1} \cos^2 \omega(t_j - \tau)} + \frac{\left( \sum_{j=0}^{n-1} h_j \sin \omega(t_j - \tau) \right)^2}{\sum_{j=0}^{n-1} \sin^2 \omega(t_j - \tau)} \right] \quad \text{[[1] eq. 3 p277].}$$

NB: This assumes *equal uncertainties* on the data. This is exactly the equation for energy one gets from a standard statistical fit which minimizes the residual signal in a least-squares sense (i.e., minimum residual energy) [4]. Such a fit is a simultaneous 2-parameter linear fit (for  $A$  and  $B$ ) to the model:

$$S_{fit}(t) = A \cos(\omega(t - \tau)) + B \sin(\omega(t - \tau)), \quad P_{true}(\omega) = A^2 + B^2.$$

$P_{true}(\omega)$  is the true estimate of the “power” at  $\omega$ , because it is proportional to the *squared* amplitude of the fitted sinusoid at frequency  $\omega$ . For large data sets, or well-randomized sample times,  $D(\omega)$  approaches being proportional to  $P_{true}(\omega)$  at all frequencies. Therefore, the parameter  $D(\omega)$  is often used as a substitute for the spectral power estimate,  $P_{true}(\omega)$ . As with most hypothesis testing, the presence of a spectral line (frequency) is deemed likely if the line’s parameter is substantially less likely than that expected from pure noise. Since both terms in the L-S formula are gaussian random variables (RVs), the Lomb-Scargle expression in brackets for pure gaussian noise is proportional to a  $\chi^2_{\nu=2}$  distribution. The factor of  $1/2$  makes the probability distribution of  $D(\omega)$  approach a unit-mean exponential [3], rather than a  $\chi^2_{\nu=2}$ . However, the normalization by  $s^2$  means that  $D(\omega)$  is exactly beta distributed (*not* F distributed, as thought for decades) [A. Schwarzenberg-Czerny, 1997].

Note that  $s^2$  is (close to) the average “energy” (squared value) of the samples (remember that the average value of all the samples has been subtracted off, so the  $h_j$  have zero average). The  $1/s^2$  in this equation makes the result independent of the signal amplitude, i.e. multiplying all the data by a constant has no effect on the periodogram. Also, for pure noise,  $D(\omega)$  is roughly independent of the number of samples,  $n$ , since  $s^2$  is independent of  $n$ , and the numerators and denominators both scale roughly like  $n$ . The numerator summations scale like  $\sqrt{n}$ , because they are sums of random variables (noise).

In contrast, if a signal is present at frequency  $\omega$ ,  $D(\omega)$  grows like  $n$ , because then the numerator summation grows like  $n$ . Thus, if a signal is present, it becomes easier to detect with a larger sample set, consistent with our intuition.

### The Meaning Behind the Math

Understanding exactly what Lomb-Scargle does, and how it works, puts you in a powerful position to know when to use it, and its limitations. Also, if you ever want to develop a novel algorithm, or have ever wondered how others develop them, Lomb-Scargle provides an interesting and informative example of the process. (However, our derivation here is very different from Lomb’s original [Lom].) The Lomb-Scargle formula may look daunting, but we can understand and derive it in just a few high-level steps:

1. Given our basis of cosine and sine, find a way to make them orthogonal.
2. Use standard orthogonal decomposition of our data into our two basis functions.
3. Normalize our coefficients, being careful to distinguish power-estimate from detection parameter.
4. Prove that the correlation amplitude of the previous steps is equivalent to the least-squares fit.

We complete these steps below, in full detail.

#### 1. Make Cosine and Sine Orthogonal

When making a LS periodogram, we are *not* performing a basis decomposition. We are separately finding correlations with each periodogram frequency, without regard to any other frequencies. For real-valued data (i.e., not complex), there are two basis functions at any frequency: cosine and sine. We need both to find the detection level (and also the “power”) at that frequency.

At any given frequency,  $\omega$ , we have two basis functions,  $\cos(\omega t)$  and  $\sin(\omega t)$ , which we write as a sum:  $A\cos(\omega t) + B\sin(\omega t)$ . Recall how a uniformly sampled DFT works:  $\omega$  is a multiple of the fundamental frequency, and our sample times are *uniformly* spaced. Then cosine and sine are naturally orthogonal:

$$\sum_{j=0}^{n-1} \cos(\omega j \Delta t) \sin(\omega j \Delta t) = 0, \quad \text{where} \quad \begin{cases} \omega \equiv \text{a multiple of fundamental frequency} \\ j \equiv \text{sample number} \\ \Delta t \equiv \text{sampling period} = 1/f_{\text{samp}} \end{cases}$$

Using this orthogonality, we find our coefficients  $A$  and  $B$  separately, using inner products:

$$A = \langle s | \cos \rangle = \frac{2}{n} \sum_{j=0}^{n-1} s_j \cos(\omega j \Delta t), \quad B = \langle s | \sin \rangle = \frac{2}{n} \sum_{j=0}^{n-1} s_j \sin(\omega j \Delta t).$$

The power at frequency  $\omega$  is then  $A^2 + B^2$ .

In contrast, for arbitrary sample times  $t_j$  (as in much observational data), or for arbitrary  $\omega$ ,  $\cos(\cdot)$  and  $\sin(\cdot)$  are not orthogonal (i.e., they are “correlated”):

$$C \equiv \sum_{j=0}^{n-1} \cos(\omega t_j) \sin(\omega t_j) \neq 0, \quad \text{where} \quad \begin{cases} \omega \equiv \text{arbitrary frequency} \\ j \equiv \text{sample number} \\ t_j \equiv \text{arbitrary sample times} \end{cases}$$

Being correlated, we cannot use simple inner-products to find  $A$  and  $B$  separately. Furthermore, the presences of other components prevents us from simply simultaneously solving for the amplitudes  $A$  and  $B$ .

Despite being correlated, cosines and sines are usually still a convenient basis, because they are the eigenfunctions of linear, time-invariant systems, and appear frequently in physical systems. So we ask: Is there a way to “orthogonalize” the cosines and sines over the *given* set of arbitrary sample times? Yes, there is, as we now show.

Consider the basis-function parameters we have to play with: amplitude, frequency, and phase. We are given the frequency, and are seeking the amplitudes. The only parameter left to adjust is phase (or equivalently, a shift in time). So we could write the correlation amplitude  $C$  above as a function of some phase shift  $\phi$ :

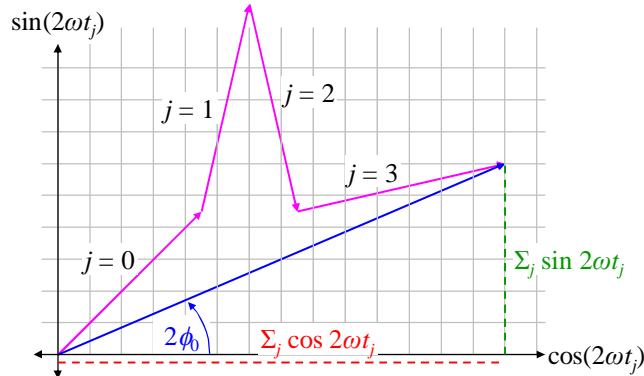
$$C(\phi) \equiv \sum_{j=0}^{n-1} \cos(\omega t_j - \phi) \sin(\omega t_j - \phi).$$

Can we find a phase shift  $\phi_0$  such that  $C(\phi_0) = 0$ , thus constructing a pair of orthogonal cosine and sine? The simplest shift I can think of is  $\pi$ :  $\cos(\omega t_j + \pi) = -\cos(\omega t_j)$ , and similarly for  $\sin(\cdot)$ . Thus a phase shift of  $\pi$  negates both cosine and sine, and the correlation is not affected:  $C(\pi) = C(0)$ . The next simplest shift is  $\pi/2$ . This converts  $\cos(\cdot) \rightarrow \sin(\cdot)$ , and  $\sin(\cdot) \rightarrow -\cos(\cdot)$ , so  $C(\pi/2) = -C(0)$ . This is great:  $C(\phi)$  is a continuous function of  $\phi$ , and it changes sign between 0 and  $\pi/2$ . This means that somewhere between 0 and  $\pi/2$ ,  $C(\phi) = 0$ , i.e. the cosine and sine are orthogonal.

The existence of a phase-shift  $\phi_0$  which makes cosine and sine orthogonal is important, because we can always find the required  $\phi_0$  numerically. Even better, it turns out that we can find a closed-form expression for  $\phi_0$ . We notice that the correlation  $C(\phi)$  can be rewritten, using a simple identity:

$$\sin 2\theta = 2 \cos \theta \sin \theta \quad \Rightarrow \quad C(\phi_0) = 0 = \sum_{j=0}^{n-1} 2 \sin(2\omega t_j - 2\phi_0), \quad \text{or} \quad \sum_{j=0}^{n-1} \sin(2\omega t_j - 2\phi_0) = 0.$$

Given the sample times  $t_j$ , how do we find  $\phi_0$ ? We can use geometry: let's set  $\phi = 0$  for now, and plot the sum of the vectors corresponding to  $(x = \cos(2\omega t_j), y = \sin(2\omega t_j))$ , for some hypothetical sample times,  $t_j$ . Each vector is unit length (see Figure 9.4).



**Figure 9.4** Sum (blue) of  $n = 4$  vectors corresponding to  $(x = \cos(2\omega t_j), y = \sin(2\omega t_j))$

[This is equivalent to plotting the complex numbers  $\exp(i2\omega t_j)$  in the complex plane.] In the example shown above, if we rotate all the vectors clockwise by  $2\phi_0$ , then the sum of the sine components will be zero. The components of the vector sum are the sums of the components, so:

$$\tan(2\phi_0) = \frac{\sum_{j=0}^{n-1} \sin 2\omega t_j}{\sum_{j=0}^{n-1} \cos 2\omega t_j} \Rightarrow C(\phi_0) \equiv \sum_{j=0}^{n-1} \cos(\omega t_j - \phi_0) \sin(\omega t_j - \phi_0) = 0.$$

In other words, we rotate each component (in the  $2\omega$  set) by  $-2\phi_0$ , which corresponds to rotating each component of our original ( $1\omega$ ) set by  $-\phi_0$ . This gives the condition we need for orthogonality.

Any phase shift, at a given frequency, can be written as a time shift. By convention, Lomb-Scargle uses a subtracted time shift, so:

$$2\omega\tau = 2\phi_0 \Rightarrow C(\phi_0) = \sum_{j=0}^{n-1} \cos(\omega(t_j - \tau)) \sin(\omega(t_j - \tau)) = 0.$$

With this time shift, as with the  $\phi_0$  phase shift, the cosines and sines are orthogonal over the given sample times. Be careful to distinguish  $\phi_0$ , the orthogonalizing phase shift, from the fitted-sinusoid phase, usually called  $\phi$ .

**2. Use orthogonal decomposition of our data into our basis functions**

Now that we have orthogonal basis functions (though not yet normalized), we can find our cosine and sine coefficients with simple correlations (aka inner-products):

$$A' = \langle h | \cos \rangle = \sum_{j=0}^{n-1} h_j \cos \omega(t_j - \tau), \quad B' = \langle h | \sin \rangle = \sum_{j=0}^{n-1} h_j \sin \omega(t_j - \tau) \quad (\text{unnormalized}),$$

where the primes indicate unnormalized coefficients. Note that, because the offset cosine and sine functions are orthogonal,  $A'$  and  $B'$  fit for both components *simultaneously*. That is, orthogonality implies that the *individual* best fits for cosine and sine are also the *simultaneous* best fit.

**3. Normalize Our Coefficients**

With all orthonormal basis decompositions, we require *normalized* basis functions to get properly scaled components. We normalize our coefficients by dividing them by the squared-norm of the (unnormalized) basis functions:

$$A = \frac{A'}{\langle \cos | \cos \rangle} = \frac{\sum_{j=0}^{n-1} h_j \cos \omega(t_j - \tau)}{\sum_{j=0}^{n-1} \cos^2 \omega(t_j - \tau)}, \quad B = \frac{B'}{\langle \sin | \sin \rangle} = \frac{\sum_{j=0}^{n-1} h_j \sin \omega(t_j - \tau)}{\sum_{j=0}^{n-1} \sin^2 \omega(t_j - \tau)} \quad (\text{normalized}).$$

These formulas are similar to the two terms in the Lomb-Scargle formula. The normalized coefficients,  $A$  and  $B$ , yield a true power estimate for a best-fit sinusoid:

$$P_{true}(\omega) = A^2 + B^2 \quad \text{from} \quad S_{fit}(\omega) = A \cos \omega(t_j - \tau) + B \sin \omega(t_j - \tau).$$

To arrive at the Lomb-Scargle detection parameter, we must consider not the true *power* estimate,  $P_{true}(\omega)$ , but the *contribution* to the total sample set “energy” (sum of squares) from our fitted sinusoid,  $S_{fit}(\omega)$ , at the given sample times. For example, for a frequency component with a given true power, if the sinusoid happens to be small at the sample times, then that component contributes a small amount to the sample-set “energy.” On the other hand, if the sinusoid happens to be large at the sample times, then that component contributes a large amount to the sample-set “energy.”

The *significance* of a frequency component at  $\omega$  is a function of the ratio of the component’s energy to that expected from pure noise. Given a component with cosine and sine amplitudes  $A$  and  $B$ , its energy in the sample set is the sums of the squares of its samples, at the given sample times:

$$\begin{aligned} E(\omega) &= \sum_{j=0}^{n-1} [A \cos \omega(t_j - \tau)]^2 + \sum_{j=0}^{n-1} [B \sin \omega(t_j - \tau)]^2 \\ &= A^2 \sum_{j=0}^{n-1} \cos^2 \omega(t_j - \tau) + B^2 \sum_{j=0}^{n-1} \sin^2 \omega(t_j - \tau) \\ &= \frac{\left( \sum_{j=0}^{n-1} h_j \cos \omega(t_j - \tau) \right)^2}{\sum_{j=0}^{n-1} \cos^2 \omega(t_j - \tau)} + \frac{\left( \sum_{j=0}^{n-1} h_j \sin \omega(t_j - \tau) \right)^2}{\sum_{j=0}^{n-1} \sin^2 \omega(t_j - \tau)}. \end{aligned}$$

This is precisely the Lomb-Scargle detection parameter.

For wideband noise, with no signal, the samples  $h_j$  are independent identically distributed (IID), with variance equal to the noise power,  $\sigma^2$ . Across many sets of noise, then, the numerators above have variance:

$$\sigma_A^2 = \sigma^2 \sum_{j=0}^{n-1} \cos^2 \omega(t_j - \tau), \quad \sigma_B^2 = \sigma^2 \sum_{j=0}^{n-1} \sin^2 \omega(t_j - \tau)$$

This means each term in  $E(\omega)$  has variance  $= \sigma^2$ .

For gaussian noise,  $A$  and  $B$  are gaussian, and  $E(\omega)$  is the sum of their squares, scaled to the *estimated* variance  $= \sigma^2$ . Therefore,  $E(\omega)/\sigma^2$  (always  $\leq 1$ ) is distributed with CDF an incomplete beta function [A. Schwarzenberg-Czerny, 1997]. (For decades, it was thought that  $E(\omega)/\sigma^2$  was  $\chi^2_{\nu=2}$  distributed, but it is easy to show that it is not:  $\chi^2$  has no upper bound, but  $E(\omega)/\sigma^2 \leq 1$ .) Nonetheless, assuming the incorrect  $\chi^2_{\nu=2}$  distribution, Lomb-Scargle divides by 2 to get a more-convenient exponential distribution with  $\mu = 1$ :

$$D(\omega) = \frac{1}{2} \cdot \frac{E(\omega)}{\sigma^2} = \frac{1}{2\sigma^2} \left[ \frac{\left( \sum_{j=0}^{n-1} h_j \cos \omega(t_j - \tau) \right)^2}{\sum_{j=0}^{n-1} \cos^2 \omega(t_j - \tau)} + \frac{\left( \sum_{j=0}^{n-1} h_j \sin \omega(t_j - \tau) \right)^2}{\sum_{j=0}^{n-1} \sin^2 \omega(t_j - \tau)} \right].$$

This is the standard (though flawed) formula for LS detection. Note again that we don't know the true  $\sigma^2$ ; we must estimate it from the samples.

N. R. Lomb first derived this result with a completely different method, using the standard "normal equations" for least-squares fitting [5].

**4. Prove that the correlation amplitude of the previous steps is equivalent to the least-squares fit**

This is a general theorem: any correlation amplitude for a component of a sequence  $s_j$  is equivalent to a least-squares fit. We prove it by contradiction. Given any single basis function,  $b_k$ , we can construct a complete, orthonormal basis set which includes it. In that case, the component of  $b_k$  is found by correlation, as usual. Call it  $A_k$ .

The least-squares residue is simply the energy of the sequence after subtracting off the  $b_k$  component. Since the basis set is orthonormal, Parseval's theorem holds. Thus, the residual energy after subtracting the  $b_k$  component from  $s_j$  is the sum of the squares of all the *other* component amplitudes. If there existed some *other* value of  $A_k$  which had less residual energy, then that would imply a *different* decomposition into the *other* basis functions. But the decomposition into an orthogonal basis is unique. Therefore, no  $A_k$  other than the one given by correlation can have a smaller residual.

The basis coefficient given by correlation is a least-squares-residual fit.

This proof holds equally well for discrete sequences  $s_j$ , and for continuous functions  $s(t)$ .

**Bandwidth Correction (aka Bandwidth Penalty)**

Determining the significance of a signal detection requires some care, since most algorithms search for any *one* of many possible signals. For example, when searching for periodic signals in noisy data, one often searches many trial frequencies, and a "hit" on any frequency counts as a detection. How do we determine the significance ( $p$  value) of such a detection?  $p$  is also called the "false alarm" probability.

All the common periodic-signal detection algorithms require bandwidth correction, because if one makes enough attempts, even an unlikely outcome will eventually happen. If one tries many frequencies, the probability that *one* of them exceeds a threshold is much higher than the probability of a *single* given frequency exceeding that threshold. From elementary statistics, if the parameters for all frequencies are independent, the probability that they are *all* not false alarm (FA) is the product of the probabilities that each one is not false alarm. For  $M$  independent parameters at various frequencies, and a given  $p$ -value, then in our gaussian white noise case (i.e., the standard Null Hypothesis of no signal):

$$\Pr(\text{all not FA}) = \Pr(\text{one not FA})^M = (1 - p_{1f})^M$$

$$\text{confidence level} \equiv 1 - p = \Pr(\text{all not FA}) = (1 - p_{1f})^M.$$

Therefore, to achieve an overall  $p$ -value for all frequencies of  $p$ , we must choose the  $p$ -value for each trial frequency such that [Schwarzenberg-Czerny (1997)]:

$$1 - p = (1 - p_{1f})^M, \quad \Rightarrow \quad p_{1f} = 1 - (1 - p)^{1/M}.$$

Since  $p$  is usually small ( $< \sim .05$ ), we can often use the binomial theorem to approximate:

$$(1-p)^{1/M} \approx 1-p/M, \quad p_{1f} \approx p/M.$$

For simulations, we may want to estimate  $M$  from  $p$  and  $p_{1f}$ . For example, we choose  $p_{1f}$ , measure  $p$ , and from that estimate  $M$ . Solving the above for  $M$ :

$$\ln(1-p) = M \ln(1-p_{1f}), \quad M = \frac{\ln(1-p)}{\ln(1-p_{1f})}.$$

Larger  $M$  is more demanding on your data.  
Being conservative on a claim of detection means favoring larger  $M$ .

In most period-searching methods (except for the DFT), we are free to search as many frequencies as we like, at as dense a trial frequency spacing as we like. We call our significance parameter  $\theta = \theta(f)$ , because it is a function of frequency. Intuitively, we expect that two very close frequencies will produce similar  $\theta$  values, and indeed, such  $\theta$ -values are correlated (in the precise statistical sense). So our problem reduces to determining  $M$ , the number of *independent* frequencies in our arbitrary set of frequencies.

The bottom line is that, for dense trial frequencies,  $M$  is approximately the same as if we had equally spaced samples, and therefore a simple DFT [Press 1988]. Such a DFT has independent frequency components. This simple-sounding result, however, requires understanding a few subtleties, especially when the trial frequencies are sparse.

We consider 3 cases, starting with the simplest:

- Uniformly space data points, uniformly spaced frequencies (i.e., a DFT).
- Arbitrarily spaced data points, uniformly spaced frequencies.
- Arbitrarily space data points, arbitrarily spaced frequencies.

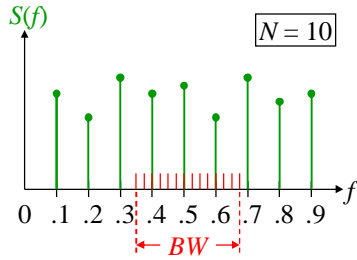
**Notation:**

$\theta$	significance parameter, such as Lomb-Scargle, Phase Dispersion Minimization, etc.
$\Delta f$	the independent frequency spacing.
$N$	number of data samples.
$M$	number of <i>independent</i> $\theta$ values over our chosen set of frequencies.
$BW$	the total range of frequencies tried: $BW \equiv f_{max} - f_{min}$ .
$T$	the total duration of samples: $T \equiv t_{max} - t_{min}$ .

**Equally spaced samples:** If our samples are equally spaced, we have the common case of a Discrete Fourier Transform (DFT). For white (i.e., uncorrelated) noise, each frequency component is independent of all the others. Furthermore, in the relevant notation (where all frequencies are positive), the maximum number of frequencies, and their spacing, is:

$$\text{max \# DFT frequencies} = N/2, \quad \Delta f = 1/T.$$

Note that the maximum number of frequencies depends *only* on the number of data points,  $N$ , and not on  $T$ . However, we may be looking for frequencies only in some range (Figure 9.2).



**Figure 9.5** Sample frequency spectrum for uniformly spaced discrete time data (here  $\Delta f = 0.1$ ).  $BW$  defines a subset of frequencies (here  $BW = 0.325$ ).

Therefore, for dense trial frequencies ( $\Delta f_{trial} < \Delta f$ ), the number of independent frequencies is approximately:

$$M \approx BW / \Delta f = (BW)T = 0.325/0.1 = 3.25 \rightarrow 4 .$$

We round  $M$  up to be conservative.

**Arbitrarily spaced samples:** In astronomy, the data times are rarely uniformly spaced. In such cases, we usually choose our trial frequency spacing,  $\Delta f_{trial}$ , to be dense, i.e., smaller than the independent frequency spacing:  $\Delta f_{trial} < \Delta f \approx 1/T$ . Then, per [Press 1988], we use the same equations as above to approximate  $\Delta f$  and  $M$ :

$$\Delta f \approx 1/T, \quad M \approx BW / \Delta f = (BW)T, \quad (\Delta f_{trial} < \Delta f) . \tag{9.4}$$

Note that this is true even if  $BW >$  Nyquist frequency, which is perfectly valid for nonuniformly spaced time samples [Press 1988].

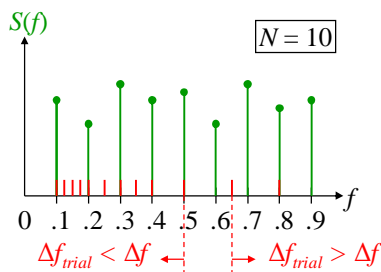
In the unusual case that our trial frequency spacing is large,  $\Delta f_{trial} > \Delta f$ , then we approximate that each frequency is independent:

$$M \approx \# \text{ trial frequencies}, \quad (\Delta f_{trial} > \Delta f) . \tag{9.5}$$

(In reality, even if  $\theta$  values separated by  $\Delta f$  are truly independent, some  $\theta$  values separated by more than  $\Delta f$  will be somewhat correlated. However, the correlation coefficient “envelope” decreases with increasing frequency spacing. Nonetheless, these correlations imply that there are parasitic cases where this approximation, eq. (9.5), fails.)

**Arbitrarily spaced trial frequencies:** One common situation leading to nonuniformly spaced trial frequencies is that of *uniformly* spaced trial *periods*. If the ratio of highest to lowest period is large (say,  $> 2$ ), then the frequency spacing is seriously nonuniform.

We may think of  $\Delta f$  as approximately the difference in frequency required to make the  $\theta$  values independent. (In reality, even if  $\theta$  values separated by  $\Delta f$  are truly independent, *some*  $\theta$  values separated by more than  $\Delta f$  will be somewhat correlated. However, the correlation coefficient “envelope” decreases with increasing frequency spacing.) In such a case, we may break up the trial frequencies into (1) regions where  $\Delta f_{trial} < \Delta f$ , and (2) regions where  $\Delta f_{trial} > \Delta f$  (Figure 9.6).



**Figure 9.6** Nonuniformly spaced trial frequencies.



The region where  $\Delta f_{\text{trial}} < \Delta f$  behaves as before, as does the region where  $\Delta f_{\text{trial}} > \Delta f$ . In the example of Figure 9.6, we have:

$$\Delta f = 0.1, \quad M \approx [(0.5 - 0.1)/0.1] + 2 = 6.$$

### Summary

Bandwidth correction requires estimating the number of *independent* frequencies. For uniformly spaced, dense trial frequencies (and arbitrarily spaced time samples), we approximate the number of independent frequencies,  $M$ , with eq. (9.4). We may think loosely of  $\Delta f$  as the difference in frequency required for  $\theta$  to become independent of its predecessor. Therefore, for nonuniformly spaced trial frequencies, we must consider two types of region: (1) where the trial frequency spacing  $\Delta f_{\text{trial}} < \Delta f$ , we use eq. (9.4); (2) where the trial frequency spacing  $\Delta f_{\text{trial}} > \Delta f$ , we approximate  $M$  as the number of trial frequencies, eq. (9.5).

### References

- [1] Press, William H. and George B. Rybicki, *Fast Algorithm for Spectral Analysis of Unevenly Sampled Data*, *Astrophysics Journal*, 338:277-280, 1989 March 1.
- [2] [http://www.ltrr.arizona.edu/~dmeko/notes\\_6.pdf](http://www.ltrr.arizona.edu/~dmeko/notes_6.pdf), retrieved 1/22/2012.
- [3] Press, William H. and Saul A. Teukolsky, *Search Algorithm for Weak Periodic Signals in Unevenly Spaced Data*, *Computers in Physics*, Nov/Dec 1988, p77.
- [4] Scargle, Jeffrey, *Studies in Astronomical Time Series Analysis. II. Statistical Aspects of Spectral Analysis of Unevenly Spaced Data*, *Astrophysical Journal*, 263:835-853, 12/15/1982.
- [5] Lomb, N. R., *Least Squares Frequency Analysis of Unequally Spaced Data*, *Astrophysics and Space Science* 39 (1976) 447-462.

Schwarzenberg-Czerny, A., *The Correct Probability Distribution for the Phase Dispersion Minimization Periodogram*, *The Astrophysical Journal*, 489:941-945, 1997 November 10.

---

## Analytic Signals and Hilbert Transforms

Given some real-valued signal,  $s(t)$ , it is often convenient to write it in “phasor form.” Such uses arise in diverse signal processing applications from communication systems to neuroscience. We describe here the meaning of “analytic signals,” and some practical computation considerations. This section relies heavily on phasor concepts, which you can learn from *Funky Electromagnetic Concepts*. We proceed along these lines:

- Mathematical definitions and review.
- The meaning of the analytic signal,  $A(t)$ .
- Instantaneous values.
- Finding  $A(t)$  from the signal  $s(t)$ , theoretically.
- The special case of zero reference frequency,  $\omega_0 = 0$ ; Hilbert Transform.
- A simple and reliable numerical computation of  $A(t)$  without Hilbert Transforms.

**Definitions, conventions, and review:** There are many different conventions in the literature for normalization and sign of the Fourier Transform (FT). We define the Fourier Transform such that our basis functions are  $e^{i\omega t}$ , and our original (possibly complex) signal  $z(t)$  is composed from them; this fully defines all the normalization and sign conventions:

$$\text{For } z(t) \text{ complex:} \quad z(t) \equiv \int_{-\infty}^{\infty} Z(\omega) e^{+i\omega t} d\omega \quad \Rightarrow \quad Z(\omega) = \frac{1}{2\pi} \int_{-\infty}^{\infty} z(t) e^{-i\omega t} dt$$

where  $Z(\omega) \equiv \mathcal{F}\{z(t)\}$  is the Fourier Transform of  $z(t)$ .

Note that we can think of the FT as a phasor-valued function of frequency, and that we use the positive time dependence  $e^{+i\omega t}$ .

For real-valued signals we use  $s(t)$  instead of  $z(t)$ . For real  $s(t)$ , the FT is **conjugate symmetric**:

$$S(-\omega) = S^*(\omega) \quad \text{for } s(t) \text{ real.}$$

This conjugate symmetry for real signals allows us to use a 1-sided FT, where we consider only positive frequencies, so that:

$$s(t) = 2\text{Re}\left\{\int_0^\infty S(\omega)e^{i\omega t} d\omega\right\}, \quad \text{which is equivalent to } s(t) = \int_{-\infty}^\infty S(\omega)e^{i\omega t} d\omega, \quad s(t) \text{ real.}$$

Note that a *complex* signal with no negative frequencies is very different from a *real* signal which we choose to write as a 1-sided FT. We rely on this distinction in the following discussion.

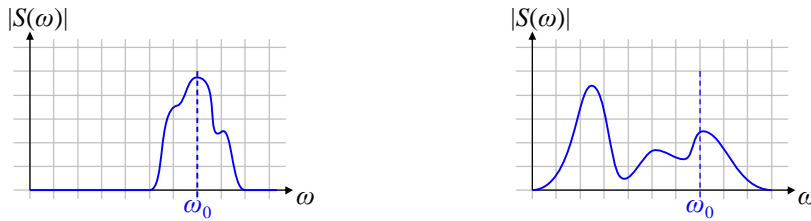
**Analytic signals:** Given a real-valued signal,  $s(t)$ , its phasor form is:

$$s(t) = \text{Re}\left\{A(t)e^{i\omega_0 t}\right\} = |A(t)|\cos(\omega_0 t + \phi(t))$$

where  $A(t) \equiv |A(t)|e^{i\phi(t)}$  is a (complex) phasor function of time (9.6)

$\omega_0 \equiv$  somewhat arbitrary reference frequency .

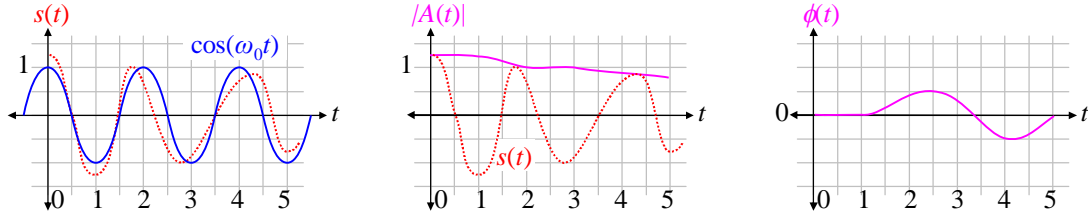
Recall that as a phasor,  $A(t)$  is complex. The phasor form of  $s(t)$  may be convenient when  $s(t)$  is **band-limited** (exists only in a well-defined range of frequencies, Figure 9.7 left), or where we are only concerned with the components of  $s(t)$  in some well-defined range of frequencies. Figure 9.7 shows two 1-sided Fourier Transform (FT) examples of  $S(\omega)$ , the FT of a hypothetical (real) signal  $s(t)$ .



**Figure 9.7** Example 1-sided FTs of a real signal  $s(t)$ : (Left) band-limited. (Right) Wideband. The  $\omega$  axis points only to the right, because we need consider only positive frequencies for a 1-sided FT.

In communication systems,  $\omega_0$  is the carrier frequency. Note that even in the band-limited case,  $\omega_0$  may be different than the band center frequency. [For example, in vestigial sideband modulation (VSB),  $\omega_0$  is close to one edge of the signal band.] Keep in mind throughout this discussion that  $\omega_0$  is often chosen to be zero, i.e. the spectrum of  $s(t)$  is kept “in place”.

We start by considering the band-limited case, because it is somewhat simpler. From Figure 9.7 (left), we see that our signal  $s(t)$  is not too different from a pure sinusoid at a reference frequency  $\omega_0$ , near the middle of the band.  $s(t)$  and  $\cos(\omega_0 t)$  might look like Figure 9.8, left.  $s(t)$  is a modulation of the pure sinusoid, varying somewhat (i.e. perturbing) the amplitude and phase at each instant in time. We define these variations as the complex function  $A(t)$ . When a signal  $s(t)$  is real, and  $A(t)$  satisfies eq. (9.6),  $A(t)$  is called the **analytic signal** for  $s(t)$ .



**Figure 9.8** (Left)  $s(t)$  (dotted), and the reference sinusoid. (Middle) Magnitude of the analytic signal  $|A(t)|$ . (Right) Phase of the analytic signal.

We can visualize  $A(t)$  by considering Figure 9.8, left. At  $t = 0$ ,  $s(t)$  is a little bigger than 1, but it is in-phase with the reference cosine; this is reflected in the amplitude  $|A(0)|$  being slightly greater than 1, and the phase  $\phi(0) = 0$ . At  $t = 1$ , the amplitude remains  $> 1$ , and  $\phi$  is still 0. At  $t = 2$ , the amplitude has dropped to 1, and the phase  $\phi(2)$  is now positive (early, or leading). This continues through  $t = 3$ . At  $t = 4$ , the amplitude drops further to  $|A(4)| < 1$ , and the phase is now negative (late, or lagging), i.e.  $\phi(4) < 0$ . At  $t = 5$ , the amplitude remains  $< 1$ , while the phase has returned to zero:  $\phi(5) = 0$ . Figure 9.8, middle and right, are plots of these amplitudes and phases.

**Instantaneous values:** When a signal has a clear oscillatory behavior, one can meaningfully define instantaneous values of phase, frequency, and amplitude. Note that the frequency of a sinusoid (in rad/s) is the rate of change of the phase (in rad) with time. A general signal  $s(t)$ , has a varying phase  $\phi(t)$ , aka an **instantaneous phase**. Therefore, we can define an **instantaneous frequency**, as well:

$$phase = \omega_0 t + \phi(t) \quad \Rightarrow \quad \omega(t) = \frac{d(phase)}{dt} = \omega_0 + \frac{d\phi}{dt}.$$

Such an instantaneous frequency is more meaningful when  $|A(t)|$  is fairly constant over one or more periods. For example, in FM radio (frequency modulation),  $|A(t)|$  is constant for all time, and *all* of the information is encoded in the instantaneous frequency.

Finally, we similarly define the **instantaneous amplitude** of a signal  $s(t)$  as  $|A(t)|$ . This is more meaningful when  $|A(t)|$  is fairly constant over one or more cycles of oscillation. The instantaneous amplitude is the “envelope” which bounds the oscillations of  $s(t)$  (Figure 9.8, middle). By construction,  $|s(t)| < |A(t)|$  everywhere.

**Finding A(t) from s(t):** Given an arbitrary  $s(t)$ , how do we find its (complex) analytic signal,  $A(t)$ ? First, we see that the defining eq. (9.6),  $s(t) = \text{Re}\{A(t)e^{i\omega_0 t}\}$ , is underdetermined, since  $A(t)$  has two real components, but is constrained by only the one equation. Therefore, if  $A(t)$  is to be unique, we must further constrain it.

As a simple starting point, suppose  $s(t)$  is a pure cosine (we will generalize shortly). Then:

$$s(t) = \cos \omega_0 t = \text{Re}\{1 e^{i\omega_0 t}\} \quad \text{where} \quad A(t) = 1.$$

If instead,  $s(t)$  has a phase offset  $\theta$ , then:

$$s(t) = \cos(\omega_0 t + \theta) = \text{Re}\{e^{i\theta} e^{i\omega_0 t}\} \quad \text{where} \quad A(t) = e^{i\theta} = \cos \theta + i \sin \theta.$$

(Note that  $\theta = 0$  reproduces the pure-cosine example above.) Thus, in the case of a pure sinusoid,  $A \equiv A(t)$  is the (constant) phasor for the sinusoid  $s(t)$ , and the imaginary part of  $A$  is the same sinusoid delayed by  $1/4$  cycle ( $90^\circ$ ):

$$\text{Re}\{A\} = \cos \theta, \quad \text{Im}\{A\} = \cos(\theta - \pi/4).$$

In Fourier space, the real and imaginary parts of  $A$  are simply related. Recall that delaying a sinusoid by  $1/4$  cycle multiplies its Fourier component by  $-i$  (for  $\omega > 0$ ). Therefore:

$$\mathcal{F}\{\text{Im}(A(t))\} = -i\mathcal{F}\{\text{Re}(A(t))\}, \quad \text{1-sided FT, } \omega > 0.$$

We now generalize our pure sinusoid example to an arbitrary signal, which can be thought of as a linear combination sinusoids. The above relation is linear, so it holds for any linear combination of sinusoids, i.e. it holds for any real signal  $s(t)$ . This means that, by construction, the imaginary part of  $A(t)$  has exactly the same *magnitude* spectrum as the real part of  $A(t)$ . Also, the imaginary part has a phase spectrum which is everywhere  $\frac{1}{4}$  cycle delayed from the phase spectrum of the real part. This is the relationship that uniquely defines the analytic signal  $A(t)$  that corresponds to a given real signal  $s(t)$  and a given reference frequency  $\omega_0$ . From this relation, we can solve for  $\text{Im}\{A(t)\}$  explicitly as a functional of  $\text{Re}\{A(t)\}$ :

$$\text{Im}\{A(t)\} = \mathcal{F}^{-1}\{-i\mathcal{F}\{\text{Re}(A(t))\}\}, \quad \text{1-sided FT, } \omega > 0. \tag{9.7}$$

This relation defines the **Hilbert Transform** (HT) from  $\text{Re}\{A(t)\}$  to  $\text{Im}\{A(t)\}$ .

The Hilbert Transform of  $s(t)$  is a function  $H(t)$  that has all the Fourier components of  $s(t)$ , but delayed in phase by  $\frac{1}{4}$  cycle ( $90^\circ$ ).

Note that the Hilbert transform takes a function of time into another function of time (in contrast to the Fourier Transform that takes a function of time into a function of frequency). Since the FT is linear, eq. (9.7) shows that the HT is also linear. The Hilbert Transform can be shown to be given by the time-domain form:

$$\mathcal{H}\{s(t)\} \equiv H(t) = \frac{1}{\pi} \text{PV} \int_{-\infty}^{\infty} dt' \frac{s(t')}{t-t'}, \quad \text{where PV} \equiv \text{Principal Value}.$$

(The integrand blows up at  $t' = t$ , which is why we need the Principal Value to make the integral well-defined.) We now easily show that the inverse Hilbert transform is the negative of the forward transform:

$$\mathcal{H}^{-1}\{H(t)\} \equiv s(t) = -\frac{1}{\pi} \text{PV} \int_{-\infty}^{\infty} dt' \frac{H(t')}{t-t'}, \quad \text{where PV} \equiv \text{Principal Value}.$$

We see this because the Hilbert Transform shifts the phase of every sinusoid by  $90^\circ$ . Therefore, two Hilbert transforms shifts the phase by  $180^\circ$ , equivalent to negating every sinusoid, which is equivalent to negating the original signal. Putting in a minus sign then restores the original signal.

Equivalently, the HT multiplies each Fourier component ( $\omega > 0$ ) by  $-i$ . Then  $\mathcal{H}\{\mathcal{H}(\cdot)\}$  multiplies each component by  $(-i)^2 = -1$ . Thus,  $\mathcal{H}\{\mathcal{H}[s(t)]\} = -s(t)$ , and therefore  $\mathcal{H}^{-1} = -\mathcal{H}$ .

**Analytic signal relative to  $\omega_0 = 0$ :** Some analysts prefer not to remove a reference frequency  $e^{i\omega_0 t}$  from the signal, and instead include all of the phase in  $A(t)$ ; this is equivalent to choosing  $\omega_0 = 0$ :

$$s(t) = \text{Re}\{A(t)\} = |A(t)|\cos(\phi(t)).$$

Since  $s(t) = \text{Re}\{A(t)\}$  is given, we can now find  $\text{Im}\{A(t)\}$  explicitly from (9.7):

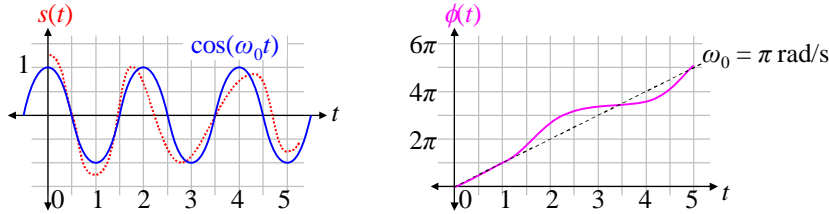
$$\text{Im}\{A(t)\} = \mathcal{F}^{-1}\{-i\mathcal{F}\{s(t)\}\} \equiv \mathcal{H}\{s(t)\} \quad \text{1-sided FT, } \omega > 0.$$

In other words:

For  $\omega_0 = 0$ ,  $A(t)$  is just the complex phasor factors for  $s(t)$ , without taking any real part.

If  $s(t)$  is dominated by a single frequency  $\omega$ , then  $\phi(t)$  contains a fairly steady phase ramp that is close to  $\phi(t) \approx \omega t$  (Figure 9.9). We can use the phase function  $\phi(t)$  to estimate the frequency  $\omega$  by simply taking the average phase *rate* over our sample interval:

$$\omega_{est} = \frac{\phi(t_{end}) - \phi(0)}{t_{end}}.$$



**Figure 9.9** Phase ramp of a perturbed sinusoid, and the estimate of  $\omega_0$ .

**Efficient numerical computation of  $A(t)$ :** The traditional way to find  $A(t)$  is to use a discrete Hilbert Transform to evaluate the defining integral. (This is a standard function in scientific software packages.) The discrete Hilbert Transform (DHT) is often implemented by taking a DFT, manipulating it, and then inverse FT back to the time domain. This can be seen by recasting our above (1-sided DFT) description of the Hilbert Transform (HT) into a 2-sided DFT form.

Recall that in the 1-sided DFT for a real signal  $s(t)$ , the frequencies are always positive,  $\omega > 0$ , and  $S(\omega)$  is just a phasor-valued function of frequency. To recover the real signal from phasors, we must take a real-part,  $\text{Re}\{ \}$ . In the 2-sided DFT, we instead arrive at the real part by adding the complex conjugate of all the phasor factors:

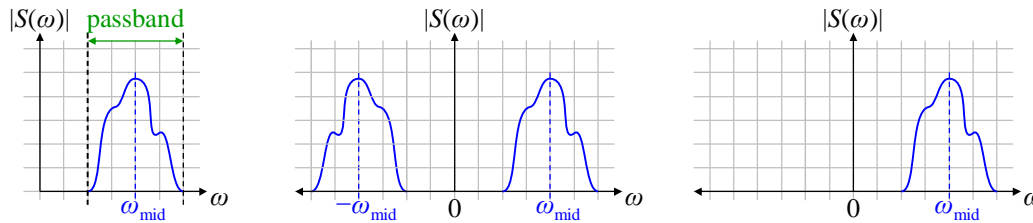
$$s(t) = 2 \int_0^\infty d\omega \text{Re}\{S(\omega)e^{+i\omega t}\} \quad \rightarrow \quad s(t) = \int_0^\infty d\omega [S(\omega)e^{+i\omega t} + S^*(\omega)e^{-i\omega t}].$$

However, to achieve a 2-sided FT, we rewrite the second term as a negative frequency. Then the integral spans both positive and negative frequencies:

$$s(t) = \int_{-\infty}^\infty S(\omega)e^{i\omega t} d\omega, \quad \text{where} \quad S(-\omega) = S^*(\omega).$$

For a complex signal,  $z(t)$ , only a 2-sided FT exists (a 1-sided FT is not generally possible). Then there is no symmetry or relation between positive and negative frequencies.

We now describe a simple, efficient, and stable, purely time-domain algorithm for finding  $A(t)$  from a band-limited  $s(t)$ . This algorithm is sometimes more efficient than the DFT-based approach. It is especially useful when the data must be downsampled (converted to a lower sampling rate by keeping only every  $n^{\text{th}}$  sample, called **decimating**). Even though  $s(t)$  is real, the algorithm uses complex arithmetic throughout.



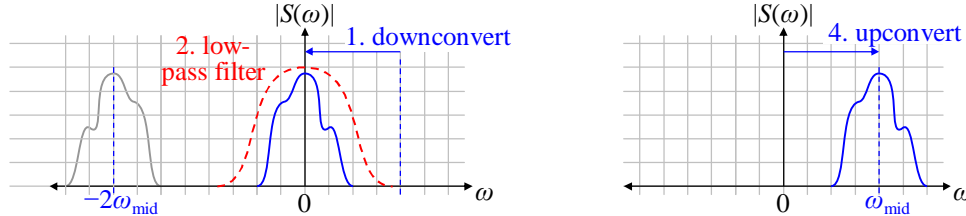
**Figure 9.10** (Left) 1-sided FT of  $s(t)$ , and (middle) its 2-sided equivalent. (Right) 2-sided FT of  $A(t)$ .

Figure 9.10 shows a 1-sided FT for a real  $s(t)$ , along with its 2-sided FT equivalent, and the 2-sided FT of the desired complex  $A(t)$ . We define  $\omega_{mid}$  as the midpoint of the signal band (this is *not*  $\omega_0$ , which we take to be zero for illustration). The question is: how do we efficiently go from the middle diagram to the right diagram? In other words, how do we keep just the positive frequency half of the 2-sided spectrum? Figure 9.11 illustrates the simple steps to achieve this:

- Rotate the spectrum down by  $\omega_{mid}$  (downconvert).
- Low-pass filter around the downconverted signal band.

- (Optional) Decimate (downsample).
- Rotate the spectrum back up by  $\omega_{mid}$  (upconvert).

This results in a complex function of time whose 2-sided spectrum has only positive frequencies; in other words, it is exactly the analytic signal  $A(t)$ .



**Figure 9.11** (Left) To find  $A(t)$ : 1. downconvert; 2. low-pass filter; (Right) 4. upconvert.

**Step 1: Downconvert:** Numerically, we downconvert in the time domain by multiplying by  $\exp(-i\omega_{mid}t)$ . This simply subtracts  $\omega_{mid}$  from the frequency of each component in the spectrum:

$$\text{For every } \omega: \quad S(\omega)e^{i\omega t} e^{-i\omega_{mid} t} = S(\omega)e^{i(\omega-\omega_{mid})t} .$$

Note that both positive and negative frequencies are shifted to the left (more negative) in frequency. In the time domain, we construct the complex downconverted signal for each sample time  $t_j$  as:

$$z_{down}(t_j) = s(t_j)\exp(-i\omega_{mid}t_j) = s(t_j)\cos(\omega_{mid}t_j) - i\sin(\omega_{mid}t_j)$$

**Step 2: Low-pass filter:** Low pass filters are easily implemented as Finite Impulse Response (FIR) filters, with symmetric filter coefficients [Ham chap. 6, 7]. In the time domain:

$$A_{down}(t_j) = 2 \sum_{k=-m}^m c_k z_{down}(t_{j+k}) \quad \text{where } 2m+1 \equiv \text{the number of filter coefficients}$$

$$c_k \equiv \text{filter coefficients}$$

The leading factor of 2 is to restore the full amplitude to  $A(t)$  after filtering out half the frequency components.

**Step 3: (Optional) Decimate:** We now have a (complex) low-pass signal whose full (2-sided) bandwidth is just that of our desired signal band. If desired, we can now downsample (decimate), by simply keeping every  $n^{\text{th}}$  sample. In other words, our low-pass filter acts as both a pass-band filter for the desired signal, and an anti-aliasing filter for downsampling. Two for the price of one.

**Step 4: Upconvert:** We now restore our complex analytic signal to a reference frequency of  $\omega_0 = 0$  by putting the spectrum back where it came from. The key distinction is that after upconverting, there will be no components of negative frequency because we filtered them out in Step 2. This provides our desired complex analytic signal:

$$A(t_j) = A_{down}(t_j)\exp(i\omega_{mid}t_j).$$

Note that the multiplications above are full complex multiplies, because both  $A_{down}$  and the exponential factor are complex. Also, if we want some nonzero  $\omega_0$ , we would simply upconvert by  $(\omega_{mid} - \omega_0)$  instead of upconverting by  $\omega_{mid}$ .

### Summary

The analytic signal for a real function  $s(t)$  is  $A(t)$ , and is the complex phasor-form of  $s(t)$  such that:

$$s(t) = \text{Re}\{A(t)\exp(i\omega_0 t)\} \quad \text{where } \omega_0 \equiv \text{reference frequency} .$$

$\omega_0$  is often chosen to be zero, so that  $s(t) = \text{Re}\{A(t)\}$ . This definition does *not* uniquely define  $A(t)$ , since  $A(t)$  has real and imaginary components, but is constrained by only one equation. The Hilbert Transform of a real function  $s(t)$  is  $H(t)$ , and comprises all the Fourier components of  $s(t)$  phase-delayed by  $\pi/4$  radians ( $90^\circ$ ). We uniquely define  $A(t)$  by saying that its imaginary part is the Hilbert Transform of its real part. This gives the imaginary part the exact same magnitude spectrum as the real part, but a shifted phase spectrum.

Analytic signals allow defining instantaneous values of frequency, phase, and amplitude for almost-sinusoidal signals. Instantaneous values are useful in many applications, including communication and neuron behavior.

[Ham]           Hamming, R. W., *Digital Filters*, Dover Publications (July 10, 1997), ISBN-13: 978-0486650883.

## 10 Tensors, Without the Tension

### Approach

We'll present tensors as follows:

1. Two physical examples: magnetic susceptibility, and deformable solids
2. A non-example: when is a matrix not a tensor?
3. Forward looking definitions (don't get stuck on these)
4. Review of vector spaces and notation (don't get stuck on this, either)
5. A short, but at first unhelpful, definition (really, really don't get stuck on this)
6. A discussion which clarifies the above definition
7. Examples, including dot products and cross-products as tensors
8. Higher rank tensors
9. Change of basis
10. Non-orthonormal systems: contravariance and covariance
11. Indefinite metrics of Special and General Relativity
12. Mixed basis linear functions (transformation matrices, the Pauli vector)

Tensors are all about vectors. They let you do things with vectors you never thought possible. We define tensors in terms of what they do (their linearity properties), and then show that linearity *implies* the transformation properties. This gets most directly to the true importance of tensors. [Most references define tensors in terms of transformations, but then fail to point out the all-important linearity properties.]

We also take a geometric approach, treating vectors and tensors as geometric objects that exist independently of their representation in any basis. Inevitably, though, there is a fair amount of unavoidable algebra.

Later, we introduce contravariance and covariance in terms of non-orthonormal coordinates, but first with a familiar positive-definite metric from classical mechanics. This makes for a more intuitive understanding of contra- and co-variance, before applying the concept to the more bizarre indefinite metrics of special and general relativity.

There is deliberate repetition of several points, because it usually takes me more than once to grok something. So I repeat:

If you don't understand something, read it again once, then keep reading. Don't get stuck on one thing. Often, the following discussion will clarify an ambiguity.

---

### Two Physical Examples

We start with two physical examples: magnetic susceptibility, and deformation of a solid. We start with matrix notation, because we assume it is familiar to you. Later we will see that matrix notation is not ideal for tensor algebra.

#### Magnetic Susceptibility

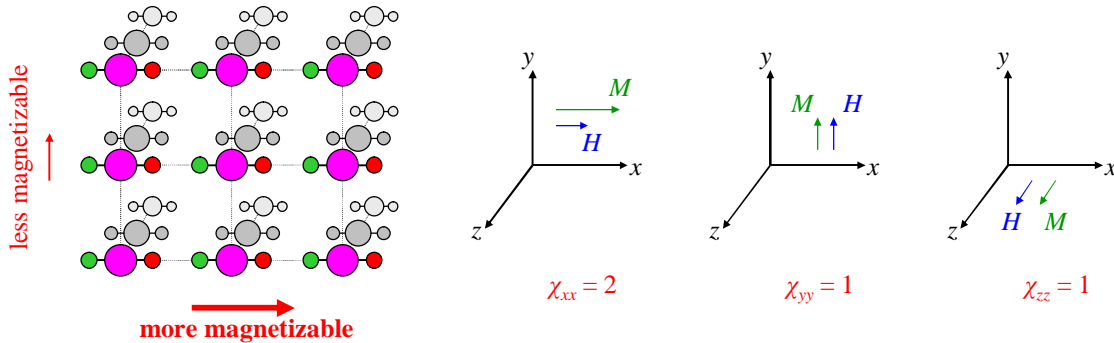
We assume you are familiar with **susceptibility** of magnetic materials: when placed in an H-field, magnetizable (susceptible) materials acquire a magnetization, which adds to the resulting B-field. In simple cases, the susceptibility  $\chi$  is a scalar, and



$\mathbf{M} = \chi \mathbf{H}$       where  $\mathbf{M}$  is the magnetization,  
 $\chi$  is the susceptibility, and  
 $\mathbf{H}$  is the applied magnetic field

The susceptibility in this simple case is the same in any direction; i.e., the material is isotropic.

However, there exist materials which are more magnetizable in some directions than others. E.g., imagine a cubic lattice of axially-symmetric molecules which are more magnetizable along the molecular axis than perpendicular to it:



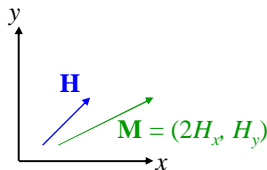
Magnetization,  $M$ , as a function of external field,  $H$ , for a material with a tensor-valued susceptibility,  $\chi$ .

In each direction, the magnetization is proportional to the applied field, but  $\chi$  is larger in the  $x$ -direction than  $y$  or  $z$ . In this example, for an arbitrary  $H$ -field, we have

$$\mathbf{M} = (M_x, M_y, M_z) = (2H_x, H_y, H_z) \quad \text{or} \quad \mathbf{M} = \chi \mathbf{H} = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{H}$$

$\chi \equiv \chi_{ij}$

Note that in general,  $\mathbf{M}$  is not parallel to  $\mathbf{H}$  (below, dropping the  $z$  axis for now):



$\mathbf{M}$  need not be parallel to  $\mathbf{H}$  for a material with a tensor-valued  $\chi$ .

But  $\mathbf{M}$  is a linear function of  $\mathbf{H}$ , which means:  $\mathbf{M}(k\mathbf{H}_1 + \mathbf{H}_2) = k\mathbf{M}(\mathbf{H}_1) + \mathbf{M}(\mathbf{H}_2)$ .

This linearity is reflected in the fact that matrix multiplication is linear:

$$\mathbf{M}(k\mathbf{H}_1 + \mathbf{H}_2) = \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} (k\mathbf{H}_1 + \mathbf{H}_2) = k \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{H}_1 + \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \mathbf{H}_2 = k\mathbf{M}(\mathbf{H}_1) + \mathbf{M}(\mathbf{H}_2)$$

The matrix notation might seem like overkill, since  $\chi$  is diagonal, but it is *only* diagonal in this basis of  $x$ ,  $y$ , and  $z$ . We'll see in a moment what happens when we change basis. First, let us understand what the matrix  $\chi_{ij}$  really means. Recall the visualization of pre-multiplying a vector by a matrix: a matrix  $\chi$  times a column vector  $\mathbf{H}$ , is a weighted sum of the columns of  $\chi$ :

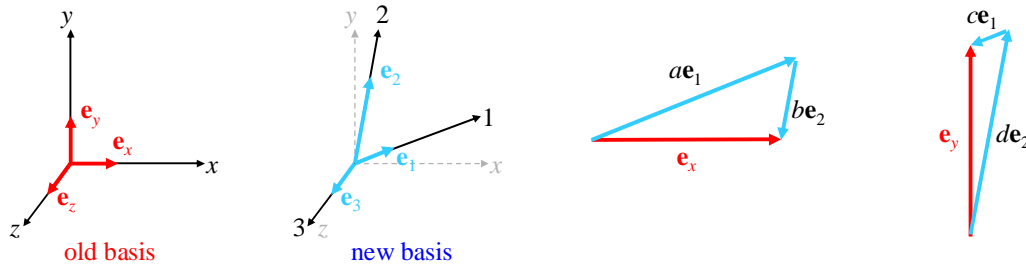
$$\chi\mathbf{H} = \begin{bmatrix} \chi_{xx} & \chi_{xy} & \chi_{xz} \\ \chi_{yx} & \chi_{yy} & \chi_{yz} \\ \chi_{zx} & \chi_{zy} & \chi_{zz} \end{bmatrix} \begin{bmatrix} H_x \\ H_y \\ H_z \end{bmatrix} \equiv H_x \begin{bmatrix} \chi_{xx} \\ \chi_{yx} \\ \chi_{zx} \end{bmatrix} + H_y \begin{bmatrix} \chi_{xy} \\ \chi_{yy} \\ \chi_{zy} \end{bmatrix} + H_z \begin{bmatrix} \chi_{xz} \\ \chi_{yz} \\ \chi_{zz} \end{bmatrix}$$

We can think of the matrix  $\chi$  as a set of 3 column vectors: the first is the magnetization vector for  $\mathbf{H} = \mathbf{e}_x$ ; the 2<sup>nd</sup> column is  $\mathbf{M}$  for  $\mathbf{H} = \mathbf{e}_y$ ; the 3<sup>rd</sup> column is  $\mathbf{M}$  for  $\mathbf{H} = \mathbf{e}_z$ . Since magnetization is linear in  $\mathbf{H}$ , the magnetization for any  $\mathbf{H}$  can be written as the weighted sum of the magnetizations for each of the basis vectors:

$$\mathbf{M}(\mathbf{H}) = H_x \mathbf{M}(\mathbf{e}_x) + H_y \mathbf{M}(\mathbf{e}_y) + H_z \mathbf{M}(\mathbf{e}_z) \quad \text{where } \mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z \text{ are the unit vectors in } x, y, z$$

This is just the matrix multiplication above:  $\mathbf{M} = \chi\mathbf{H}$ . (We're writing all indexes as subscripts for now; later on we'll see that  $\mathbf{M}$ ,  $\chi$ , and  $\mathbf{H}$  should be indexed as  $M^i$ ,  $\chi^i_j$ , and  $H^i$ .)

Now let's change bases from  $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ , to some  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ , defined below. We use a simple transformation, but the 1-2-3 basis is *not* orthonormal:



Transformation to a non-orthogonal, non-normal basis.  $\mathbf{e}_1$  and  $\mathbf{e}_2$  are in the  $x$ - $y$  plane, but are neither orthogonal nor normal. For simplicity, we choose  $\mathbf{e}_3 = \mathbf{e}_z$ . Here,  $b$  and  $c$  are negative.

To find the transformation equations to the new basis, we first write the old basis vectors in the new basis. We've chosen for simplicity a transformation in the  $x$ - $y$  plane, with the  $z$ -axis unchanged:

$$\mathbf{e}_x = a\mathbf{e}_1 + b\mathbf{e}_2 \quad \mathbf{e}_y = c\mathbf{e}_1 + d\mathbf{e}_2 \quad \mathbf{e}_z = \mathbf{e}_3$$

Now write a vector,  $\mathbf{v}$ , in the old basis, and substitute out the old basis vectors for the new basis. We see that the new components are a linear combination of the old components:

$$\begin{aligned} \mathbf{v} &= v_x \mathbf{e}_x + v_y \mathbf{e}_y + v_z \mathbf{e}_z = v_x (a\mathbf{e}_1 + b\mathbf{e}_2) + v_y (c\mathbf{e}_1 + d\mathbf{e}_2) + v_z \mathbf{e}_3 \\ &= (av_x + cv_y)\mathbf{e}_1 + (bv_x + dv_y)\mathbf{e}_2 + v_z \mathbf{e}_3 = v_1 \mathbf{e}_1 + v_2 \mathbf{e}_2 + v_3 \mathbf{e}_3 \\ \Rightarrow \quad v_1 &= av_x + cv_y, \quad v_2 = bv_x + dv_y, \quad v_3 = v_z \end{aligned}$$

Recall that matrix multiplication is defined to be the operation of linear transformation, so we can write this basis transformation in matrix form:

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix} = \begin{pmatrix} a & c & 0 \\ b & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} = v_x \begin{bmatrix} a \\ b \\ 0 \end{bmatrix} + v_y \begin{bmatrix} c \\ d \\ 0 \end{bmatrix} + v_z \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$\mathbf{e}_x$                        $\mathbf{e}_y$                        $\mathbf{e}_z$

The columns of the transformation matrix are the old basis vectors written in the new basis.

This is illustrated explicitly on the right hand side, which is just  $v_x \mathbf{e}_x + v_y \mathbf{e}_y + v_z \mathbf{e}_z$ .

Finally, we look at how the susceptibility matrix  $\chi_{ij}$  transforms to the new basis. We saw above that the columns of  $\chi$  are the  $\mathbf{M}$  vectors for  $\mathbf{H}$  = each of the basis vectors. So right away, we must transform each column of  $\chi$  with the transformation matrix above, to convert it to the new basis. Since matrix multiplication  $\mathbf{A} \cdot \mathbf{B}$  is distributive across the columns of  $\mathbf{B}$ , we can write the transformation of all 3 columns in a single expression by pre-multiplying with the above transformation matrix:

$$\text{Step 1 of } \chi^{new} \equiv \chi \text{ in new basis} = \begin{pmatrix} a & c & 0 \\ b & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \chi = \begin{pmatrix} a & c & 0 \\ b & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 2a & c & 0 \\ 2b & d & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

But we're not done. This first step expressed the column vectors in the new basis, but the columns of the RHS (right hand side) are still the  $\mathbf{M}$ 's for basis vectors  $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ . Instead, we need the columns of  $\chi^{new}$  to be the  $\mathbf{M}$  vectors for  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ . Please don't get bogged down yet in the details, but we do this transformation similarly to how we transformed the column vectors. We transform the contributions to  $\mathbf{M}$  due to  $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$  to that due to  $\mathbf{e}_1$  by writing  $\mathbf{e}_1$  in terms of  $\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z$ :

$$\mathbf{e}_1 = e\mathbf{e}_x + f\mathbf{e}_y \quad \Rightarrow \quad \mathbf{M}(\mathbf{H} = \mathbf{e}_1) = e\mathbf{M}(\mathbf{H} = \mathbf{e}_x) + f\mathbf{M}(\mathbf{H} = \mathbf{e}_y)$$

Similarly,

$$\mathbf{e}_2 = g\mathbf{e}_x + h\mathbf{e}_y \quad \Rightarrow \quad \mathbf{M}(\mathbf{H} = \mathbf{e}_2) = g\mathbf{M}(\mathbf{H} = \mathbf{e}_x) + h\mathbf{M}(\mathbf{H} = \mathbf{e}_y)$$

$$\mathbf{e}_3 = \mathbf{e}_z \quad \Rightarrow \quad \mathbf{M}(\mathbf{H} = \mathbf{e}_3) = \mathbf{M}(\mathbf{H} = \mathbf{e}_z)$$

Essentially, we need to transform among the columns, i.e. transform the rows of  $\chi$ . These two transformations (once of the columns, and once of the rows) is the essence of a rank-2 tensor:

A tensor matrix (rank-2 tensor) has columns that are vectors, and simultaneously, its rows are also vectors. Therefore, transforming to a new basis requires two transformations: once for the rows, and once for the columns (in either order).

[Aside: The details (which you can skip at first): We just showed that we transform using the *inverse* of our previous transformation. The reason for the inverse is related to the up/down indexes mentioned earlier; please be patient. In matrix notation, we write the row transformation as post-multiplying by the transpose of the needed transformation:

$$\text{Final } \chi^{new} = \begin{pmatrix} a & c & 0 \\ b & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e & f & 0 \\ g & h & 0 \\ 0 & 0 & 1 \end{pmatrix}^T = \begin{pmatrix} a & c & 0 \\ b & d & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} e & g & 0 \\ f & h & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

]

[Another aside: A direction-dependent susceptibility requires  $\chi$  to be promoted from a scalar to a rank-2 tensor (skipping any rank-1 tensor). This is necessary because a rank-0 tensor (a scalar) and a rank-2 tensor can both act on a vector ( $\mathbf{H}$ ) to produce a vector ( $\mathbf{M}$ ). There is no sense to a rank-1 (vector) susceptibility, because there is no simple way a rank-1 tensor (a vector) can act on another vector  $\mathbf{H}$  to produce an output vector  $\mathbf{M}$ . More on this later.]

### Mechanical Strain

A second example of a tensor is the mechanical strain tensor. When I push on a deformable material, it deforms. A simple model is just a spring, with Hooke’s law:

$$\Delta x = + \frac{1}{k} F_{\text{applied}}$$

We write the formula with a plus sign, because (unlike freshman physics spring questions) we are interested in how a body deforms when *we* apply a force *to* it. For an isotropic material, we can push in any direction, and the deformation is parallel to the force. This makes the above equation a vector equation:

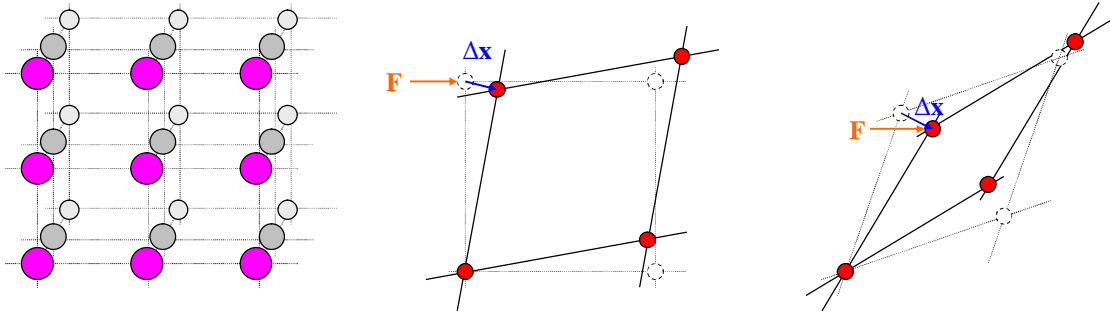
$$\Delta \mathbf{x} = s \mathbf{F} \quad \text{where} \quad s = \frac{1}{k} \equiv \text{the strain constant}$$

**Strain** is defined as the displacement of a given point under force. [**Stress** is the force per unit area applied to a body. Stress produces strain.] In an isotropic material, the stress constant is a simple scalar. Note that if we transform to another basis for our vectors, the stress constant is unchanged. That’s the definition of a scalar:

A scalar is a number that is the same in any coordinate system. A scalar is a rank-0 tensor.

The scalar is unchanged even in a non-ortho-normal coordinate system.

But what if our material is a bunch of microscopic blobs connected by stiff rods, like atoms in a crystal?



(Left) A constrained deformation crystal structure. (Middle) The deformation vector,  $\Delta \mathbf{x}$ , is not parallel to the force. (Right) More extreme geometries lead to a larger angle between the force and displacement.

The diagram shows a 2D example: pushing in the  $x$ -direction results in both  $x$  and  $y$  displacements. The same principle could result in a 3D  $\Delta \mathbf{x}$ , with some component into the page. For small deformations, the deformation is *linear* with the force: pushing twice as hard results in twice the displacement. Pushing with the sum of two (not necessarily parallel) forces results in the sum of the individual displacements. But the displacement is *not proportional* to the force (because the displacement is not parallel to it). In fact, each component of force results in a deformation *vector*. Mathematically:

$$\Delta \mathbf{x} = F_x \begin{bmatrix} s_{xx} \\ s_{yx} \\ s_{zx} \end{bmatrix} + F_y \begin{bmatrix} s_{xy} \\ s_{yy} \\ s_{zy} \end{bmatrix} + F_z \begin{bmatrix} s_{xz} \\ s_{yz} \\ s_{zz} \end{bmatrix} = \underbrace{\begin{pmatrix} s_{xx} & s_{xy} & s_{xz} \\ s_{yx} & s_{yy} & s_{yz} \\ s_{zx} & s_{zy} & s_{zz} \end{pmatrix}}_{\mathbf{s}} \begin{bmatrix} F_x \\ F_y \\ F_z \end{bmatrix} = \mathbf{s} \mathbf{F}$$

Much like the anisotropy of the magnetization in the previous example, the anisotropy of the strain requires us to use a rank-2 tensor to describe it. The *linearity* of the strain with force allows us to write the strain tensor as a matrix. Linearity also guarantees that we can change to another basis using a method

similar to that shown above for the susceptibility tensor. Specifically, we must transform both the columns and the rows of the strain tensor  $\mathbf{s}$ . Furthermore, the linearity of deformation with force also insures that we can use non-orthonormal bases, just as well as orthonormal ones.

### When Is a Matrix Not a Tensor?

I would say that most matrices are *not* tensors. A matrix *is* a tensor when its rows and columns are both vectors. This implies that there is a vector space, basis vectors, and the possibility of changing basis. As a counter example, consider the following graduate physics problem:

Two pencils, an eraser, and a ruler cost \$2.20. Four pencils, two erasers, and a ruler cost \$3.45. Four pencils, an eraser, and two rulers cost \$3.85. How much does each item cost?

We can write this as simultaneous equations, and as shorthand in matrix notation:

$$\begin{aligned} 2p + e + r &= 220 \\ 4p + 2e + r &= 345 \\ 4p + e + 2r &= 385 \end{aligned} \quad \text{or} \quad \begin{pmatrix} 2 & 1 & 1 \\ 4 & 2 & 1 \\ 4 & 1 & 1 \end{pmatrix} \begin{bmatrix} p \\ e \\ r \end{bmatrix} = \begin{bmatrix} 220 \\ 345 \\ 385 \end{bmatrix}$$

It is possible to use a matrix for this problem because the problem takes linear combinations of the costs of 3 items. Matrix multiplication is defined as the process of linear combinations, which is the same process as linear transformations. However, the above matrix is *not* a tensor, because there are no vectors of school supplies, no bases, and no linear combinations of (say) part eraser and part pencil. Therefore, the matrix has no well-defined transformation properties. Hence, it is a lowly matrix, but no tensor.

However, later (in “We Don’t Need No Stinking Metric”) we’ll see that under the right conditions, we *can* form a vector space out of seemingly unrelated quantities.

### Heading In the Right Direction

An ordinary vector associates a *number* with each direction of space:

$$\mathbf{v} = v_x \hat{\mathbf{x}} + v_y \hat{\mathbf{y}} + v_z \hat{\mathbf{z}}$$

The vector  $\mathbf{v}$  associates the number  $v_x$  with the  $x$ -direction; it associates the number  $v_y$  with the  $y$ -direction, and the number  $v_z$  with the  $z$ -direction.

The above tensor examples illustrate the basic nature of a rank-2 tensor:

A rank-2 tensor associates a *vector* with each direction of space:

$$\mathbf{T} = \begin{bmatrix} T_{xx} \\ T_{yx} \\ T_{zx} \end{bmatrix} \hat{\mathbf{x}} + \begin{bmatrix} T_{xy} \\ T_{yy} \\ T_{zy} \end{bmatrix} \hat{\mathbf{y}} + \begin{bmatrix} T_{xz} \\ T_{yz} \\ T_{zz} \end{bmatrix} \hat{\mathbf{z}}$$

### Some Definitions and Review

These definitions will make more sense as we go along. Don’t get stuck on these:

“ordinary” vector = contravariant vector = contravector =  $(^1_0)$  tensor

1-form = covariant vector = covector =  $(^0_1)$  tensor. (Yes, there are 4 different ways to say the same thing.)

covariant the same. E.g., General Relativity says that the mathematical form of the laws of physics are covariant (i.e., the same) with respect to arbitrary coordinate transformations. This is a *completely different meaning* of “covariant” than the one above.

rank The number of indexes of a tensor;  $T^{ij}$  is a rank-2 tensor;  $R^i_{jkl}$  is a rank-4 tensor. Rank is unrelated to the dimension of the vector space in which the tensor operates.

MVE mathematical vector element. Think of it as a vector for now.

Caution: a rank  $\binom{0}{1}$  tensor is a 1-form, but a rank  $\binom{0}{2}$  tensor is not always a 2-form. [Don't worry about it, but just for completeness, a 2-form (or any  $n$ -form) has to be fully anti-symmetric in all pairs of vector arguments.]

**Notation:**

$(a, b, c)$  is a row vector;  $(a, b, c)^T$  is a column vector (the transpose of a row vector).

To satisfy our pathetic word processor, we write  $\binom{m}{n}$ , even though the 'm' is supposed to be directly above the 'n'.

**T** is a tensor, without reference to any basis or representation.

$T^{ij}$  is the matrix of components of **T**, contravariant in both indexes, with an understood basis.

**T(v, w)** is the result of **T** acting on **v** and **w**.

**v** or  $\vec{v}$  are two equivalent ways to denote a vector, without reference to any basis or representation. Note that a vector is a rank-1 tensor.

$\tilde{\mathbf{a}}$  or  $a \sim$  are two equivalent ways to denote a covariant vector (aka 1-form), without reference to any basis or representation

$a_i$  the components of the covector (1-form) **a**, in an understood basis.

**Vector Space Summary**

Briefly, a **vector space** comprises a **field** of scalars, a **group** of vectors, and the operation of scalar multiplication of vectors (details below). Quantum mechanical vector spaces have two additional characteristics: they define a dot product between two vectors, and they define linear operators which act on vectors to produce other vectors.

Before understanding tensors, it is very helpful, if not downright necessary, to understand vector spaces. *Funky Quantum Concepts* has a more complete description of vector spaces. Here is a *very* brief summary: a **vector space** comprises a **field** of scalars, a **group** of vectors, and the operation of scalar multiplication of vectors. The scalars can be any mathematical "field," but are usually the real numbers, or the complex numbers (e.g., quantum mechanics). For a given vector space, the vectors are a class of things, which can be one of many possibilities (physical vectors, matrices, kets, bras, tensors, ...). In particular, the vectors are *not* necessarily lists of scalars, nor need they have anything to do with physical space. Vector spaces have the following properties, which allow solving simultaneous linear equations both for unknown scalars, and unknown vectors:

Scalars	Mathematical Vectors
Scalars form a commutative group (closure, unique identity, inverses) under operation +.	Vectors form a commutative group (closure, unique identity, inverses) under operation +.
Scalars, excluding 0, form a commutative group under operation ( · ).	
Distributive property of ( · ) over +.	
Scalar multiplication of vector produces another vector.	
Distributive property of scalar multiplication over both scalar + and vector +.	

With just the scalars, you can solve ordinary scalar linear equations such as:

$$\left. \begin{aligned} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n &= c_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n &= c_2 \\ \vdots & \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n &= c_n \end{aligned} \right\} \text{ written in matrix form as } \mathbf{ax} = \mathbf{c}$$

All the usual methods of linear algebra work to solve the above equations: Cramer’s rule, Gaussian elimination, etc. With the whole vector space, you can solve simultaneous linear *vector* equations for unknown vectors, such as

$$\left. \begin{aligned} a_{11}\mathbf{v}_1 + a_{12}\mathbf{v}_2 + \dots + a_{1n}\mathbf{v}_n &= \mathbf{w}_1 \\ a_{21}\mathbf{v}_1 + a_{22}\mathbf{v}_2 + \dots + a_{2n}\mathbf{v}_n &= \mathbf{w}_2 \\ \vdots & \\ a_{n1}\mathbf{v}_1 + a_{n2}\mathbf{v}_2 + \dots + a_{nn}\mathbf{v}_n &= \mathbf{w}_n \end{aligned} \right\} \text{ written in matrix form as } \mathbf{av} = \mathbf{w}$$

where **a** is again a matrix of scalars. The same methods of linear algebra work just as well to solve vector equations as scalar equations.

Vector spaces may also have these properties:

Dot product produces a scalar from two vectors.
Linear operators act on vectors to produce other vectors.

The key points of mathematical vectors are (1) we can form linear combinations of them to make other vectors, and (2) any vector can be written as a linear combination of basis vectors:

$$\mathbf{v} = (v^1, v^2, v^3) = v^1\mathbf{e}_1 + v^2\mathbf{e}_2 + v^3\mathbf{e}_3 \quad \text{where } \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \text{ are basis vectors, and } v^1, v^2, v^3 \text{ are the components of } \mathbf{v} \text{ in the } \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3 \text{ basis.}$$

Note that  $v^1, v^2, v^3$  are *numbers*, while  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  are *vectors*. There is a (kind of bogus) reason why basis vectors are written with subscripts, and vector components with superscripts, but we’ll get to that later.

The **dimension** of a vector space,  $N$ , is the number of basis vectors needed to construct every vector in the space.

Do not confuse the dimension of physical space (typically 1D, 2D, 3D, or (in relativity) 4D), with the dimension of the mathematical objects used to work a problem.

For example, a 3×3 matrix is an element of the vector space of 3×3 matrices. This is a 9-D vector space, because there are 9 basis matrices needed to construct an arbitrary matrix.

Given a basis, components are equivalent to the vector. Components alone (without a basis) are insufficient to be a vector.

[Aside: Note that for position vectors defined by  $\mathbf{r} = (r, \theta, \phi)$ ,  $r, \theta$ , and  $\phi$  are *not* the components of a vector. The tip off is that with two vectors, you can *always* add their components to get another vector. Clearly,  $\mathbf{r}_1 + \mathbf{r}_2 \neq (r_1 + r_2, \theta_1 + \theta_2, \phi_1 + \phi_2)$ , so  $(r, \theta, \phi)$  *cannot be* the components of a vector. This failure to add is due to  $\mathbf{r}$  being a displacement vector from the origin, where there is no consistent basis: e.g., what is  $\mathbf{e}_r$  at the origin? At points *off* the origin, there *is* a consistent basis:  $\mathbf{e}_r, \mathbf{e}_\theta$ , and  $\mathbf{e}_\phi$  are well-defined.]

### When Vectors Collide

There now arises a collision of terminology: to a physicist, “vector” usually means a physical vector in 3- or 4-space, but to a mathematician, “vector” means an element of a mathematical vector-space. These are two different meanings, but they share a common aspect: linearity (i.e., we can form linear



combinations of vectors to make other vectors, and any vector can be written as a linear combination of basis vectors). Because of that linearity, we can have general rank- $n$  tensors whose components are arbitrary elements of a mathematical vector-space. To make the terminology confusion worse, an  $\binom{m}{n}$  tensor whose components are simple numbers is itself a “vector-element” of the vector-space of  $\binom{m}{n}$  tensors.

Mathematical vector-elements of a vector space are much more general than physical vectors (e.g. force, or velocity), though physical vectors and tensors are elements of mathematical vector spaces. To be clear, we’ll use **MVE** to refer to a mathematical vector-element of a vector space, and “vector” to mean a normal physics vector (3-vector or 4-vector). Recall that MVEs are usually written as a set of components in some basis, just like vectors are. In the beginning, we choose all the input MVEs to be vectors.

If you’re unclear about what an MVE is, just think of it as a physical vector for now, like “force.”

### “Tensors” vs. “Symbols”

There are lots of tensors: metric tensors, electromagnetic tensors, Riemann tensors, etc. There are also “symbols:” Levi-Civita symbols, Christoffel symbols, etc. What’s the difference? “Symbols” aren’t tensors. Symbols look like tensors, in that they have components indexed by multiple indices, they are referred to basis vectors, and are summed with tensors. But they are defined to have *specific* components, which may depend on the basis, and therefore symbols don’t change basis (transform) the way tensors do. Hence, symbols are *not* geometric entities, with a meaning in a manifold, independent of coordinates. For example, the Levi-Civita symbol is defined to have specific constant components in *all* bases. It doesn’t follow the usual change-of-basis rules. Therefore, it cannot be a tensor.

### Notational Nightmare

If you come from a differential geometry background, you may wonder about some insanely confusing notation. It is a fact that “ $dx$ ” and “ $\mathbf{d}x$ ” are two different things:

$$dx = (dx, dy, dz) \quad \text{is a vector, but}$$

$$\mathbf{d}x = \nabla x(\mathbf{r}) \quad \text{is a 1-form}$$

We don’t use the second notation (or exterior derivatives) in this chapter, but we might in the Differential Geometry chapter.

## Tensors? What Good Are They?

### A Short, Complicated Definition

It is very difficult to give a short definition of a tensor that is useful to anyone who doesn’t already know what a tensor is. Nonetheless, you’ve got to start somewhere, so we’ll give a short definition, to point in the right direction, but it may not make complete sense at first (don’t get hung up on this, skip if needed):

A **tensor** is an operator on one or more **mathematical vector elements** (MVEs), *linear in each operand*, which produces another mathematical vector element.

The key point is this (which we describe in more detail in a moment):

Linearity in all the operands is the essence of a tensor.

I should add that the basis vectors for all the MVEs must be the same (or tensor products of the same) for an operator to qualify as a tensor. But that’s too much to put in a “short” definition. We clarify this point later.

Note that a **scalar** (i.e., a coordinate-system-invariant number, but for now, just a number) satisfies the definition of a “mathematical vector element.”

Many definitions of tensors dwell on the transformation properties of tensors. This is mathematically valid, but such definitions give no insight into the use of tensors, or why we like them. Note that to satisfy

the transformation properties, all the input vectors and output tensors must be expressed in the same basis (or tensor products of that basis with itself).

Some coordinate systems require distinguishing between **contravariant** and **covariant** components of tensors; superscripts denote contravariant components; subscripts denote covariant components. However, orthonormal positive definite systems, such as the familiar Cartesian, spherical, and cylindrical systems, do not require such a distinction. So for now, let's ignore the distinction, even though the following notation properly represents both contravariant and covariant components. Thus, in the following text, contravariant components are written with superscripts, and covariant components are written with subscripts, but we don't care right now. Just think of them all as components in an arbitrary coordinate system.

## Building a Tensor

Oversimplified, a tensor operates on vectors to produce a scalar or a vector. Let's construct a tensor which accepts (operates on) two 3-vectors to produce a scalar. (We'll see later that this is a rank-2 tensor.) Let the tensor **T** act on vectors **a** and **b** to produce a scalar, *s*; in other words, this tensor is a scalar function of two vectors:

$$s = \mathbf{T}(\mathbf{a}, \mathbf{b})$$

Call the first vector  $\mathbf{a} = (a^1, a^2, a^3)$  in some basis, and the second vector  $\mathbf{b} = (b^1, b^2, b^3)$  (in the same basis). A tensor, by definition, must be linear in both **a** and **b**; if we double **a**, we double the result, if we triple **b**, we triple the result, etc. Also,

$$\mathbf{T}(\mathbf{a} + \mathbf{c}, \mathbf{b}) = \mathbf{T}(\mathbf{a}, \mathbf{b}) + \mathbf{T}(\mathbf{c}, \mathbf{b}), \quad \text{and} \quad \mathbf{T}(\mathbf{a}, \mathbf{b} + \mathbf{d}) =$$

So the result *must* involve at least the product of a component of **a** with a component of **b**. Let's say the tensor takes  $a^2 b^1$  as that product, and additionally multiplies it by a constant,  $T_{21}$ . Then we have built a tensor acting on **a** and **b**, and it is linear in both:

$$\mathbf{T}(\mathbf{a}, \mathbf{b}) = T_{21} a^2 b^1. \quad \text{Example:} \quad \mathbf{T}(\mathbf{a}, \mathbf{b}) = 7 a^2 b^1$$

But, if we add to this some *other* weighted product of some *other* pair of components, the result is still a tensor: it is still linear in both **a** and **b**:

$$\mathbf{T}(\mathbf{a}, \mathbf{b}) = T_{13} a^1 b^3 + T_{21} a^2 b^1. \quad \text{Example:} \quad \mathbf{T}(\mathbf{a}, \mathbf{b}) = 4 a^1 b^3 + 7 a^2 b^1$$

In fact, we can extend this to the weighted sum of *all* combinations of components, one each from **a** and **b**. Such a sum is still linear in both **a** and **b**:

$$\mathbf{T}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^3 \sum_{j=1}^3 T_{ij} a^i b^j \quad \text{Example:} \quad T_{ij} = \begin{bmatrix} -2 & 6 & 4 \\ 7 & 5 & -1 \\ -6 & 0 & 8 \end{bmatrix}$$

Further, nothing else can be added to this that is linear in **a** and **b**.

A tensor is the most general linear function of **a** and **b** that exists, i.e. any linear function of **a** and **b** can be written as a 3x3 matrix.

(We'll see that the **rank** of a tensor is equal to the number of its indices; **T** is a rank-2 tensor.) The  $T_{ij}$  are the **components** of the tensor (in the basis of the vectors **a** and **b**.) At this point, we consider the components of **T**, **a**, and **b** all as just numbers.

Why does a tensor have a separate weight for each *combination* of components, one from each input mathematical vector element (MVE)? Couldn't we just weight each input MVE as a whole? No, because that would restrict tensors to only *some* linear functions of the inputs.

Any linear function of the input vectors can be represented as a tensor.

Note that tensors, just like vectors, can be written as components in some basis. And just like vectors, we can transform the components from one basis to another. Such a transformation does *not* change the

tensor itself (nor does it change a vector); it simply changes how we represent the tensor (or vector). More on transformations later.

Tensors don't have to produce scalar results!

Some tensors accept one or more vectors, and produce a vector for a result. Or they produce some rank- $r$  tensor for a result. In general, a rank- $n$  tensor accepts ' $m$ ' vectors as inputs, and produces a rank ' $n-m$ ' tensor as a result. Since any tensor is an element of a mathematical vector space, tensors can be written as linear combinations of other (same rank & type) tensors. So even when a tensor produces another (lower rank) tensor as an output, the tensor is still a linear function of all its input vectors. It's just a tensor-valued function, instead of a scalar-valued function. For example, the force on a charge: a B-field operates on a vector,  $q\mathbf{v}$ , to produce a vector,  $\mathbf{f}$ . Thus, we can think of the B-field as a rank-2 tensor which acts on a vector to produce a vector; it's a vector-valued function of one vector.

Also, in general, tensors aren't limited to taking just vectors as inputs. Some tensors take rank-2 tensors as inputs. For example, the quadrupole moment tensor operates on the 2<sup>nd</sup> derivative matrix of the potential (the rank-2 "Hessian" tensor) to produce the (scalar) work stored in the quadrupole of charges. And a density matrix in quantum mechanics is a rank-2 tensor that acts on an operator matrix (rank-2 tensor) to produce the ensemble average of that operator.

---

## Tensors in Action

Let's consider how rank-0, rank-1, and rank-2 tensors operate on a single vector. Recall that in "tensor-talk," a scalar is an invariant number, i.e. it is the same number in any coordinate system.

**Rank-0:** A rank-0 tensor is a scalar, i.e. a coordinate-system-independent number. Multiplying a vector by a rank-0 tensor (a scalar), produces a new vector. Each component of the vector contributes to the corresponding component of the result, and each component is weighted *equally* by the scalar,  $a$ :

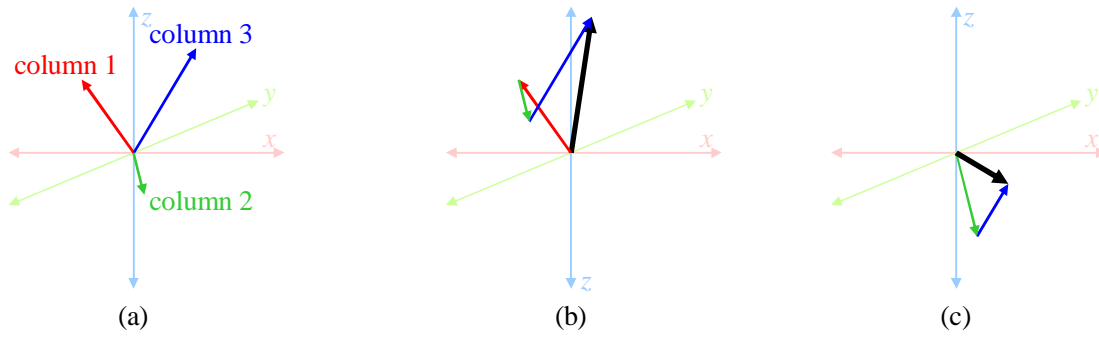
$$\vec{v} = v^x \mathbf{i} + v^y \mathbf{j} + v^z \mathbf{k} \quad \Rightarrow \quad a\vec{v} = av^x \mathbf{i} + av^y \mathbf{j} + av^z \mathbf{k}$$

**Rank-1:** A rank-1 tensor  $\mathbf{a}$  operates on (contracts with) a vector to produce a scalar. Each component of the input vector contributes a *number* to the result, but each component is weighted *separately* by the corresponding component of the tensor  $\mathbf{a}$ :

$$\tilde{\mathbf{a}}(\mathbf{v}) = a_x v^x + a_y v^y + a_z v^z = \sum_{i=1}^3 a_i v^i$$

Note that a vector is itself a rank-1 tensor. Above, instead of considering  $\mathbf{a}$  acting on  $\mathbf{v}$ , we can equivalently consider that  $\mathbf{v}$  acts on  $\mathbf{a}$ :  $\mathbf{a}(\mathbf{v}) = \mathbf{v}(\mathbf{a})$ . Both  $\mathbf{a}$  and  $\mathbf{v}$  are of equal standing.

**Rank-2:** Filling one slot of a rank-2 tensor with a vector produces a new vector. Each component of the input vector contributes a *vector* to the result, and each input vector component weights a *different* vector.



(a) A hypothetical rank-2 tensor with an  $x$ -vector (red), a  $y$ -vector (green), and a  $z$ -vector (blue). (b) The tensor acting on the vector  $(1, 1, 1)$  producing a vector (heavy black). Each component (column) vector of the tensor is weighted by 1, and summed. (c) The tensor acting on the vector  $(0, 2, 0.5)$ , producing a vector (heavy black). The  $x$ -vector is weighted by 0, and so does not contribute; the  $y$ -vector is weighted by 2, so contributes double; the  $z$ -vector is weighted by 0.5, so contributes half.

$$\mathbf{B}(\_, \mathbf{v}) = B^i_j v^j = \begin{bmatrix} B^x_x & B^x_y & B^x_z \\ B^y_x & B^y_y & B^y_z \\ B^z_x & B^z_y & B^z_z \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} = v^x \begin{bmatrix} B^x_x \\ B^y_x \\ B^z_x \end{bmatrix} + v^y \begin{bmatrix} B^x_y \\ B^y_y \\ B^z_y \end{bmatrix} + v^z \begin{bmatrix} B^x_z \\ B^y_z \\ B^z_z \end{bmatrix}$$

$$= \mathbf{B}_x v^x + \mathbf{B}_y v^y + \mathbf{B}_z v^z = \left( \sum_{j=1}^3 B^x_j v^j \right) \mathbf{i} + \left( \sum_{j=1}^3 B^y_j v^j \right) \mathbf{j} + \left( \sum_{j=1}^3 B^z_j v^j \right) \mathbf{k}$$

The columns of  $\mathbf{B}$  are the vectors which are weighted by each of the input vector components,  $v^j$ ; or equivalently, the columns of  $\mathbf{B}$  are the vector weights for each of the input vector components

**Example of a simple rank-2 tensor: the moment-of-inertia tensor,  $\mathbf{I}_{ij}$ .** Every blob of matter has one. We know from mechanics that if you rotate an arbitrary blob around an arbitrary axis, the angular momentum vector of the blob does *not* in general line up with the axis of rotation. So what is the angular momentum vector of the blob? It is a vector-valued linear function of the angular velocity vector, i.e. given the angular velocity vector, you can operate on it with the moment-of-inertia tensor, to get the angular momentum vector. Therefore, by the definition of a tensor as a linear operation on a vector, the relationship between angular momentum vector and angular velocity vector can be given as a tensor; it is the moment-of-inertia tensor. It takes as an input the angular velocity vector, and produces as output the angular momentum vector, therefore it is a rank-2 tensor:

$$\mathbf{I}(\boldsymbol{\omega}, \_) = \mathbf{L}, \quad \mathbf{I}(\boldsymbol{\omega}, \boldsymbol{\omega}) = \mathbf{L} \cdot \boldsymbol{\omega} = 2KE$$

[Since  $\mathbf{I}$  is constant in the blob frame, it rotates in the lab frame. Thus, in the lab frame, the above equations are valid only at a single instant in time. In effect,  $\mathbf{I}$  is a function of time,  $\mathbf{I}(t)$ .]

[?? This may be a bad example, since  $\mathbf{I}$  is only a **Cartesian tensor** [L&L3, p ??], which is not a real tensor. Real tensors can't have finite displacements on a curved manifold, but blobs of matter have finite size. If you want to get the kinetic energy, you have to use the metric to compute  $\mathbf{L} \cdot \boldsymbol{\omega}$ . Is there a simple example of a real rank-2 tensor??]

Note that some rank-2 tensors operate on *two* vectors to produce a scalar, and some (like  $\mathbf{I}$ ) can either act on one vector to produce a vector, or act on two vectors to produce a scalar (twice the kinetic energy). More of that, and higher rank tensors, later.

## Tensor Fields

A vector is a single mathematical object, but it is quite common to define a *field* of vectors. A **field** in this sense is a function of space. A **vector field** defines a vector for each point in a space. For example, the electric field is a vector-valued function of space: at each point in space, there is an electric field vector.

Similarly, a tensor is a single mathematical object, but it is quite common to define a *field* of tensors. At each point in space, there is a tensor. The metric tensor field is a tensor-valued function of space: at each point, there is a metric tensor. Almost universally, the word “field” is omitted when calling out tensor fields: when you say “metric tensor,” everyone is expected to know it is a tensor *field*. When you say “moment of inertia tensor,” everyone is expected to know it is a single tensor (not a field).

## Dot Products and Cross Products as Tensors

Symmetric tensors are associated with elementary dot products, and anti-symmetric tensors are associated with elementary cross-products.

A dot product is a linear operation on two vectors:  $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$ , which produces a scalar. Because the dot product is a linear function of two vectors, it can be written as a tensor. (Recall that *any* linear function of vectors can be written as a tensor.) Since it takes two rank-1 tensors, and produces a rank-0 tensor, the dot product is a rank-2 tensor. Therefore, we can achieve the same result as a dot product with a rank-2 symmetric tensor that accepts two vectors and produces a scalar; call this tensor  $\mathbf{g}$ :

$$\mathbf{g}(\mathbf{A}, \mathbf{B}) = \mathbf{g}(\mathbf{B}, \mathbf{A})$$

‘ $\mathbf{g}$ ’ is called the **metric tensor**: it produces the dot product (aka scalar product) of two vectors. Quite often, the metric tensor varies as a function of the generalized coordinates of the system; then it is a metric tensor field. It happens that the dot product is symmetric:  $\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A}$ ; therefore,  $\mathbf{g}$  is symmetric. If we write the components of  $\mathbf{g}$  as a matrix, the matrix will be symmetric, i.e. it will equal its own transpose. (Do I need to expand on this??)

On the other hand, a cross product is an anti-symmetric linear operation on two vectors, which produces another vector:  $\mathbf{A} \times \mathbf{B} = -\mathbf{B} \times \mathbf{A}$ . Therefore, we can associate one vector, say  $\mathbf{B}$ , with a rank-2 anti-symmetric tensor, that accepts one vector and produces another vector:

$$\mathbf{B}(\_, \mathbf{A}) = -\mathbf{B}(\mathbf{A}, \_)$$

For example, the Lorentz force law:  $\mathbf{F} = \mathbf{v} \times \mathbf{B}$ . We can write  $\mathbf{B}$  as a  $\binom{1}{1}$  tensor:

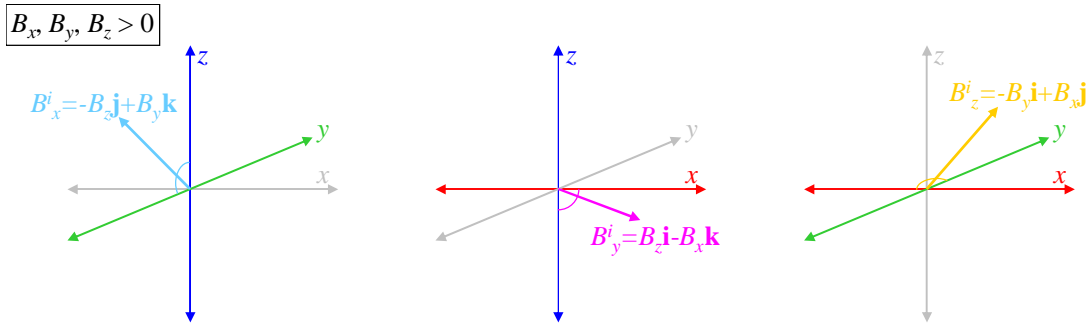
$$\mathbf{F} = \mathbf{v} \times \mathbf{B} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ v^x & v^y & v^z \\ B_x & B_y & B_z \end{vmatrix} = \mathbf{B}(\_, \mathbf{v}) = B^i_j v^j = \begin{bmatrix} 0 & B_z & -B_y \\ -B_z & 0 & B_x \\ B_y & -B_x & 0 \end{bmatrix} \begin{bmatrix} v^x \\ v^y \\ v^z \end{bmatrix} = \begin{bmatrix} B_z v^y - B_y v^z \\ -B_z v^x + B_x v^z \\ B_y v^x - B_x v^y \end{bmatrix}$$

We see again how a rank-2 tensor,  $\mathbf{B}$ , contributes a *vector* for each *component* of  $\mathbf{v}$ :

$$B^i_x \mathbf{e}_i = -B_z \mathbf{j} + B_y \mathbf{k} \text{ (the first column of } B \text{) is weighted by } v^x.$$

$$B^i_y \mathbf{e}_i = B_z \mathbf{i} - B_x \mathbf{k} \text{ (the 2}^{\text{nd}} \text{ column of } B \text{) is weighted by } v^y.$$

$$B^i_z \mathbf{e}_i = -B_y \mathbf{i} + B_x \mathbf{j} \text{ (the 3}^{\text{rd}} \text{ column of } B \text{) is weighted by } v^z.$$



A rank-2 tensor acting on a vector to produce their cross-product.

TBS: We can also think of the cross product as a fully anti-symmetric rank-3 tensor, which acts on 2 vectors to produce a vector (their cross product). This is the **anti-symmetric symbol**  $\epsilon_{ijk}$  (not a tensor).

Note that both the dot product and cross-product are linear on both of their operands. For example:

$$(\alpha \mathbf{A} + \gamma \mathbf{C}) \cdot \mathbf{B} = \alpha (\mathbf{A} \cdot \mathbf{B}) + \gamma (\mathbf{C} \cdot \mathbf{B})$$

$$\mathbf{A} \cdot (\beta \mathbf{B} + \eta \mathbf{D}) = \beta (\mathbf{A} \cdot \mathbf{B}) + \eta (\mathbf{A} \cdot \mathbf{D})$$

Linearity in all the operands is the essence of a tensor.

Note also that a “rank” of a tensor **contracts** with (is summed over) a “rank” of one of its operands to eliminate both of them: one rank of the B-field tensor contracts with one input vector, leaving one surviving rank of the B-field tensor, which is the vector result. Similarly, one rank of the metric tensor,  $\mathbf{g}$ , contracts with the first operand vector; another rank of  $\mathbf{g}$  contracts with the second operand vector, leaving a rank-0 (scalar) result.

**The Danger of Matrices**

There are some dangers to thinking of tensors as matrices: (1) it doesn’t work for rank 3 or higher tensors, and (2) non-commutation of matrix multiplication is harder to follow than the more-explicit summation convention. Nonetheless, the matrix conventions are these:

- contravariant components and basis covectors (“up” indexes) → column vector. E.g.,

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ v_3 \end{bmatrix}, \quad \text{basis 1-forms: } \begin{bmatrix} \mathbf{e}^1 \\ \mathbf{e}^2 \\ \mathbf{e}^3 \end{bmatrix}$$

- covariant components and basis contravectors (“down” indexes) → row vector

$$\mathbf{w} = (w_1, w_2, w_3), \quad \text{basis vectors: } (\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$$

Matrix rows and columns are indicated by spacing of the indexes, and are independent of their “upness” or “downness.” The first matrix index is always the row; the second, the column:

$$T^r_c \quad T_r^c \quad T_{rc} \quad T^{rc} \quad \text{where } r = \text{row index, } c = \text{column index}$$

**Reading Tensor Component Equations**

Tensor equations can be written as equations with tensors as operators (written in bold):

$$\text{KE} = \frac{1}{2} \mathbf{I}(\boldsymbol{\omega}, \boldsymbol{\omega})$$

Or, they can be written in component form:

$$(1) \text{ KE} = \frac{1}{2} I_{ij} \omega^i \omega^j$$

We'll be using lots of tensor equations written in component form, so it is important to know how to read them. Note that some standard notations almost *require* component form: In GR, the Ricci tensor is  $R^{\mu\nu}$ , and the Ricci scalar is  $R$ :

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}$$

In component equations, tensor indexes are written explicitly. There are two kinds of tensor indexes: dummy (aka summation) indexes, and free indexes. Dummy indexes appear exactly twice in any term. Free indexes appear only once in each term, and the same free indexes must appear in each term (except for scalar terms). In the above equation, both  $\mu$  and  $\nu$  are free indexes, and there are no dummy indexes. In eq. (1) above,  $i$  and  $j$  are both dummy indexes and there are no free indexes.

Dummy indexes appear exactly twice in any term are used for implied summation, e.g.

$$KE = \frac{1}{2} I_{ij} \omega^i \omega^j \quad \equiv \quad KE = \frac{1}{2} \sum_{i=1}^3 \sum_{j=1}^3 I_{ij} \omega^i \omega^j$$

Free indexes are a shorthand for writing several equations at once. Each free index takes on all possible values for it. Thus,

$$C^i = A^i + B^i \quad \equiv \quad C^x = A^x + B^x, \quad C^y = A^y + B^y, \quad C^z = A^z + B^z \quad (3 \text{ equations})$$

and

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} \quad \equiv$$

$G_{00} = R_{00} - \frac{1}{2} R g_{00},$	$G_{01} = R_{01} - \frac{1}{2} R g_{01},$	$G_{02} = R_{02} - \frac{1}{2} R g_{02},$	$G_{03} = R_{03} - \frac{1}{2} R g_{03}$
$G_{10} = R_{10} - \frac{1}{2} R g_{10},$	$G_{11} = R_{11} - \frac{1}{2} R g_{11},$	$G_{12} = R_{12} - \frac{1}{2} R g_{12},$	$G_{13} = R_{13} - \frac{1}{2} R g_{13}$
$G_{20} = R_{20} - \frac{1}{2} R g_{20},$	$G_{21} = R_{21} - \frac{1}{2} R g_{21},$	$G_{22} = R_{22} - \frac{1}{2} R g_{22},$	$G_{23} = R_{23} - \frac{1}{2} R g_{23}$
$G_{30} = R_{30} - \frac{1}{2} R g_{30},$	$G_{31} = R_{31} - \frac{1}{2} R g_{31},$	$G_{32} = R_{32} - \frac{1}{2} R g_{32},$	$G_{33} = R_{33} - \frac{1}{2} R g_{33}$

(16 equations).

It is common to have both dummy and free indexes in the same equation. Thus the GR statement of conservation of energy and momentum uses  $\mu$  as a dummy index, and  $\nu$  as a free index:

$$\nabla_{\mu} T^{\mu\nu} = 0 \quad \equiv \quad \sum_{\mu=0}^3 \nabla_{\mu} T^{\mu 0} = 0, \quad \sum_{\mu=0}^3 \nabla_{\mu} T^{\mu 1} = 0, \quad \sum_{\mu=0}^3 \nabla_{\mu} T^{\mu 2} = 0, \quad \sum_{\mu=0}^3 \nabla_{\mu} T^{\mu 3} = 0$$

(4 equations). Notice that scalars apply to all values of free indexes, and don't need indexes of their own. However, any free indexes must match on all tensor terms. It is nonsense to write something like:

$$A^{ij} = B^i + C^j \quad (\text{nonsense})$$

However, it is reasonable to have

$$A^{ij} = B^i C^j \quad \text{E.g., angular momentum: } M^{ij} = r^i p^j - r^j p^i$$

### Adding, Subtracting, Differentiating Tensors

Since tensors are linear operations, you can add or subtract any two tensors that take the same type arguments and produce the same type result. Just add the tensor components individually.

$$S = T + U \quad \text{E.g.} \quad S^{ij} = T^{ij} + U^{ij}, \quad i, j = 1, \dots, N$$

You can also scalar multiply a tensor. Since these properties of tensors are the defining requirements for a vector space, all the tensors of given rank and index types compose a vector space, and every tensor is an MVE in its space.

This implies that a tensor field can be differentiated (or integrated), and in particular, it has a gradient.

### Higher Rank Tensors

When considering higher rank tensors, it may be helpful to recall that multi-dimensional matrices can be thought of as lower-dimensional matrices with each element itself a vector or matrix. For example, a 3 x 3 matrix can be thought of as a “column vector” of 3 row-vectors. Matrix multiplication works out the same whether you consider the 3 x 3 matrix as a 2-D matrix of numbers, or a 1-D column vector of row vectors:

$$(x \ y \ z) \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = (ax + dy + gz \quad bx + ey + hz \quad cx + fy + iz)$$

or



$$(x \ y \ z) \begin{bmatrix} (a,b,c) \\ (d,e,f) \\ (g,h,i) \end{bmatrix} = x(a,b,c) + y(d,e,f) + z(g,h,i) = (ax + dy + gz \quad bx + ey + hz \quad cx + fy + iz)$$

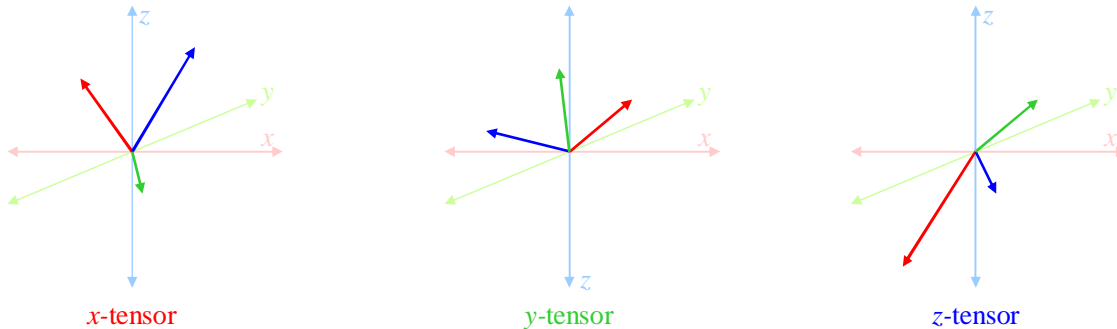
Using this same idea, we can compare the gradient of a scalar field, which is a  $(^0_1)$  tensor field (a 1-form), with the gradient of a rank-2 (say  $(^0_2)$ ) tensor field, which is a  $(^0_3)$  tensor field. First, the gradient of a scalar field is a  $(^0_1)$  tensor field with 3 components, where each component is a number-valued function:

$$\nabla f = \mathbf{D} = \frac{\partial f}{\partial x} \boldsymbol{\omega}^1 + \frac{\partial f}{\partial y} \boldsymbol{\omega}^2 + \frac{\partial f}{\partial z} \boldsymbol{\omega}^3, \quad \boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \boldsymbol{\omega}_3 \text{ are basis (co)vectors}$$

$$\mathbf{D} \text{ can be written as } (D_1, D_2, D_3), \text{ where } D_1 = \frac{\partial f}{\partial x}, \quad D_2 = \frac{\partial f}{\partial y}, \quad D_3 = \frac{\partial f}{\partial z}$$

The gradient operates on an infinitesimal displacement vector to produce the change in the function when you move through the given displacement:  $df = \mathbf{D}(d\mathbf{r}) = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz$ .

Now let  $\mathbf{R}$  be a  $(^0_2)$  tensor field, and  $\mathbf{T}$  be its gradient.  $\mathbf{T}$  is a  $(^0_3)$  tensor field, but can be thought of as a  $(^0_1)$  tensor field where each component is itself a  $(^0_2)$  tensor.



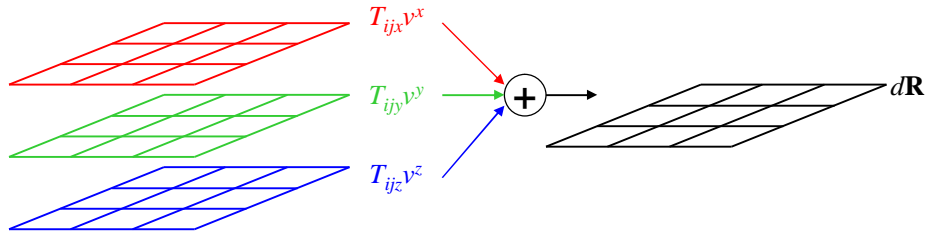
A rank-3 tensor considered as a set of 3 rank-2 tensors: an x-tensor, a y-tensor, and a z-tensor.

The gradient operates on an infinitesimal displacement vector to produce the change in the  $(^0_2)$  tensor field when you move through the given displacement.



$$\mathbf{T} = \nabla \mathbf{R} = \frac{\partial \mathbf{R}}{\partial x} \boldsymbol{\omega}^1 + \frac{\partial \mathbf{R}}{\partial y} \boldsymbol{\omega}^2 + \frac{\partial \mathbf{R}}{\partial z} \boldsymbol{\omega}^3$$

$$= \begin{pmatrix} \begin{bmatrix} T_{11x} & T_{12x} & T_{13x} \\ T_{21x} & T_{22x} & T_{23x} \\ T_{31x} & T_{32x} & T_{33x} \end{bmatrix} & \begin{bmatrix} T_{11y} & T_{12y} & T_{13y} \\ T_{21y} & T_{22y} & T_{23y} \\ T_{31y} & T_{32y} & T_{33y} \end{bmatrix} & \begin{bmatrix} T_{11z} & T_{12z} & T_{13z} \\ T_{21z} & T_{22z} & T_{23z} \\ T_{31z} & T_{32z} & T_{33z} \end{bmatrix} \end{pmatrix}$$



$$d\mathbf{R} = \mathbf{T}(\mathbf{v}) = \sum_{k=x,y,z} T_{ijk} v^k \quad \leftrightarrow \quad (dR)_{ij} = T_{ijk} v^k$$

Note that if  $\mathbf{R}$  had been a  $(^2_0)$  (fully *contravariant*) tensor, then its gradient would be a  $(^2_1)$  mixed tensor. Taking the gradient of any field simply adds a covariant index, which can then be contracted with a displacement vector to find the change in the tensor field when moving through the given displacement.

The contraction considerations of the previous section still apply: a rank of an tensor operator contracts with a rank of one of its inputs to eliminate both. In other words, each rank of input tensors eliminates one rank of the tensor operator. The rank of the result is the number of surviving ranks from the tensor operator:

$$rank(tensor) = \left( \sum rank(inputs) \right) + rank(result)$$

or 
$$rank(result) = rank(tensor) - \left( \sum rank(inputs) \right)$$

**Tensors of Mathematical Vector Elements:** The operation of a tensor on vectors involves multiplying components (one from the tensor, and one from each input vector), and then summing. E.g.,

$$\mathbf{T}(\mathbf{a}, \mathbf{b}) = T_{11} a^1 b^1 + \dots + T_{ij} a^i b^j + \dots$$

Similar to the above example, the  $T_{ij}$  components could themselves be a vector of a mathematical vector space (i.e., could be MVEs), while the  $a^i$  and  $b^j$  components are scalars of that vector space. In the example above, we could say that each of the  $T_{ijx}$ ,  $T_{iyy}$ , and  $T_{ijz}$  is a rank-2 tensor (an MVE in the space of rank-2 tensors), and the components of  $\mathbf{v}$  are scalars in that space (in this case, real numbers).

### Tensors In General

In complete generality then, a tensor  $\mathbf{T}$  is a linear operation on one or more MVEs:  
 $\mathbf{T}(\mathbf{a}, \mathbf{b}, \dots)$ .

Linearity implies that  $\mathbf{T}$  can be written as a numeric weight for each *combination* of components, one component from each input MVE. Thus, the “linear operation” performed by  $\mathbf{T}$  is equivalent to a weighted sum of all combinations of components of the input MVEs. (Since  $\mathbf{T}$  and the  $\mathbf{a}, \mathbf{b}, \dots$  are simple objects, not functions, there is no concept of derivative or integral operations. Derivatives and integrals are linear operations *on functions*, but not linear functions of MVEs.)

Given the components of the inputs  $\mathbf{a}, \mathbf{b}, \dots$ , and the components of  $\mathbf{T}$ , we can contract  $\mathbf{T}$  with (operate with  $\mathbf{T}$  on) the inputs to produce a MVE result. Note that all input MVEs have to have the same basis.

Also,  $\mathbf{T}$  may have units, so the output units are arbitrary. Note that in generalized coordinates, different components of a tensor may have different units (much like the vector parameters  $r$  and  $\theta$  have different units).

## Change of Basis: Transformations

Since tensors are linear operations on MVEs, we can represent a tensor by components. If we know a tensor's operations on all combinations of basis vectors, we have fully defined the tensor. Consider a rank-2 tensor  $\mathbf{T}$  acting on two vectors,  $\mathbf{a}$  and  $\mathbf{b}$ . We expand  $\mathbf{T}$ ,  $\mathbf{a}$ , and  $\mathbf{b}$  into components, using the linearity of the tensor:

$$\begin{aligned} \mathbf{T}(\mathbf{a}, \mathbf{b}) &= \mathbf{T}(a^1\mathbf{i} + a^2\mathbf{j} + a^3\mathbf{k}, b^1\mathbf{i} + b^2\mathbf{j} + b^3\mathbf{k}) \\ &= a^1b^1\mathbf{T}(\mathbf{i}, \mathbf{i}) + a^2b^1\mathbf{T}(\mathbf{j}, \mathbf{i}) + a^3b^1\mathbf{T}(\mathbf{k}, \mathbf{i}) \\ &\quad + a^1b^2\mathbf{T}(\mathbf{i}, \mathbf{j}) + a^2b^2\mathbf{T}(\mathbf{j}, \mathbf{j}) + a^3b^2\mathbf{T}(\mathbf{k}, \mathbf{j}) \\ &\quad + a^1b^3\mathbf{T}(\mathbf{i}, \mathbf{k}) + a^2b^3\mathbf{T}(\mathbf{j}, \mathbf{k}) + a^3b^3\mathbf{T}(\mathbf{k}, \mathbf{k}) \end{aligned}$$

Define  $T_{ij} = \mathbf{T}(\mathbf{e}_i, \mathbf{e}_j)$ , where  $\mathbf{e}_1 = \mathbf{i}$ ,  $\mathbf{e}_2 = \mathbf{j}$ ,  $\mathbf{e}_3 = \mathbf{k}$

then 
$$\mathbf{T}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^3 \sum_{j=1}^3 a^i b^j \mathbf{T}(\mathbf{e}_i, \mathbf{e}_j) = \sum_{i=1}^3 \sum_{j=1}^3 T_{ij} a^i b^j$$

The tensor's values on all combinations of input basis vectors are the **components** of the tensor (in the basis of the input vectors.)

Now let's transform  $\mathbf{T}$  to another basis. To change from one basis to another, we need to know how to find the new basis vectors from the old ones, or equivalently, how to transform components in the old basis to components in the new basis. We write the new basis with primes, and the old basis without primes.

Because vector spaces demand linearity, any change of basis can be written as a linear transformation of the basis vectors or components, so we can write (eq. #s from Talman):

$$\mathbf{e}'_i = \sum_{k=1}^N \Lambda^k_i \mathbf{e}_k = \Lambda^k_i \mathbf{e}_k \quad [\text{Tal 2.4.5}]$$

$$v'^i = \sum_{k=1}^N (\Lambda^{-1})^i_k v^k = (\Lambda^{-1})^i_k v^k \quad [\text{Tal 2.4.8}]$$

where the last form uses the summation convention. There is a very important difference between equations 2.4.5 and 2.4.8. The first is a set of 3 *vector* equations, expressing each of the *new* basis vectors in the *old* basis

Aside: Let's look more closely at the difference between equations 2.4.5 and 2.4.8. The first is a set of 3 *vector* equations, expressing each of the *new* basis vectors in the *old* basis. Basis vectors are vectors, and hence can themselves be expressed in any basis:

$$\left. \begin{aligned} \mathbf{e}'_1 &= \Lambda^1_1 \mathbf{e}_1 + \Lambda^2_1 \mathbf{e}_2 + \Lambda^3_1 \mathbf{e}_3 \\ \mathbf{e}'_2 &= \Lambda^1_2 \mathbf{e}_1 + \Lambda^2_2 \mathbf{e}_2 + \Lambda^3_2 \mathbf{e}_3 \\ \mathbf{e}'_3 &= \Lambda^1_3 \mathbf{e}_1 + \Lambda^2_3 \mathbf{e}_2 + \Lambda^3_3 \mathbf{e}_3 \end{aligned} \right\} \quad \text{or more simply} \quad \left\{ \begin{aligned} \mathbf{e}'_1 &= a^1 \mathbf{e}_1 + a^2 \mathbf{e}_2 + a^3 \mathbf{e}_3 \\ \mathbf{e}'_2 &= b^1 \mathbf{e}_1 + b^2 \mathbf{e}_2 + b^3 \mathbf{e}_3 \\ \mathbf{e}'_3 &= c^1 \mathbf{e}_1 + c^2 \mathbf{e}_2 + c^3 \mathbf{e}_3 \end{aligned} \right.$$

where the  $a$ 's are the components of  $\mathbf{e}'_1$  in the old basis, the  $b$ 's are the components of  $\mathbf{e}'_2$  in the old basis, and the  $c$ 's are the components of  $\mathbf{e}'_3$  in the old basis.

In contrast, equation 2.4.8 is a set of 3 *number* equations, relating the components of a *single vector*, taking its *old* components into the *new* basis. In other words, in the first equation, we are taking new basis vectors and expressing them in the old basis (new  $\rightarrow$  old). In the second equation, we are taking old components and converting them to the new basis (old  $\rightarrow$  new). The two equations go in opposite directions: the first takes new to

old, the second takes old to new. So it is natural that the two equations use inverse matrices to achieve those conversions. However, because of the inverse matrices in these equations, vector *components* are said to transform “contrary” (oppositely) to basis *vectors*, so they are called contravariant vectors.

I think it is misleading to say that contravariant vectors transform “oppositely” to basis vectors. In fact, that is impossible. Basis vectors are contravectors, and transform like any other contravector. A vector of (1, 0, 0) (in some basis) *is* a basis vector. It may also happen to be the value of some physical vector. In both cases, the expression of the vector (1, 0, 0) (old basis) in the new-basis is the same.

Now we can use 2.4.5 to evaluate the components of **T** in the primed basis:

$$T'_{ij} = T(\mathbf{e}'_i, \mathbf{e}'_j) = T(\Lambda^k_i \mathbf{e}_k, \Lambda^l_j \mathbf{e}_l) = \sum_{k=1}^N \sum_{l=1}^N \Lambda^k_i \Lambda^l_j T(\mathbf{e}_k, \mathbf{e}_l) = \sum_{k=1}^N \sum_{l=1}^N \Lambda^k_i \Lambda^l_j T_{kl}$$

Notice that there is one use of the transformation matrix  $\Lambda$  for each index of **T** to be transformed.

### Matrix View of Basis Transformation

The concept of tensors seems clumsy at first, but it’s a very fundamental concept. Once you get used to it, tensors are essentially simple things (though it took me 3 years to understand how “simple” they are). The rules for transformations are pretty direct. Transforming a rank-*n* tensor requires using the transformation matrix *n* times. A vector is rank-1, and transforms by a simple matrix multiply, or in tensor terms, by a summation over indices. Here, since we must distinguish row basis from column basis, we put the primes on the indices, to indicate which index is in the new basis, and which is in the old basis.

$$\mathbf{a}' = \Lambda \mathbf{a} \quad \leftrightarrow \quad \begin{bmatrix} a^{0'} \\ a^{1'} \\ a^{2'} \\ a^{3'} \end{bmatrix} = \begin{pmatrix} \Lambda^{0'0} & \Lambda^{0'1} & \Lambda^{0'2} & \Lambda^{0'3} \\ \Lambda^{1'0} & \Lambda^{1'1} & \Lambda^{1'2} & \Lambda^{1'3} \\ \Lambda^{2'0} & \Lambda^{2'1} & \Lambda^{2'2} & \Lambda^{2'3} \\ \Lambda^{3'0} & \Lambda^{3'1} & \Lambda^{3'2} & \Lambda^{3'3} \end{pmatrix} \begin{bmatrix} a^0 \\ a^1 \\ a^2 \\ a^3 \end{bmatrix} \quad \leftrightarrow \quad a^{\mu'} = \Lambda^{\mu'}_{\nu} a^{\nu}$$

Notice that when you sum over (contract over) two indices, they disappear, and you’re left with the unsummed index. So above when we sum over old-basis indices, we’re left with a new-basis vector.

Rank-2 example: The electromagnetic field tensor **F** is rank-2, and transforms using the transformation matrix twice, by two summations over indices, transforming both stress-energy indices. This is clumsy to write in matrix terms, because you have to use the transpose of the transformation matrix to transform the rows; this transposition has no physical significance. In the rank-2 (or higher) case, the tensor notation is both simpler, and more physically meaningful:

$$\mathbf{F}' = \Lambda \mathbf{F} \Lambda^T \quad \leftrightarrow \quad \begin{bmatrix} F^{0'0'} & F^{0'1'} & F^{0'2'} & F^{0'3'} \\ F^{1'0'} & F^{1'1'} & F^{1'2'} & F^{1'3'} \\ F^{2'0'} & F^{2'1'} & F^{2'2'} & F^{2'3'} \\ F^{3'0'} & F^{3'1'} & F^{3'2'} & F^{3'3'} \end{bmatrix} = \begin{pmatrix} \Lambda^{0'0} & \Lambda^{0'1} & \Lambda^{0'2} & \Lambda^{0'3} \\ \Lambda^{1'0} & \Lambda^{1'1} & \Lambda^{1'2} & \Lambda^{1'3} \\ \Lambda^{2'0} & \Lambda^{2'1} & \Lambda^{2'2} & \Lambda^{2'3} \\ \Lambda^{3'0} & \Lambda^{3'1} & \Lambda^{3'2} & \Lambda^{3'3} \end{pmatrix} \begin{bmatrix} F^{00} & F^{01} & F^{02} & F^{03} \\ F^{10} & F^{11} & F^{12} & F^{13} \\ F^{20} & F^{21} & F^{22} & F^{23} \\ F^{30} & F^{31} & F^{32} & F^{33} \end{bmatrix} \begin{pmatrix} \Lambda^{0'0} & \Lambda^{1'0} & \Lambda^{2'0} & \Lambda^{3'0} \\ \Lambda^{0'1} & \Lambda^{1'1} & \Lambda^{2'1} & \Lambda^{3'1} \\ \Lambda^{0'2} & \Lambda^{1'2} & \Lambda^{2'2} & \Lambda^{3'2} \\ \Lambda^{0'3} & \Lambda^{1'3} & \Lambda^{2'3} & \Lambda^{3'3} \end{pmatrix} \quad \leftrightarrow \quad F^{\mu'\nu'} = \Lambda^{\mu'}_{\nu} \Lambda^{\nu'}_{\rho} F^{\mu\nu}$$

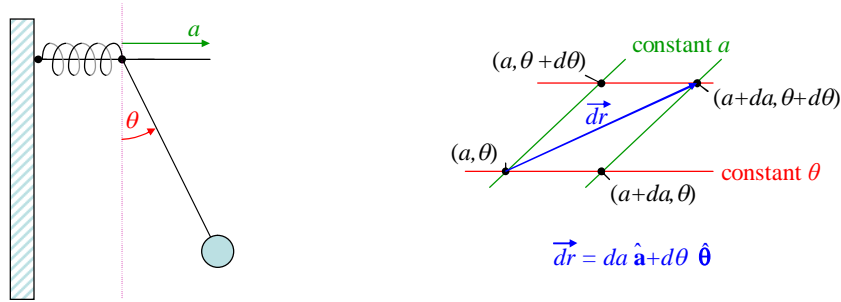
In general, you have to transform every index of a tensor, each index requiring one use of the transformation matrix.

---

### Non-Orthonormal Systems: Contravariance and Covariance

Many systems *cannot* be represented with orthonormal coordinates, e.g. the (surface of a) sphere. Dealing with non-orthonormality *requires* a more sophisticated view of tensors, and introduces the concepts of contravariance and covariance.

Consider the following problem from classical mechanics: a pendulum is suspended from a pivot point which slides horizontally on a spring. The generalized coordinates are (*a*,  $\theta$ ).



To compute kinetic energy, we need to compute  $|\mathbf{v}|^2$ , conveniently done in some orthogonal coordinates, say  $x$  and  $y$ . We start by converting the generalized coordinates to the orthonormal  $x$ - $y$  coordinates, to compute the length of a physical displacement from the changes in generalized coordinates:

$$\begin{aligned} x &= a + l \sin \theta, & dx &= da + l \cos \theta d\theta \\ y &= l \cos \theta, & dy &= -l \sin \theta d\theta \\ \Rightarrow & & ds^2 &= dx^2 + dy^2 = da^2 + 2l \cos \theta da d\theta + l^2 \cos^2 \theta d\theta^2 + l^2 \sin^2 \theta d\theta^2 \\ & & &= da^2 + 2l \cos \theta da d\theta + l^2 d\theta^2 \end{aligned}$$

We have just computed the metric tensor field, which is a function of position in the  $(a, \theta)$  configuration space. We can write the metric tensor field components by inspection:

$$\begin{aligned} \text{Let } x^1 &= a, x^2 = \theta \\ ds^2 &= \sum_{i=1}^2 \sum_{j=1}^2 g_{ij} dx^i dx^j = da^2 + 2l \cos \theta da d\theta + l^2 d\theta^2 \quad \Rightarrow \quad g_{ij} = \begin{pmatrix} 1 & l \cos \theta \\ l \cos \theta & l^2 \end{pmatrix} \end{aligned}$$

Then  $|\mathbf{v}|^2 = ds^2/dt^2$ . A key point here is that the *same* metric tensor computes a physical displacement from generalized coordinate displacements, or a physical velocity from generalized coordinate velocities, or a physical acceleration from generalized coordinate accelerations, etc., because time is the same for any generalized coordinate system (no Relativity here!). Note that we symmetrize the cross-terms of the metric,  $g_{ij} = g_{ji}$ , which is necessary to insure that  $\mathbf{g}(\mathbf{v}, \mathbf{w}) = \mathbf{g}(\mathbf{w}, \mathbf{v})$ .

Now consider the scalar product of two vectors. The same metric tensor (field) helps compute the scalar product (dot product) of any two (infinitesimal) vectors, from their generalized coordinates:

$$d\mathbf{v} \cdot d\mathbf{w} = \mathbf{g}(d\mathbf{v}, d\mathbf{w}) = g_{ij} dv^i dw^j$$

Since the metric tensor takes two input vectors, is linear in both, and produces a scalar result, it is a rank-2 tensor. Also, since  $\mathbf{g}(\mathbf{v}, \mathbf{w}) = \mathbf{g}(\mathbf{w}, \mathbf{v})$ ,  $\mathbf{g}$  is a **symmetric** tensor.

Now, let's define a scalar field as a function of the generalized coordinates; say, the potential energy:

$$U = \frac{k}{2} a^2 - mg \cos \theta$$

It is quite useful to know the gradient of the potential energy:

$$\mathbf{D} = \nabla U = \frac{\partial U}{\partial a} \boldsymbol{\omega}^a + \frac{\partial U}{\partial \theta} \boldsymbol{\omega}^\theta \quad \Rightarrow \quad dU = \mathbf{D}(d\mathbf{r}) = \frac{\partial U}{\partial a} da + \frac{\partial U}{\partial \theta} d\theta$$

The gradient takes an infinitesimal displacement vector  $d\mathbf{r} = (da, d\theta)$ , and produces a differential in the value of potential energy,  $dU$  (a scalar). Further,  $dU$  is a linear function of the displacement vector. Hence, by definition, the gradient at each point in  $a$ - $\theta$  space is a rank-1 tensor, i.e. the gradient is a tensor field.

Do we need to use the metric (computed earlier) to make the gradient operate on  $d\mathbf{r}$ ? No! The gradient operates directly on  $d\mathbf{r}$ , without the need for any “assistance” by a metric. So the gradient is a rank-1 tensor that can directly contract with a vector to produce a scalar. This is markedly different from the dot product case above, where the first vector (a rank-1 tensor) could *not* contract directly with an input vector to produce a scalar. So clearly,

There are two kinds of rank-1 tensors: those (like the gradient) that can contract directly with an input vector, and those that need the metric to “help” them operate on an input vector.

Those tensors that can operate directly on a vector are called **covariant tensors**, and those that need help are called **contravariant**, for reasons we will show soon. To indicate that  $\mathbf{D}$  is covariant, we write its components with subscripts, instead of superscripts. Its basis vectors are covariant vectors, related to  $\mathbf{e}_1$ ,  $\mathbf{e}_2$ , and  $\mathbf{e}_3$ :

$$\mathbf{D} = D_i \omega^i = D_a \omega^a + D_\theta \omega^\theta \quad \text{where } \omega^r, \omega^\theta \text{ are covariant basis vectors}$$

In general, covariant tensors result from differentiation operators on other (scalar or) tensor fields: gradient, covariant derivative, exterior derivative, Lie derivative, etc.

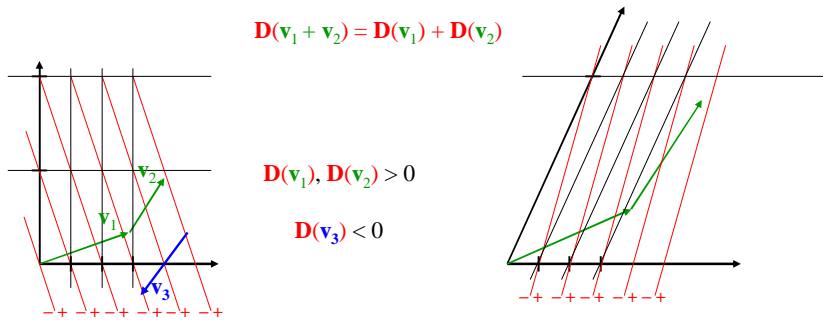
Note that just as we can say that  $\mathbf{D}$  acts on  $d\mathbf{r}$ , we can say that  $d\mathbf{r}$  is a rank-1 tensor that acts on  $\mathbf{D}$  to produce  $dU$ :

$$\mathbf{D}(d\mathbf{r}) = d\mathbf{r}(\mathbf{D}) = \sum_i \frac{\partial U}{\partial x^i} dx^i = \frac{\partial U}{\partial a} da + \frac{\partial U}{\partial \theta} d\theta$$

The contractions are the same with either acting on the other, so the definitions are symmetric.

Interestingly, when we compute small oscillations of a system of particles, we need both the potential matrix, which is the gradient of the gradient of the potential field, and the “mass” matrix, which really gives us kinetic energy. The potential matrix is fully covariant, and we need no metric to compute it. The kinetic energy matrix requires us to compute absolute magnitudes of  $|\mathbf{v}|^2$ , and so requires us to compute the metric.

We know that a vector, which is a rank-1 tensor, can be visualized as an arrow. How do we visualize this covariant tensor, in a way that reveals how it operates on a vector (an arrow)? We use a set of equally spaced parallel planes. Let  $\mathbf{D}$  be a covariant tensor (aka 1-form):



Visualization of a covariant vector (1-form) as oriented parallel planes. The 1-form is a linear operator on vectors (see text).

The value of  $\mathbf{D}$  on a vector,  $\mathbf{D}(\mathbf{v})$ , is the number of planes “pierced” by the vector when laid on the parallel planes. Clearly,  $\mathbf{D}(\mathbf{v})$  depends on the magnitude and direction of  $\mathbf{v}$ . It is also a linear function of  $\mathbf{v}$ : the sum of planes pierced by two different vectors equals the number of planes pierced by their vector sum.

There is an orientation to the planes. One side is negative, and the other positive. Vectors crossing in the negative to the positive direction “pierce” a positive number of planes. Vectors crossing in the positive to negative direction “pierce” a negative number of planes.

Note also we could redraw the two axes arbitrarily oblique (non-orthogonal), and rescale the axes arbitrarily, but keeping the intercept values of the planes with the axes unchanged (thus stretching the arrows and planes). The number of planes pierced would be the same, so the two diagrams above are equivalent. Hence, this geometric construction of the operation of a covector on a contravector is completely general, and even applies to vector spaces which have no metric (aka “non-metric” spaces). All you need for the construction is a set of arbitrary basis vectors (not necessarily orthonormal), and the values  $\mathbf{D}(e_i)$  on each, and you can draw the parallel planes that illustrate the covector.

The “direction” of  $\mathbf{D}$ , analogous to the direction of a vector, is *normal to* (perpendicular to) the planes used to graphically represent  $\mathbf{D}$ .

### What Goes Up Can Go Down: Duality of Contravariant and Covariant Vectors

Recall the dot product is given by

$$d\mathbf{v} \cdot d\mathbf{w} = \mathbf{g}(d\mathbf{v}, d\mathbf{w}) = g_{ij} dv^i dw^j$$

If I fill only one slot of  $\mathbf{g}$  with  $\mathbf{v}$ , and leave the 2<sup>nd</sup> slot empty, then  $\mathbf{g}(\mathbf{v}, \_)$  is a linear function of one vector, and can be directly contracted with that vector; hence  $\mathbf{g}(\mathbf{v}, \_)$  is a rank-1 covariant vector. For any given contravariant vector  $v^j$ , I can define this “dual” covariant vector,  $\mathbf{g}(\mathbf{v}, \_)$ , which has  $N$  components I’ll call  $v_i$ .

$$v_i = \mathbf{g}(\mathbf{v}, \_) = g_{ik} v^k$$

So long as I have a metric, the contravariant and covariant forms of  $\mathbf{v}$  contain equivalent information, and are thus two ways of expressing the same vector (geometric object).

The covariant representation can contract directly with a contravariant vector, and the contravariant representation can contract directly with a covariant vector, to produce the dot product of the two vectors. Therefore, we can use the metric tensor to “lower” the components of a contravariant vector into their covariant equivalents.

Note that the metric tensor itself has been written with two covariant (lower) indexes, because it contracts directly with two contravariant vectors to produce their scalar-product.

Why do I need two forms of the same vector? Consider the vector “force.”

$$\mathbf{F} = m\mathbf{a} \quad \text{or} \quad F^i = ma^i \quad \text{(naturally contravariant)}$$

Since position  $x^j$  is naturally contravariant, so is its derivative  $v^j$ , and 2<sup>nd</sup> derivative,  $a^i$ . Therefore, force is “naturally” contravariant. But force is also the gradient of potential energy:

$$\mathbf{F} = -\nabla U \quad \text{or} \quad F_i = -\frac{\partial}{\partial x^i} U \quad \text{(naturally covariant)}$$

Oops! Now “force” is naturally *covariant*! But it’s the same force as above. So which is more natural for “force?” Neither. Use whichever one you need. Nurture supersedes nature.

The inverse of the metric tensor matrix is the contravariant metric tensor,  $g^{ij}$ . It contracts directly with two covariant vectors to produce their scalar product. Hence, we can use  $g^{ij}$  to “raise” the index of a covariant vector to get its contravariant components.

$$v^i = \mathbf{g}(\mathbf{v}, \_) = g^{ik} v_k \quad \quad \quad g^{ik} g_{kj} = g^i_j$$

Notice that raising and lowering works on the metric tensor itself. Note that in general, even for symmetric tensors,  $T_i^j \neq T_j^i$ , and  $T_i^j \neq T_j^i$ .

For rank-2 or higher tensors, each index is separately of the contravariant or covariant type. Each index may be raised or lowered separately from the others. Each lowering requires a contraction with the fully covariant metric tensor; each raising requires a contraction with the fully contravariant metric tensor.

In Euclidean space with orthonormal coordinates, the metric tensor is the identity matrix. Hence, the covariant and contravariant components of any vector are identical. This is why there is no distinction made in elementary treatments of vector mathematics; displacements, gradients, everything, are simply called “vectors.”

The space of covectors is a vector space, i.e. it satisfies the properties of a vector space. However, it is called “dual” to the vector space of contravectors, because covectors operate on contravectors to produce scalar invariants. A thing is **dual** to another thing if the dual can act on the original thing to produce a scalar, and vice versa. E.g., in QM, bras are dual to kets. “Vectors in the dual space” are covectors.

Just like basis contravectors, basis covectors always have components (in their own basis)

$$\omega^1 = (1,0,0\dots), \quad \omega^2 = (0,1,0\dots), \quad \omega^3 = (0,0,1\dots), \quad \text{etc.}$$

and we can write an arbitrary covector as  $\tilde{\mathbf{f}} = f_1\omega^1 + f_2\omega^2 + f_3\omega^3 + \dots$ .

TBS: construction and units of a dual covector from its contravector.

### The *Real* Summation Convention

The summation convention says repeated indexes in an arithmetic expression are implicitly summed (contracted). We now understand that only a contravariant/covariant pair can be meaningfully summed. Two covariant or two contravariant indexes require contracting with the metric tensor to be meaningful. Hence, the *real* Einstein summation convention is that any two matching indexes, one “up” (contravariant) and one “down” (covariant), are implicitly summed (contracted). Two matching contravariant or covariant indexes are meaningless, and not allowed.

Now we can see why basis contravectors are written  $\mathbf{e}_1, \mathbf{e}_2, \dots$  (with subscripts), and basis covectors are written  $\omega^1, \omega^2, \dots$  (with superscripts). It is purely a trick to comply with the *real* summation convention that requires summations be performed over one “up” index and one “down” index. Then we can write a vector as a linear combination of the basis vectors, using the summation convention:

$$\mathbf{v} = v^1\mathbf{e}_1 + v^2\mathbf{e}_2 + v^3\mathbf{e}_3 = v^i\mathbf{e}_i \qquad \tilde{\mathbf{a}} = a_1\omega^1 + a_2\omega^2 + a_3\omega^3 = a_i\omega^i$$

Note well that there is nothing “covariant” about  $\mathbf{e}_i$ , even though it has a subscript; there is nothing “contravariant” about  $\omega^i$ , even though it has a superscript. It’s just a notational trick.

### Transformation of Covariant Indexes

It turns out that the components of a covariant vector transform with the same matrix as used to express the new (primed) basis vectors in the old basis:

$$f'_k = f_j \Lambda^j_k \qquad \text{[Tal 2.4.11]}$$

Again, somewhat bogusly, eq. 2.4.11 is said to “transform covariantly with” (the same as) the basis vectors, so ‘ $f'_i$ ’ is called a **covariant** vector.

For a rank-2 tensor such as  $T_{ij}$ , each index of  $T_{ij}$  transforms “like” the basis vectors (i.e., covariantly with the basis vectors). Hence, each index of  $T_{ij}$  is said to be a “covariant” index. Since both indexes are covariant,  $T_{ij}$  is sometimes called “fully covariant.”

## Indefinite Metrics: Relativity

In short, a covariant index of a tensor is one which can be contracted with (summed over) a contravariant index of an input MVE to produce a meaningful resultant MVE.

In relativity, the metric tensor has some negative signs. The scalar-product is a frame-invariant “interval.” No problem. All the math, raising, and lowering, works just the same. In special relativity, the metric ends up simply putting minus signs where you need them to get SR intervals. The covariant form of a vector has the minus signs “pre-loaded,” so it contracts directly with a contravariant vector to produce a scalar.

Let's use the sign convention where  $\eta_{\mu\nu} = \text{diag}(-1, 1, 1, 1)$ . When considering the dual 1-forms for Minkowski space, the only unusual aspect is that the 1-form for time increases in the *opposite* direction as the vector for time. For the space components, the dual 1-forms increase in the *same* direction as the vectors. This means that

$$\omega^t \mathbf{e}_t = -1, \quad \omega^x \mathbf{e}_x = 1, \quad \omega^y \mathbf{e}_y = 1, \quad \omega^z \mathbf{e}_z = 1$$

as it should for the Minkowski metric.

---

## Is a Transformation Matrix a Tensor?

Sort of. When applied to a vector, it converts components from the “old” basis to the “new” basis. It is clearly a linear function of its argument. However, a tensor usually has all its inputs and outputs in the *same* basis (or tensor products of that basis). But a transformation matrix is specifically constructed to take inputs in one basis, and produce outputs in a *different* basis. Essentially, the columns are indexed by the old basis, and the rows are indexed by the new basis. It basically works like a tensor, but the transformation rule is that to transform the columns, you use a transformation matrix for the old basis; to transform the rows, you use the transformation matrix for the new basis.

Consider a vector

$$\mathbf{v} = v^1 \mathbf{e}_1 + v^2 \mathbf{e}_2 + v^3 \mathbf{e}_3$$

This is a vector equation, and despite its appearance, it is true in *any* basis, not just the  $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$  basis. If we write  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$  as vectors in some new  $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$  basis, the vector equation above still holds:

$$\mathbf{e}_1 = (\mathbf{e}_1)^x \mathbf{e}_x + (\mathbf{e}_1)^y \mathbf{e}_y + (\mathbf{e}_1)^z \mathbf{e}_z$$

$$\mathbf{e}_2 = (\mathbf{e}_2)^x \mathbf{e}_x + (\mathbf{e}_2)^y \mathbf{e}_y + (\mathbf{e}_2)^z \mathbf{e}_z$$

$$\mathbf{e}_3 = (\mathbf{e}_3)^x \mathbf{e}_x + (\mathbf{e}_3)^y \mathbf{e}_y + (\mathbf{e}_3)^z \mathbf{e}_z$$

$$\mathbf{v} = v^1 \mathbf{e}_1 + v^2 \mathbf{e}_2 + v^3 \mathbf{e}_3$$

$$= v^1 \underbrace{\left[ (\mathbf{e}_1)^x \mathbf{e}_x + (\mathbf{e}_1)^y \mathbf{e}_y + (\mathbf{e}_1)^z \mathbf{e}_z \right]}_{\mathbf{e}_1} + v^2 \underbrace{\left[ (\mathbf{e}_2)^x \mathbf{e}_x + (\mathbf{e}_2)^y \mathbf{e}_y + (\mathbf{e}_2)^z \mathbf{e}_z \right]}_{\mathbf{e}_2} + v^3 \underbrace{\left[ (\mathbf{e}_3)^x \mathbf{e}_x + (\mathbf{e}_3)^y \mathbf{e}_y + (\mathbf{e}_3)^z \mathbf{e}_z \right]}_{\mathbf{e}_3}$$

The vector  $\mathbf{v}$  is just a weighted sum of basis vectors, and therefore the columns of the transformation matrix are the old basis vectors expressed in the new basis. E.g., to transform the components of a vector from the  $(\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$  to the  $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$  basis, the transformation matrix is

$$\begin{bmatrix} (\mathbf{e}_1)^x & (\mathbf{e}_2)^x & (\mathbf{e}_3)^x \\ (\mathbf{e}_1)^y & (\mathbf{e}_2)^y & (\mathbf{e}_3)^y \\ (\mathbf{e}_1)^z & (\mathbf{e}_2)^z & (\mathbf{e}_3)^z \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1 \cdot \mathbf{e}_x & \mathbf{e}_2 \cdot \mathbf{e}_x & \mathbf{e}_3 \cdot \mathbf{e}_x \\ \mathbf{e}_1 \cdot \mathbf{e}_y & \mathbf{e}_2 \cdot \mathbf{e}_y & \mathbf{e}_3 \cdot \mathbf{e}_y \\ \mathbf{e}_1 \cdot \mathbf{e}_z & \mathbf{e}_2 \cdot \mathbf{e}_z & \mathbf{e}_3 \cdot \mathbf{e}_z \end{bmatrix}$$

You can see directly that the first column is  $\mathbf{e}_1$  written in the  $x$ - $y$ - $z$  basis; the 2<sup>nd</sup> column is  $\mathbf{e}_2$  in the  $x$ - $y$ - $z$  basis; and the 3<sup>rd</sup> column is  $\mathbf{e}_3$  in the  $x$ - $y$ - $z$  basis.

---

## How About the Pauli Vector?

In quantum mechanics, the Pauli vector is a vector of three 2x2 matrices: the Pauli matrices. Each 2x2 complex-valued matrix corresponds to a spin-1/2 operator in some  $x$ ,  $y$ , or  $z$  direction. It is a 3<sup>rd</sup> rank object in the tensor product space of  $\mathbf{R}^3 \otimes \mathbf{C}^2 \otimes \mathbf{C}^2$ , i.e.  $xyz \otimes \text{spinor} \otimes \text{spinor}$ . The  $xyz$  rank is clearly in a different basis than the complex spinor ranks, since  $xyz$  is a completely different vector space than spin-1/2



spinor space. However, it is a linear operator on various objects, so each rank transforms according to the transformation matrix for its basis.

$$\vec{\sigma} = \left( \begin{matrix} x & y & z \\ \left[ \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix} \right], & \left[ \begin{matrix} 0 & -i \\ i & 0 \end{matrix} \right], & \left[ \begin{matrix} 1 & 0 \\ 0 & -1 \end{matrix} \right] \end{matrix} \right)$$

It's interesting to note that the term **tensor product** produces, in general, an object of mixed bases, and often, mixed vector spaces. Nonetheless, the term “tensor” seems to be used most often for mathematical objects whose ranks are all in the same basis.

## Cartesian Tensors

**Cartesian tensors** aren't quite tensors, because they don't transform into non-Cartesian coordinates properly. (Note that despite their name, Cartesian tensors are *not* a special kind of tensor; they aren't really tensors. They're tensor wanna-be's.) Cartesian tensors have two failings that prevent them from being true tensors: they don't distinguish between contravariant and covariant components, and they treat finite displacements in space as vectors. In non-orthogonal coordinates, you must distinguish contravariant and covariant components. In non-Cartesian coordinates, only infinitesimal displacements are vectors. Details:

Recall that in Cartesian coordinates, there is no distinction between contravariant and covariant components of a tensor. This allows a certain sloppiness that one can only get away with if one sticks to Cartesian coordinates. This means that Cartesian “tensors” only transform reliably by rotations from one set of Cartesian coordinates to a new, rotated set of Cartesian coordinates. Since both the new and old bases are Cartesian, there is no need to distinguish contravariant and covariant components in either basis, and the transformation (to a rotated coordinate system) “works.”

For example, the moment of inertia “tensor” is a Cartesian tensor. There is no problem in its first use, to compute the angular momentum of a blob of mass given its angular velocity:

$$\mathbf{I}(\boldsymbol{\omega}, \_) = \mathbf{L} \quad \Rightarrow \quad L^i = I_j^i \omega^j \quad \Rightarrow$$

$$\begin{bmatrix} L^x \\ L^y \\ L^z \end{bmatrix} = \begin{bmatrix} I^x_x & I^x_y & I^x_z \\ I^y_x & I^y_y & I^y_z \\ I^z_x & I^z_y & I^z_z \end{bmatrix} \begin{bmatrix} \omega^x \\ \omega^y \\ \omega^z \end{bmatrix} = \omega^x \begin{bmatrix} I^x_x \\ I^y_x \\ I^z_x \end{bmatrix} + \omega^y \begin{bmatrix} I^x_y \\ I^y_y \\ I^z_y \end{bmatrix} + \omega^z \begin{bmatrix} I^x_z \\ I^y_z \\ I^z_z \end{bmatrix}$$

But notice that if **I** accepts a contravariant vector, then **I**'s components for that input vector must be covariant. However, **I** produces a contravariant output, so its output components are contravariant. So far, so good.

But now we want to find the kinetic energy. Well,  $\frac{1}{2} I \omega^2 = \frac{1}{2} \mathbf{L} \cdot \boldsymbol{\omega} = \left( \frac{1}{2} \mathbf{I}(\boldsymbol{\omega}, \_) \right) \cdot \boldsymbol{\omega}$ . But we have a dot product of two contravariant vectors. To evaluate that dot product, in a general coordinate system, we have to use the metric:

$$KE = \frac{1}{2} I_j^i \omega^j \omega_i = \frac{1}{2} I_j^i \omega^j g_{ik} \omega^k \quad \neq \frac{1}{2} I_j^i \omega^j \omega^i$$

However, in Cartesian coordinates, the metric matrix is the identity matrix, the contravariant components equal the covariant components, and the final “not-equals” above becomes an “equals.” Hence, we neglect the distinction between contravariant components and covariant components, and “incorrectly” sum the components of **I** on the components of  $\boldsymbol{\omega}$ , even though both are contravariant in the 2<sup>nd</sup> sum.

In general coordinates, the direct sum for the dot product doesn't work, and you must use the metric tensor for the final dot product.

Example of failure of finite displacements: TBS: The electric quadrupole tensor acts on two copies of the finite displacement vector to produce the electric potential at that displacement. Even in something as simple as polar coordinates, this method fails.

### The Real Reason Why the Kronecker Delta Is Symmetric

TBS: Because it a mixed tensor,  $\delta^\alpha_\beta$ . Symmetry can only be assessed by comparing interchange of two indices of the same "up-" or "down-ness" (contravariance or covariance). We can lower, say  $\alpha$ , in  $\delta^\alpha_\beta$  with the metric:

$$\delta_{\alpha\beta} = g_{\alpha\gamma} \delta^\gamma_\beta = g_{\alpha\beta}$$

The result the metric, which is always symmetric. Hence,  $\delta^\alpha_\beta$  is a symmetric tensor, but *not* because its matrix looks symmetric. In general, a mixed rank-2 symmetric tensor does *not* have a symmetric matrix representation. Only when both indices are up or both down is its matrix symmetric.

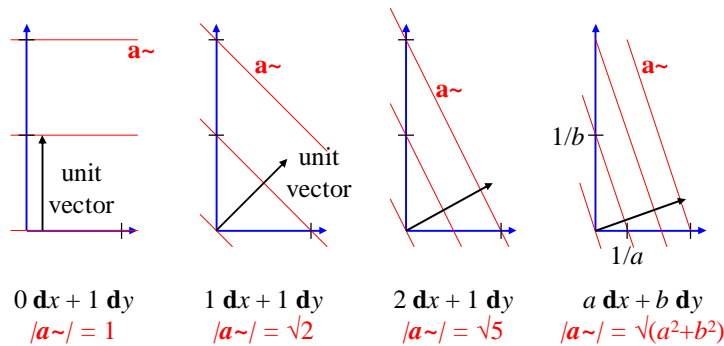
The Kronecker delta is a special case that does not generalize.

Things are not always what they seem.

### Tensor Appendices

#### Pythagorean Relation for 1-forms

Demonstration that 1-forms satisfy the Pythagorean relation for magnitude:



Examples of 3 1-forms, and a generic 1-form. Here,  $\mathbf{d}x$  is the  $x$  basis 1-form, and  $\mathbf{d}y$  is the  $y$  basis 1-form.

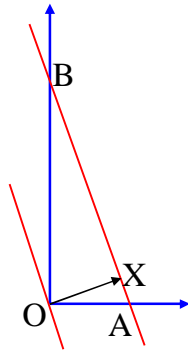
From the diagram above, a max-crossing vector (perpendicular to the planes of  $\mathbf{a}\sim$ ) has  $(x, y)$  components  $(1/b, 1/a)$ . Dividing by its magnitude, we get a unit vector:

$$\text{max crossing unit vector } \mathbf{u} = \frac{\frac{1}{b} \hat{\mathbf{x}} + \frac{1}{a} \hat{\mathbf{y}}}{\sqrt{\frac{1}{b^2} + \frac{1}{a^2}}}. \quad \text{Note that} \quad \mathbf{d}x(\hat{\mathbf{x}}) = 1, \text{ and } \mathbf{d}y(\hat{\mathbf{y}}) = 1$$

The magnitude of a 1-form is the scalar resulting from the 1-form's action on a max-crossing unit vector:

$$|\tilde{a}| = \tilde{a}(u) = \frac{(a\mathbf{d}x + b\mathbf{d}y)\left(\frac{1}{b}\hat{\mathbf{x}} + \frac{1}{a}\hat{\mathbf{y}}\right)}{\sqrt{\frac{1}{b^2} + \frac{1}{a^2}}} = \frac{\left(\frac{a}{b} + \frac{b}{a}\right)}{\sqrt{\frac{1}{b^2} + \frac{1}{a^2}}} = \frac{(a^2 + b^2)}{ab\sqrt{\frac{1}{b^2} + \frac{1}{a^2}}} = \frac{(a^2 + b^2)}{\sqrt{a^2 + b^2}} = \sqrt{a^2 + b^2}$$

Here's another demonstration that 1-forms satisfy the Pythagorean relation for magnitude. The magnitude of a 1-form is the inverse of the plane spacing:



$$\begin{aligned} \Delta OXA \sim \Delta BOA &\Rightarrow \frac{OX}{OA} = \frac{BO}{BA} \\ \Rightarrow OX &= \frac{(BO)(OA)}{BA} = \frac{\frac{1}{b} \cdot \frac{1}{a}}{\sqrt{\frac{1}{b^2} + \frac{1}{a^2}}} \\ |\tilde{a}| = \frac{1}{OX} &= \frac{\sqrt{\frac{1}{b^2} + \frac{1}{a^2}}}{\frac{1}{b} \cdot \frac{1}{a}} = ab\sqrt{\frac{1}{b^2} + \frac{1}{a^2}} = \sqrt{a^2 + b^2} \end{aligned}$$

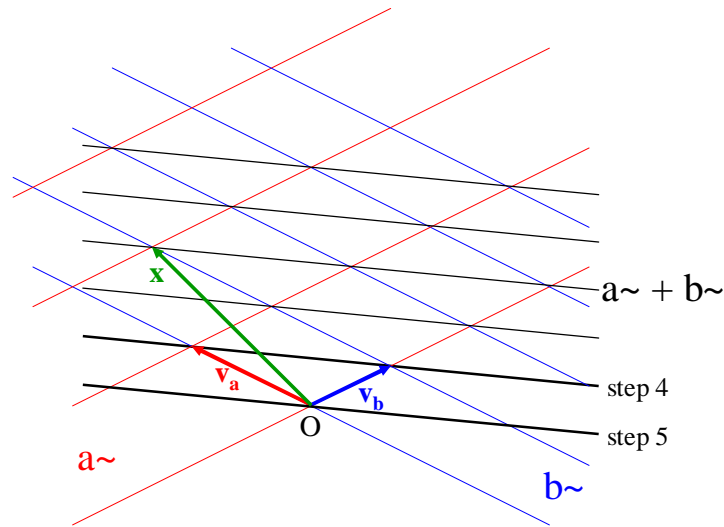
**Geometric Construction Of The Sum Of Two 1-Forms:**

Example of  $\tilde{a} + \tilde{b}$

$\tilde{a}(\mathbf{x}) = 2$
$\tilde{b}(\mathbf{x}) = 1$
$(\tilde{a} + \tilde{b})(\mathbf{x}) = 3$

Construction of  $\tilde{a} + \tilde{b}$

$\tilde{a}(\mathbf{v}_a) = 1$	$\tilde{a}(\mathbf{v}_b) = 0$
$\tilde{b}(\mathbf{v}_a) = 0$	$\tilde{b}(\mathbf{v}_b) = 1$
$(\tilde{a} + \tilde{b})(\mathbf{v}_a) = 1$	$(\tilde{a} + \tilde{b})(\mathbf{v}_b) = 1$



To construct the sum of two 1-forms,  $\tilde{a} + \tilde{b}$ :

1. Choose an origin at the intersection of a plane of  $\tilde{a}$  and a plane of  $\tilde{b}$ .
2. Draw vector  $\mathbf{v}_a$  from the origin along the planes of  $\tilde{b}$ , so  $\tilde{b}(\mathbf{v}_a) = 0$ , and of length such that  $\tilde{a}(\mathbf{v}_a) = 1$ . [This is the dual vector of  $\tilde{a}$ .]

3. Similarly, draw  $\mathbf{v}_b$  from the origin along the planes of  $\mathbf{a}$ , so  $\mathbf{a} \cdot (\mathbf{v}_b) = 0$ , and  $\mathbf{b} \cdot (\mathbf{v}_b) = 1$ . [This is the dual vector of  $\mathbf{b}$ .]
4. Draw a plane through the heads of  $\mathbf{v}_a$  and  $\mathbf{v}_b$  (black above). This defines the orientation of  $(\mathbf{a} + \mathbf{b})$ .
5. Draw a parallel plane through the common point (the origin). This defines the spacing of planes of  $(\mathbf{a} + \mathbf{b})$ .
6. Draw all other planes parallel, and with the same spacing. This is the geometric representation of  $(\mathbf{a} + \mathbf{b})$ .

Now we can easily draw the test vector  $\mathbf{x}$ , such that  $\mathbf{a} \cdot (\mathbf{x}) = 2$ , and  $\mathbf{b} \cdot (\mathbf{x}) = 1$ .

### “Fully Anti-symmetric” Symbols Expanded

Everyone hears about them, but few ever see them. They are quite sparse: the 3-D fully anti-symmetric symbol has 6 nonzero values out of 27; the 4-D one has 24 nonzero values out of 256.

3-D, from the 6 permutations,  $ijk$ : 123+, 132-, 312+, 321-, 231+, 213-

$$\epsilon_{ijk} = \begin{matrix} k=1 & k=2 & k=3 \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

4-D, from the 24 permutations,  $\alpha\beta\gamma\delta$ :

0123+    0132-    0312+    0321-    0231+    0213-  
 1023-    1032+    1302-    1320+    1230-    1203+  
 2013+    2031-    2301+    2310-    2130+    2103-  
 3012-    3021+    3201-    3210+    3120-    3102+

$$\epsilon_{\alpha\beta\gamma\delta} = \begin{matrix} \beta=0 & \beta=1 & \beta=2 & \beta=3 \\ \alpha=0 & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \alpha=1 & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \alpha=2 & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \alpha=3 & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, & \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{matrix}$$

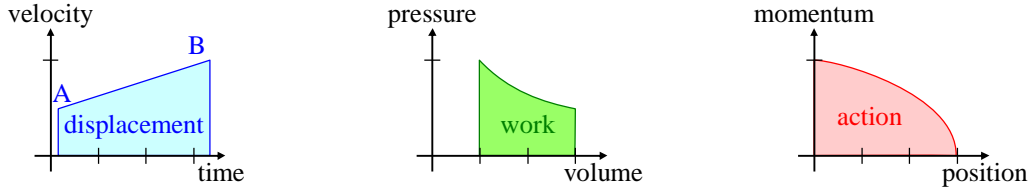
## Metric? We Don't Need No Stinking Metric!

### Examples of Useful, Non-metric Spaces

Non-metric spaces are everywhere. A non-metric space has no concept of “distance” between arbitrary points, or even between arbitrary “nearby” points (points with infinitesimal coordinate differences). However:

Non-metric spaces have no concept of “distance,”  
but many still have a well-defined concept of “area,” in the sense of an integral.

For example, consider a plot of velocity (of a particle in 1D) vs. time (below, left).



Some useful non-metric spaces: (left) velocity vs. time; (middle) pressure vs. volume; (right) momentum vs. position. In each case, there is no distance, but there *is* area.

The area under the velocity curve is the total displacement covered. The area under the  $P$ - $V$  curve is the work done by an expanding fluid. The area under the momentum-position curve ( $p$ - $q$ ) is the action of the motion in classical mechanics. Though the points in each of these plots exist on 2D manifolds, the two coordinates are incomparable (they have different units). It is meaningless to ask what is the distance between two arbitrary points on the plane. For example, points A and B on the  $v$ - $t$  curve differ in both velocity and time, so how could we define a distance between them (how can we add m/s and seconds)?

In the above cases, we have one coordinate value as a function of the other, e.g. velocity as a function of time. We now consider another case: rather than consider the function as one of the coordinates in a manifold, we consider the manifold as comprising only the independent variables. Then, the function is defined *on* that manifold. As usual, keeping track of the units of all the quantities will help in understanding both the physical and mathematical principles.

For example, the speed of light in air is a function of 3 independent variables: temperature, pressure, and humidity. At 633 nm, the effects amount to speed changes of about +1 ppm per kelvin, -0.4 ppm per mm-Hg pressure, and +0.01 ppm per 1% change in relative humidity (RH) (see <http://patapsco.nist.gov/mel/div821/Wavelength/Documentation.asp#CommentsRegardingInputstotheEquations>):

$$s(T, P, H) = s_0 + aT - bP + cH .$$

where  $a \approx 300$  (m/s)/k,  $b \approx 120$  (m/s)/mm-Hg, and  $c \approx 3$  (m/s)/% are positive constants, and the function  $s$  is the speed of light at the given conditions, in m/s. Our manifold is the set of TPH triples, and  $s$  is a function on that manifold. We can consider the TPH triple as a (contravariant, column) vector:  $(T, P, H)^T$ . These vectors constitute a 3D vector space over the field of reals.  $s(\cdot)$  is a real function on that vector space.

Note that the 3 components of a vector each have different units: the temperature is measured in kelvins (K), the pressure in mm-Hg, and the relative humidity in %. Note also that there is no metric on  $(T, P, H)$  space (which is bigger, 1 K or 1 mm-Hg?). However, the gradient of  $s$  is still well defined:

$$\nabla s = \frac{\partial s}{\partial T} \tilde{\mathbf{d}}T + \frac{\partial s}{\partial P} \tilde{\mathbf{d}}P + \frac{\partial s}{\partial H} \tilde{\mathbf{d}}H = a \tilde{\mathbf{d}}T - b \tilde{\mathbf{d}}P + c \tilde{\mathbf{d}}H .$$

What are the units of the gradient? As with the vectors, each component has different units: the first is in (m/s) per kelvin; the second in (m/s) per mm-Hg; the third in (m/s) per %. The gradient has different units than the vectors, and is not a part of the original vector space. The gradient,  $\nabla s$ , operates on a vector  $(T, P, H)^T$  to give the change in speed from one set of conditions, say  $(T_0, P_0, H_0)$  to conditions incremented by the vector  $(T_0 + T, P_0 + P, H_0 + H)$ .

One often thinks of the gradient as having a second property: it specifies the “direction” of steepest increase of the function,  $s$ . But:

Without a metric, “steepest” is not defined.

Which is steeper, moving one unit in the temperature direction, or one unit in the humidity direction? In desperation, we might ignore our units of measure, and choose the Euclidean metric (thus equating one unit of temperature with one unit of pressure and one unit of humidity); then the gradient produces a “direction” of steepest increase. However, with no justification for such a choice of metric, the result is probably meaningless.

What about basis vectors? The obvious choice is, including units,  $(1 \text{ K}, 0 \text{ mm-Hg}, 0 \%)^T$ ,  $(0 \text{ K}, 1 \text{ mm-Hg}, 0 \%)^T$ , and  $(0 \text{ K}, 0 \text{ mm-Hg}, 1 \%)^T$ , or omitting units:  $(1, 0, 0)$ ,  $(0, 1, 0)$ , and  $(0, 0, 1)$ . Note that these are *not* unit vectors, because there is no such thing as a “unit” vector, because there is no metric by which to measure one “unit.” Also, if I ascribe units to the basis vectors, then the *components* of an arbitrary vector in that basis are dimensionless.

Now let’s change the basis: suppose now I measure temperature in some unit equal to  $\frac{1}{2}$  K (almost the Rankine scale). Now all my temperature measurements “double”, i.e.  $T_{new} = 2 T_{old}$ . In other words,  $(\frac{1}{2} \text{ K}, 0, 0)^T$  is a different basis than  $(1 \text{ K}, 0, 0)^T$ . As expected for a covariant component, the temperature component of the gradient  $(\nabla s)_T$  is cut in half if the basis vector “halves.” So when the half-size gradient component operates on the double-size temperature vector component, the product remains invariant, i.e., the speed of light is a function of temperature, not of the units in which you measure temperature.

The above basis change was a simple change of scale of one component in isolation. The other common basis change is a “rotation” of the axes, “mixing” the old basis vectors.

Can we rotate axes when the units are different for each component? Surprisingly, we can.



We simply define new basis vectors as linear combinations of old ones, which is all that a rotation does. For example, suppose we measured the speed of light on 3 different days, and the environmental conditions were different on those 3 days. We choose those measurements as our basis, say  $\mathbf{e}_1 = (300 \text{ K}, 750 \text{ mm-Hg}, 20\%)$ ,  $\mathbf{e}_2 = (290 \text{ K}, 760 \text{ mm-Hg}, 30 \%)$ , and  $\mathbf{e}_3 = (290 \text{ K}, 770 \text{ mm-Hg}, 10 \%)$ . These basis vectors are not orthogonal, but are (of course) linearly independent. Suppose I want to know the speed of light at  $(296 \text{ K}, 752 \text{ mm-Hg}, 18 \%)$ . I decompose this into my new basis and get  $(0.6, 0.6, -0.2)$ . I compute the speed of light function in the new basis, and then compute its gradient, to get  $d_1 \tilde{\mathbf{e}}^1 + d_2 \tilde{\mathbf{e}}^2 + d_3 \tilde{\mathbf{e}}^3$ . I then operate on the vector with the gradient to find the change in speed:  $\Delta s = \nabla s(0.6, 0.6, -0.2) = 0.6 d_1 + 0.6 d_2 - 0.2 d_3$ .

We could extend this to a more complex function, and then the gradient is not constant. For example, a more accurate equation for the speed of light is

$$s(T, P, H) = c_0 - f \frac{P}{T} + gH \left( (T - 273)^2 + 160 \right)$$

where  $f \approx 7.86 \times 10^{-4}$  and  $g \approx 1.5 \times 10^{-11}$  are constants. Now the gradient is a function of position (in TPH space), and there is still no metric.

Comment on the metric: In desperation, you might define a metric, i.e. the length of a vector, to be  $\Delta s$ , the change in the speed of light due to the environmental changes defined by that vector. However, such a metric is in general non-Euclidean (not a Pythagorean relationship), indefinite (non-zero vectors can have zero or negative “lengths”), and still doesn’t define a meaningful dot product. Our more-accurate equation for the speed of light provides examples of these failures.

**References:**

- [Knu] Knuth, Donald, *The Art of Computer Programming, Vol. 2: Seminumerical Algorithms*, 2nd Ed., p. 117.
- [Mic] Michelsen, Eric L., *Funky Quantum Concepts*, unpublished.  
<http://physics.ucsd.edu/~emichels/FunkyQuantumConcepts.pdf>.
- [Sch] Schutz, Bernard, *A First Course in General Relativity*, Cambridge University Press, 1990.
- [Sch2] Schutz, Bernard, *Geometrical Methods of Mathematical Physics*, Cambridge University Press, 1980.
- [Tal] Talman, Richard, *Geometric Mechanics*, John Wiley and Sons, 2000.

# 11 Differential Geometry

## Manifolds

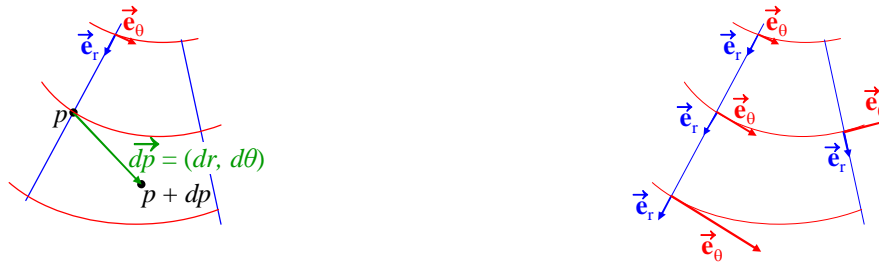
A **manifold** is a “space”: a set of points with coordinate labels. We are free to choose coordinates many ways, but a manifold must be able to have coordinates that are real numbers. We are familiar with “metric manifolds”, where there is a concept of distance. However, there are many useful manifolds which have no metric, e.g. phase space (see “We Don’t Need No Stinking Metric” above).

Even when a space is non-metric, it still has concepts of “locality” and “continuity.”

Such locality and continuity are defined in terms of the coordinates, which are real numbers. It may also have a “volume”, e.g. the oft-mentioned “phase-space volume.” It may seem odd that there’s no definition of “distance,” but there is one of “volume.” **Volume** in this case is simply defined in terms of the coordinates,  $dV = dx_1 dx_2 dx_3 \dots$ , and has no absolute meaning.

## Coordinate Bases

Coordinate bases are basis vectors derived from coordinates on the manifold. They are extremely useful, and built directly on basic multivariate calculus. Coordinate bases can be defined a few different ways. Perhaps the simplest comes from considering a small displacement vector on a manifold. We use 2D polar coordinates in  $(r, \theta)$  as our example. A **coordinate basis** can be defined as the basis in which the components of an infinitesimal displacement vector are just the differentials of the coordinates:



(Left) Coordinate bases: the components of the displacement vector are the differentials of the coordinates. (Right) Coordinate basis vectors around the manifold.

Note that  $e_\theta$  (the  $\theta$  basis vector) far from the origin must be bigger than near, because a small change in angle,  $d\theta$ , causes a bigger displacement vector far from the origin than near. The advantage of a coordinate basis is that it makes dot products, such as a gradient dotted into a displacement, appear in the simplest possible form:

$$\text{Given } f(r, \theta), \quad df = \nabla f(r, \theta) \cdot d\mathbf{p} = \left( \frac{\partial f}{\partial r} + \frac{\partial f}{\partial \theta} \right) \cdot (dr, d\theta) = \frac{\partial f}{\partial r} dr + \frac{\partial f}{\partial \theta} d\theta$$

The last equality is assured from elementary multivariate calculus.

The basis vectors are defined by differentials, but are themselves finite vectors. Any physical vector, finite or infinitesimal, can be expressed in the coordinate basis, e.g., velocity, which is finite.

**“Vectors” as derivatives:** There is a huge confusion about writing basis “vectors” as derivatives. From our study of tensors (earlier), we know that a vector can be considered an operators on a 1-form, which produces a scalar. We now describe how vector fields can be considered operators on scalar functions, which produce scalar fields. I don’t like this view, since it is fairly arbitrary, confuses the much more consistent tensor view, and is easily replaced with tensor notation.

We will see that in fact, the derivative “basis vectors” are operators which create 1-forms (dual-basis components), not traditional basis vectors. The vector basis is then *implicitly* defined as the dual of the dual-basis, which is always the coordinate basis. In detail:



We know from the “Tensors” chapter that the gradient of a scalar field is a 1-form with partial derivatives as its components. For example:

$$\nabla f(x, y, z) = \left( \frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z} \right) = \frac{\partial f}{\partial x} \omega^1 + \frac{\partial f}{\partial y} \omega^2 + \frac{\partial f}{\partial z} \omega^3, \quad \text{where } \omega^1, \omega^2, \omega^3 \text{ are basis 1-forms}$$

Many texts *define* vectors in terms of their action on scalar functions (aka scalar fields), e.g. [Wald p15]. Given a point  $(x, y, z)$ , and a function  $f(x, y, z)$ , the definition of a vector  $\mathbf{v}$  amounts to

$$\mathbf{v} \equiv (v^x, v^y, v^z) \quad \text{such that} \quad \mathbf{v}[f(x, y, z)] \equiv \mathbf{v} \cdot \nabla f = v^x \frac{\partial f}{\partial x} + v^y \frac{\partial f}{\partial y} + v^z \frac{\partial f}{\partial z} \quad (\text{a scalar field})$$

Roughly, the action of  $\mathbf{v}$  on  $f$  produces a scaled directional derivative of  $f$ : Given some small displacement  $dt$ , as a fraction of  $|\mathbf{v}|$  and in the direction of  $\mathbf{v}$ ,  $\mathbf{v}$  tells you how much  $f$  will change when moving from  $(x, y, z)$  to  $(x + v^x dt, y + v^y dt, z + v^z dt)$ :

$$df = \mathbf{v}[f] dt \quad \text{or} \quad \frac{df}{dt} = \mathbf{v}[f]$$

If  $t$  is time, and  $\mathbf{v}$  is a velocity, then  $\mathbf{v}[f]$  is the time rate of change of  $f$ . While this notation is compact, I'd rather write it simply as the dot product of  $\mathbf{v}$  and  $\nabla f$ , which is more explicit, and consistent with tensors:

$$df = \mathbf{v} \cdot \nabla f dt \quad \text{or} \quad \frac{df}{dt} = \mathbf{v} \cdot \nabla f$$

The definition of  $\mathbf{v}$  above requires an auxiliary function  $f$ , which is messy. We remove  $f$  by redefining  $\mathbf{v}$  as an operator:

$$\mathbf{v} \equiv \left( v^x \frac{\partial}{\partial x} + v^y \frac{\partial}{\partial y} + v^z \frac{\partial}{\partial z} \right) \quad (\text{an operator})$$

Given this form, it *looks like*  $\partial/\partial x$ ,  $\partial/\partial y$ , and  $\partial/\partial z$  are some kind of “basis vectors.” Indeed, standard terminology is to refer to  $\partial/\partial x$ ,  $\partial/\partial y$ , and  $\partial/\partial z$  as the “coordinate basis” for vectors, but they are really operators for creating *1-forms*! Then

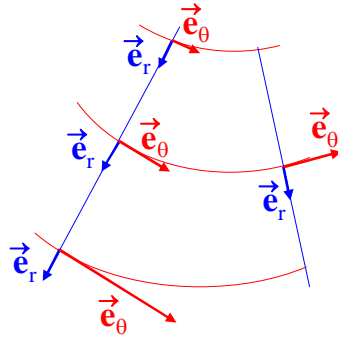
$$\mathbf{v}[f] = v^x \frac{\partial f}{\partial x} + v^y \frac{\partial f}{\partial y} + v^z \frac{\partial f}{\partial z} = \sum_{i=x,y,z} v^i (\nabla f)_i \quad (\text{a scalar field})$$

The vector  $\mathbf{v}$  contracts directly with the 1-form  $\nabla f$  (without need of any metric), hence  $\mathbf{v}$  is a vector *implicitly* defined in the basis dual to the 1-form  $\nabla f$ .

Note that if  $\mathbf{v} = \mathbf{v}(x, y, z)$  is a vector field, then

$$\mathbf{v}[f(x, y, z)] \equiv \mathbf{v}(x, y, z) \cdot \nabla f(x, y, z) \quad (\text{a scalar field})$$

These derivative operators can be drawn as basis vectors in the usual manner, as arrows on the manifold. They are just the coordinate basis vectors shown earlier. For example, consider polar coordinates  $(r, \theta)$ :



Examples of coordinate basis vectors around the manifold.  $\mathbf{e}_r$  happens to be unit magnitude everywhere, but  $\mathbf{e}_\theta$  is not.

The manifold in this case is simply the flat plane,  $\mathbb{R}^2$ . The  $r$ -coordinate basis vectors are all the same size, but have different directions at different places. The  $\theta$  coordinate basis vectors get larger with  $r$ , and also vary in direction around the manifold.

### Covariant Derivatives

Notation: Due to word-processor limitations, the following two notations are equivalent:

$$\vec{h}(\cdot) \equiv \mathbf{h}(\cdot), \quad \vec{r} \equiv \mathbf{r}.$$

This description is similar to one in [Sch].

We start with the familiar concepts of derivatives, and see how that evolves into the covariant derivative. Given a real-valued function of one variable,  $f(x)$ , we want to know how  $f$  varies with  $x$  near a value,  $a$ . The answer is the derivative of  $f(x)$ , where

$$df = f'(a) dx \text{ and therefore } f(a + dx) \approx f(a) + df = f(a) + f'(a) dx$$

Extending to two variables,  $g(x, y)$ , we'd like to know how  $g$  varies in the 2-D neighborhood around a point  $(a, b)$ , given a displacement vector  $d\mathbf{r} = (dx, dy)$ . We can compute its gradient:

$$\vec{\nabla} g = \frac{\partial g}{\partial x} \tilde{d}\mathbf{x} + \frac{\partial g}{\partial y} \tilde{d}\mathbf{y} \quad \text{and therefore} \quad g(a + dx, b + dy) \approx g(a, b) + \vec{\nabla} g(d\vec{r})$$

The gradient is also called a directional derivative, because the rate at which  $g$  changes depends on the direction in which you move away from the point  $(a, b)$ .

The gradient extends to a vector valued function (a vector **field**)  $\mathbf{h}(x, y) = h^x(x, y)\mathbf{i} + h^y(x, y)\mathbf{j}$ :

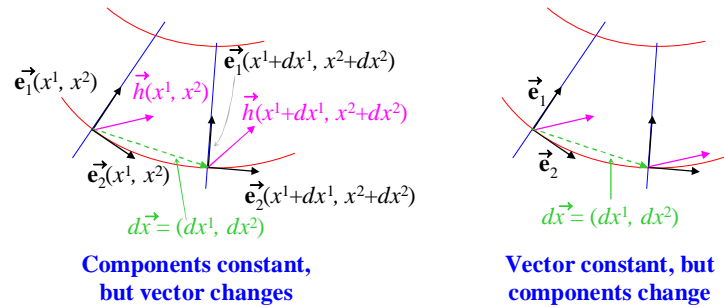
$$\begin{aligned} \vec{\nabla} \vec{h} &= \frac{\partial \vec{h}}{\partial x} \tilde{d}\mathbf{x} + \frac{\partial \vec{h}}{\partial y} \tilde{d}\mathbf{y} \\ \frac{\partial \vec{h}}{\partial x} &= \frac{\partial h^x}{\partial x} \mathbf{i} + \frac{\partial h^y}{\partial x} \mathbf{j} \quad \text{and} \quad \frac{\partial \vec{h}}{\partial y} = \frac{\partial h^x}{\partial y} \mathbf{i} + \frac{\partial h^y}{\partial y} \mathbf{j} \\ d\vec{h} = \vec{\nabla} \vec{h}(d\vec{r}) &= \frac{\partial \vec{h}}{\partial x} dx + \frac{\partial \vec{h}}{\partial y} dy = \begin{bmatrix} \frac{\partial h^x}{\partial x} & \frac{\partial h^x}{\partial y} \\ \frac{\partial h^y}{\partial x} & \frac{\partial h^y}{\partial y} \end{bmatrix} \begin{bmatrix} dx \\ dy \end{bmatrix} = dx \begin{bmatrix} \frac{\partial h^x}{\partial x} \\ \frac{\partial h^y}{\partial x} \end{bmatrix} + dy \begin{bmatrix} \frac{\partial h^x}{\partial y} \\ \frac{\partial h^y}{\partial y} \end{bmatrix} \end{aligned}$$

We see that the columns of  $\nabla \mathbf{h}$  are vectors which are weighted by  $dx$  and  $dy$ , and then summed to produce a vector result. Therefore,  $\nabla \mathbf{h}$  is linear in the displacement vector  $d\mathbf{r} = (dx, dy)$ . This linearity insures that it transforms like a duck . . . I mean, like a tensor. Thus  $\nabla \mathbf{h}$  is a rank-2 ( $^1_1$ ) tensor: it takes a single vector input, and produces a vector result.

So far, all this has been in rectangular coordinates. Now we must consider what happens in curvilinear coordinates, such as polar. Note that we're still in a simple, flat space. (We'll get to curved spaces later). Our goal is still to find the change in the vector value of  $\mathbf{h}(\ )$ , given an infinitesimal vector change of position,  $d\mathbf{x} = (dx^1, dx^2)$ . We use the same approach as above, where a vector valued function comprises two (or  $n$ ) real-valued component functions:  $\vec{h}(x^1, x^2) = h^1(x^1, x^2)\vec{e}_1 + h^2(x^1, x^2)\vec{e}_2$ . However, in this general case, the basis vectors are themselves functions of position (previously the basis vectors were constant everywhere). So  $\mathbf{h}(\ )$  is really

$$\vec{h}(x^1, x^2) = h^1(x^1, x^2)\vec{e}_1(x^1, x^2) + h^2(x^1, x^2)\vec{e}_2(x^1, x^2)$$

Hence, partial derivatives of the *component functions alone* are no longer sufficient to define the change in the vector value of  $\mathbf{h}(\ )$ ; we must also account for the change in the basis vectors.



Note that a component of the derivative is distinctly *not* the same as the derivative of the component (see diagram above). Therefore, the  $i$ th component of the derivative depends on *all* the components of the vector field.

We compute partial derivatives of the vector field  $\mathbf{h}(x^1, x^2)$  using the product rule:

$$\begin{aligned} \frac{\partial \vec{h}}{\partial x^1} &= \frac{\partial h^1}{\partial x^1} \vec{e}_1(x^1, x^2) + h^1(x^1, x^2) \frac{\partial \vec{e}_1}{\partial x^1} + \frac{\partial h^2}{\partial x^1} \vec{e}_2(x^1, x^2) + h^2(x^1, x^2) \frac{\partial \vec{e}_2}{\partial x^1} \\ &= \sum_{j=1}^n \left( \frac{\partial h^j}{\partial x^1} \vec{e}_j(x^1, x^2) + h^j(x^1, x^2) \frac{\partial \vec{e}_j}{\partial x^1} \right) \end{aligned}$$

This is a *vector* equation: all terms are vectors, each with components in all  $n$  basis directions. This is equivalent to  $n$  numerical component equations. Note that  $(\partial \mathbf{h} / \partial x_1)$  has components in both (or all  $n$ ) directions. Of course, we can write similar equations for the components of the derivative in any basis direction,  $\mathbf{e}_k$ :

$$\begin{aligned} \frac{\partial \vec{h}}{\partial x^k} &= \frac{\partial h^1}{\partial x^k} \vec{e}_1(x^1, x^2) + h^1(x^1, x^2) \frac{\partial \vec{e}_1}{\partial x^k} + \frac{\partial h^2}{\partial x^k} \vec{e}_2(x^1, x^2) + h^2(x^1, x^2) \frac{\partial \vec{e}_2}{\partial x^k} \\ &= \sum_{j=1}^n \left( \frac{\partial h^j}{\partial x^k} \vec{e}_j(x^1, x^2) + h^j(x^1, x^2) \frac{\partial \vec{e}_j}{\partial x^k} \right) \end{aligned}$$

Because we must frequently work with components and component equations, rather than whole vector equations, let us now consider only the  $i$ th component of the above:

$$\left(\frac{\partial \bar{h}}{\partial x^k}\right)^i = \frac{\partial h^i}{\partial x^k} + \sum_{j=1}^n h^j(x^1, x^2) \left(\frac{\partial \bar{\mathbf{e}}_j}{\partial x^k}\right)^i$$

The first term moves out of the summation because each of the first terms in the summation of eq. (1) are vectors, and each points exactly in the  $\mathbf{e}_j$  direction. Only the  $j = i$  term contributes to the  $i$ th component; the purely  $\mathbf{e}_j$  directed vector contributes nothing to the  $i$ th component when  $j \neq i$ .

Recall that these equations are true for *any arbitrary* coordinate system; we have made no assumptions about unit length or orthogonal basis vectors. Note that

$$\frac{\partial \bar{h}}{\partial x_k} = (\nabla \bar{h})_k = \text{the } k\text{th (covariant) component of } \nabla \bar{h}$$

Since  $\nabla \bar{h}$  is a rank-2 tensor, the  $k$ th covariant component of  $\nabla \bar{h}$  is the  $k$ th column of  $\nabla \bar{h}$ :

$$\nabla \bar{h} = \begin{bmatrix} \left(\frac{\partial \bar{h}}{\partial x^1}\right)^1 & \left(\frac{\partial \bar{h}}{\partial x^2}\right)^1 \\ \left(\frac{\partial \bar{h}}{\partial x^1}\right)^2 & \left(\frac{\partial \bar{h}}{\partial x^2}\right)^2 \end{bmatrix}$$

Since the change in  $\mathbf{h}(\ )$  is linear with small changes in position,

$$d\bar{h} = \nabla \bar{h}(d\bar{x}), \quad \text{where } d\bar{x} = (dx^1, dx^2)$$

Going back to Equations (1) and (2), we can now write the full covariant derivative of  $\mathbf{h}(\ )$  in 3 ways: vector, verbose component, and compact component:

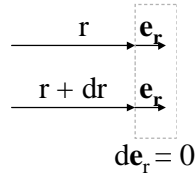
$$\begin{aligned} (\nabla \bar{h})_k &\equiv \nabla_k \bar{h} = \frac{\partial \bar{h}}{\partial x^k} + \sum_{j=1}^n h^j(x^1, x^2) \frac{\partial \bar{\mathbf{e}}_j}{\partial x^k} = \frac{\partial \bar{h}}{\partial x^k} + \sum_{j=1}^n h^j(x^1, x^2) \sum_{i=1}^n \Gamma^i_{jk} \bar{\mathbf{e}}_i \\ (\nabla_k \bar{h})^i &= \frac{\partial h^i}{\partial x^k} + \sum_{j=1}^n h^j(x^1, x^2) \left(\frac{\partial \bar{\mathbf{e}}_j}{\partial x^k}\right)^i \\ (\nabla_k \bar{h})^i &= \frac{\partial h^i}{\partial x^k} + h^j \Gamma^i_{jk}, \quad \text{where } \Gamma^i_{jk} \equiv \left(\frac{\partial \bar{\mathbf{e}}_j}{\partial x^k}\right)^i \Rightarrow \Gamma^i_{jk} \bar{\mathbf{e}}_i = \frac{\partial \bar{\mathbf{e}}_j}{\partial x^k} \end{aligned}$$

Aside: Some mathematicians complain that you can't define the Christoffel symbols as derivatives of basis vectors, because you can't compare vectors from two different points of a manifold without already having the Christoffel symbols (aka the "connection"). Physicists, including Schutz [Sch], say that physics *defines* how to compare vectors at different points of a manifold, and thus you can *calculate* the Christoffel symbols. In the end, it doesn't really matter. Either way, by physics or by fiat, the Christoffel symbols are, in fact, the derivatives of the basis vectors.

---

## Christoffel Symbols

Christoffel symbols are the covariant derivatives of the basis vector fields. TBS.



Derivatives of  $e_r$

### Visualization of n-Forms

- TBS: 1-forms as oriented planes
- 2-forms (in 3 or more space) as oriented parallelograms
- 3-forms (in 3 or more space) as oriented parallelepipeds
- 4-forms (in 4-space): how are they oriented??

### Review of Wedge Products and Exterior Derivative

This is a quick insert that needs proper work. ??

#### 1-D

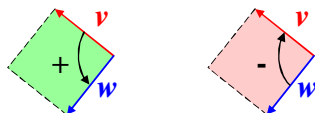
I don't know of any meaning for a wedge-product in 1-D, or even a vector. Also, the 1-D exterior derivative is a degenerate case, because the "exterior" of a line segment is just the 2 endpoints, and all functions are scalar functions. In all higher dimensions, the "exterior" or boundary of a region is a *closed* path/ surface/ volume/ hyper-volume. In 1-D the boundary of a line segment cannot be closed. So instead of integrating around a closed exterior (aka boundary), we simply take the difference in the function value at the endpoints, divided by a differential displacement. This is simply the ordinary derivative of a function,  $f'(x)$ .

#### 2-D

The exterior derivative of a scalar function  $f(x, y)$  follows the 1-D case, and is similarly degenerate, where the "exterior" is simply the two endpoints of a differential displacement. Since the domain is a 2-D space, the displacements are vectors, and there are 2 derivatives, one for displacements in  $x$ , and one for displacements in  $y$ . Hence the exterior derivative is just the one-form "gradient" of the function:

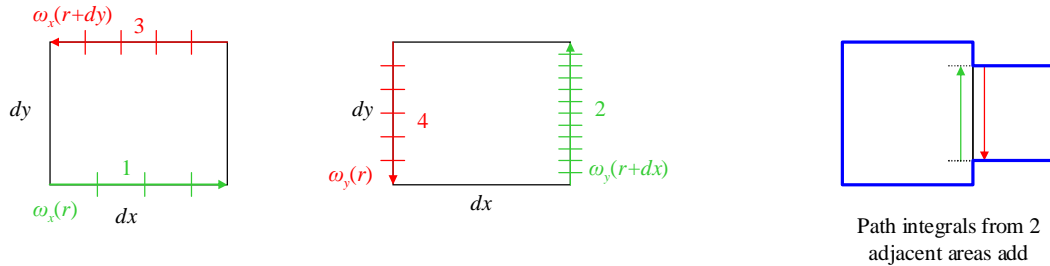
$$\tilde{d}f(x, y) = \text{"gradient"} = \frac{\partial f}{\partial x} \tilde{d}x + \frac{\partial f}{\partial y} \tilde{d}y$$

In 2-D, the **wedge product**  $\tilde{d}x \wedge \tilde{d}y$  is a two-form, which accepts two vectors to produce the signed area of the parallelogram defined by them. A *signed* area can be + or -; a counter-clockwise direction is positive, and clockwise is negative.



$$\begin{aligned} \tilde{\mathbf{d}}\mathbf{x} \wedge \tilde{\mathbf{d}}\mathbf{y}(\vec{v}, \vec{w}) &= \text{signed area defined by } (\vec{v}, \vec{w}) = -\tilde{\mathbf{d}}\mathbf{x} \wedge \tilde{\mathbf{d}}\mathbf{y}(\vec{w}, \vec{v}) \\ &= \tilde{\mathbf{d}}\mathbf{x}(\vec{v})\tilde{\mathbf{d}}\mathbf{y}(\vec{w}) - \tilde{\mathbf{d}}\mathbf{y}(\vec{v})\tilde{\mathbf{d}}\mathbf{x}(\vec{w}) \\ &= \det \begin{vmatrix} \tilde{\mathbf{d}}\mathbf{x}(\vec{v}) & \tilde{\mathbf{d}}\mathbf{x}(\vec{w}) \\ \tilde{\mathbf{d}}\mathbf{y}(\vec{v}) & \tilde{\mathbf{d}}\mathbf{y}(\vec{w}) \end{vmatrix} = \det \begin{vmatrix} v^x & w^x \\ v^y & w^y \end{vmatrix} \end{aligned}$$

The exterior derivative of a 1-form is the ratio of the closed path integral of the 1-form to the area of the parallelogram of two vectors, for infinitesimal vectors. This is very similar to the definition of curl, only applied to a 1-form instead of a vector field.



Consider the horizontal and vertical contributions to the path integral separately:

$$\begin{aligned} \tilde{\omega}(x, y) &= \omega_x(x, y)\tilde{\mathbf{d}}\mathbf{x} + \omega_y(x, y)\tilde{\mathbf{d}}\mathbf{y} & \vec{r} &= (x, y) & d\vec{r} &= (dx, dy) \\ \int_1 \tilde{\omega}(d\vec{r}) + \int_3 \tilde{\omega}(d\vec{r}) &= \omega_x(r)dx - \omega_x(r + dy)dx = -\frac{\partial \omega_x}{\partial y} dy dx \\ \int_2 \tilde{\omega}(d\vec{r}) + \int_4 \tilde{\omega}(d\vec{r}) &= \omega_y(r + dx)dy - \omega_y(r)dy = \frac{\partial \omega_y}{\partial x} dx dy \end{aligned}$$

The horizontal (segments 1 & 3) integrals are linear in  $dx$ , because that is the length of the path. They are linear in  $dy$ , because  $dy$  is proportional to the difference in  $\omega_x$ . Hence, the contribution is linear in both  $dx$  and  $dy$ , and therefore proportional to the area  $(dx)(dy)$ .

A similar argument holds for the vertical contribution, segments 2 & 4. Therefore, the path integral varies proportionately to the area enclosed by two orthogonal vectors.

It is easy to show this is true for any two vectors, and any shaped area bounded by an infinitesimal path. For example, when you butt up two rectangles, the path integral around the combined boundary equals the sum of the individual path integrals, because the contributions from the common segment cancel from each rectangle, and hence omitting them does not change the path integral. The area integrals clearly add.

### 3-D

In 3-D, the wedge product

$$\begin{aligned} \tilde{\mathbf{d}}\mathbf{x} \wedge \tilde{\mathbf{d}}\mathbf{y} \wedge \tilde{\mathbf{d}}\mathbf{z}(\vec{u}, \vec{v}, \vec{w}) &= \text{signed volume defined by } (\vec{u}, \vec{v}, \vec{w}) = -\tilde{\mathbf{d}}\mathbf{x} \wedge \tilde{\mathbf{d}}\mathbf{y} \wedge \tilde{\mathbf{d}}\mathbf{z}(\vec{w}, \vec{v}, \vec{u}), \text{ etc.} \\ &= \det \begin{vmatrix} \tilde{\mathbf{d}}\mathbf{x}(\vec{u}) & \tilde{\mathbf{d}}\mathbf{x}(\vec{v}) & \tilde{\mathbf{d}}\mathbf{x}(\vec{w}) \\ \tilde{\mathbf{d}}\mathbf{y}(\vec{u}) & \tilde{\mathbf{d}}\mathbf{y}(\vec{v}) & \tilde{\mathbf{d}}\mathbf{y}(\vec{w}) \\ \tilde{\mathbf{d}}\mathbf{z}(\vec{u}) & \tilde{\mathbf{d}}\mathbf{z}(\vec{v}) & \tilde{\mathbf{d}}\mathbf{z}(\vec{w}) \end{vmatrix} = \det \begin{vmatrix} u^x & v^x & w^x \\ u^y & v^y & w^y \\ u^z & v^z & w^z \end{vmatrix} \end{aligned}$$

is a 3-form which can either:

1. accept 2 vectors to produce an *oriented* area; it doesn't have a sign, it has a *direction*. Analogous to the cross-product. Or,

2. accept 3 vectors to produce a *signed* volume.

The exterior derivative of a scalar or 1-form field is essentially the same as in the 2-D case, except that now the areas defined by vectors are oriented instead of simply signed. In this case, the “exterior” is a closed surface; the “interior” is a volume.

## 12 Math Tricks

Here are some math “tricks” that either come up a lot and are worth knowing about, or are just fun and interesting.

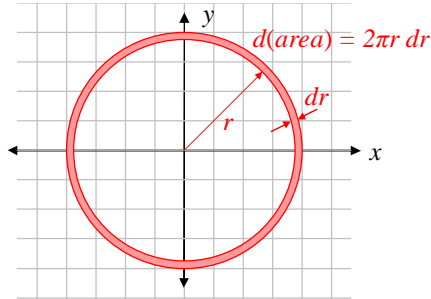
### Math Tricks That Come Up A Lot

#### The Gaussian Integral

$\int_{-\infty}^{\infty} e^{-ax^2} dx$  You can look this up anywhere, but here goes: we’ll evaluate the basic integral,  $\int_{-\infty}^{\infty} e^{-x^2} dx$ , and throw in the ‘a’ at the end by a simple change of variable. First, we square the integral, then rewrite the second factor calling the dummy integration variable y instead of x:

$$\left( \int_{-\infty}^{\infty} dx e^{-x^2} \right)^2 = \left( \int_{-\infty}^{\infty} dx e^{-x^2} \right) \left( \int_{-\infty}^{\infty} dy e^{-y^2} \right) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy e^{-(x^2+y^2)}$$

This is just a double integral over the entire x-y plane, so we can switch to polar coordinates. Note that the exponential integrand is constant at constant r, so we can replace the differential area dx dy with 2πr dr:



$$\begin{aligned} \text{Let } r^2 = x^2 + y^2 \quad \Rightarrow \quad \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy e^{-(x^2+y^2)} &= \int_0^{\infty} dr \, 2\pi r e^{-r^2} \\ &= -\pi \left[ e^{-r^2} \right]_0^{\infty} = \pi \end{aligned}$$

$$\left( \int_{-\infty}^{\infty} dx e^{-x^2} \right)^2 = \pi \quad \Rightarrow \quad \int_{-\infty}^{\infty} dx e^{-x^2} = \sqrt{\pi}, \quad \text{and} \quad \boxed{\int_{-\infty}^{\infty} dx e^{-ax^2} = \sqrt{\frac{\pi}{a}}}$$

### Math Tricks That Are Fun and Interesting

$$\int \frac{dx}{\sin x}$$

#### Continuous Infinite Crossings

The following function has an infinite number of zero crossings near the origin, but is everywhere continuous (even at x = 0). That seems bizarre to me. Recall the definition:

$$f(x) \text{ is } \mathbf{continuous} \text{ at } a \text{ iff } \lim_{x \rightarrow a} f(x) = f(a)$$

Then let



$$f(x) = \begin{cases} x \sin\left(\frac{1}{x}\right), & x \neq 0 \\ 0, & x = 0 \end{cases}$$

$$\lim_{x \rightarrow 0} f(x) = 0 = f(0) \quad f(0) \text{ is continuous}$$

Picture

---

## Phasors

Phasors are complex numbers that represent sinusoids. The phasor defines the magnitude and phase of the sinusoid, but not its frequency. See *Funky Electromagnetic Concepts* for a full description.

---

## Future Funky Mathematical Physics Topics

1. Finish theoretical importance of IBP
2. Finish Legendre transformations
3. Sturm-Liouville
4. Pseudo-tensors (ref. Jackson).
5. Tensor densities
6.  $f(z) = \int_{-\infty}^{\infty} dx \exp(-x^2)/x-z$  has no poles, but has a branch cut. Where is the branch cut, and what is the change in  $f(z)$  across it?

## 13 Appendices

---

### References

- [A&S] Abramowitz and Stegun, ??
- [Chu] Churchill, Ruel V., Brown, James W., and Verhey, Roger F., *Complex Variables and Applications*, 1974, McGraw-Hill. ISBN 0-07-010855-2.
- [Det] Dettman, John W., *Applied Complex Variables*, 1965, Dover. ISBN 0-486-64670-X.
- [F&W] Fetter, Alexander L. and John Dirk Walecka, *Theoretical Mechanics for Particles and Continua*, McGraw-Hill Companies, February 1, 1980. ISBN-13: 978-0070206588.
- [Jac] Jackson, *Classical Electrodynamics*, 3<sup>rd</sup> ed.
- [M&T] Marion & Thornton, 4th ed.
- [One] O’Neill, Barrett, *Elementary Differential Geometry*, 2<sup>nd</sup> ed., 1997, Academic Press. ISBN 0-12-526745-2.
- [Sch] Schutz, Bernard F., *A First Course in General Relativity*, Cambridge University Press (January 31, 1985), ISBN 0521277035.
- [Sch2] Schutz, Bernard F., *Geometrical Methods of Mathematical Physics*, Cambridge University Press ??, ISBN
- [Schwa 1998] Schwarzenberg-Czerny, A., “The distribution of empirical periodograms: Lomb–Scargle and PDM spectra,” *Monthly Notices of the Royal Astronomical Society*, vol 301, p831–840 (1998).
- [Sea] Sean, *Sean’s Applied Math Book*, 1/24/2004.  
<http://www.its.caltech.edu/~sean/book.html>.
- [Tal] Talman, Richard, *Geometric Mechanics*, Wiley-Interscience; 1<sup>st</sup> edition (October 4, 1999), ISBN 0471157384
- [Tay] Taylor, Angus E., *General Theory of Functions and Integration*, 1985, Dover. ISBN 0-486-64988-1.
- [W&M] Walpole, Ronald E. and Raymond H. Myers, *Probability and Statistics for Engineers and Scientists*, 3<sup>rd</sup> edition, 1985, Macmillan Publishing Company, ISBN 0-02-424170-9.
- [Wyl] Wyld, H. W., *Mathematical Methods for Physics*, 1999, Perseus Books Publishing, LLC, ISBN 0-7382-0125-1.

---

### Glossary

Definitions of common mathematical physics terms. “Special” mathematical definitions are noted by “(math)”. These are technical mathematical terms that you shouldn’t have to know, but will make reading math books a lot easier because they are very common. These definitions try to be conceptual and accurate, but comprehensible to “normal” people (including physicists, but not mathematicians).

- 1-1 A mapping from a set A to a set B is 1-1 if every value of B under the map has only one value of A that maps to it. In other words, given the value of B under the map, we can uniquely find the value of A which maps to it. However, see “1-1 correspondence.” See also “injection.”
- 1-1 correspondence A mapping, between two sets A and B, is a 1-1 correspondence if it uniquely associates each value of A with a value of B, and each value of B with a value of A. Synonym: bijection.
- accumulation point syn. for limit point.

adjoint <sup>1</sup>	The adjoint of an operator produces a bra from a bra in the same way the original operator produces a ket from a ket: $\hat{O} \psi\rangle =  \phi\rangle \Rightarrow \langle\psi \hat{O}^\dagger = \langle\phi , \forall  \psi\rangle$ . The adjoint of an operator is the operator which preserves the inner product of two vectors as $\langle\mathbf{v}  \cdot (\mathbf{O} \mathbf{w}\rangle) = (\mathbf{O}^\dagger \mathbf{v}\rangle)^\dagger \cdot  \mathbf{w}\rangle$ . The adjoint of an operator matrix is the conjugate-transpose. This has nothing to do with matrix adjoints (below).
adjoint <sup>2</sup>	In matrices, the transpose of the cofactor matrix is called the adjoint of a matrix. This has nothing to do with linear operator adjoints (above).
adjugate	the transpose of the cofactor matrix: $\text{adj}(\mathbf{A})_{ij} = C_{ji} = (-1)^{i+j} M_{ji}$ , where $M_{ji}$ is the transpose of the minor matrix: $M_{ij} = \det(\mathbf{A} \text{ deleting row } i \text{ and column } j)$ .
analytic	A function is analytic in some domain iff it has continuous derivatives to all orders, i.e. is infinitely differentiable. For complex functions of complex variables, if a function has a continuous first derivative in some region, then it has continuous derivatives to all orders, and is therefore analytic.
analytic geometry	the use of coordinate systems along with algebra and calculus to study geometry. Aka “coordinate geometry”
bijection	Both an “injection” and a “surjection,” i.e. 1-1 and “onto.” A mapping between sets A and B is a bijection iff it uniquely associates a value of A with every value of B. Synonym: 1-1 correspondence.
BLUE	In statistics, Best Linear Unbiased Estimator.
branch point	A branch point is a point in the domain of a complex function $f(z)$ , $z$ also complex, with this property: when $z$ traverses a closed path around the branch point, following continuous values of $f(z)$ , $f(z)$ has a different value at the end of the path than at the beginning, even though the beginning and end point are the same point in the domain. Example TBS: square root around the origin.
boundary point	(math) see “limit point.”
C or $\mathbb{C}$	the set of complex numbers.
closed	(math) contains any limit points. For finite regions, a closed region includes its boundary. Note that in math talk, a set can be both open and closed! The surface of a sphere is open (every point has a neighborhood in the surface), and closed (no excluded limit points; in fact, no limit points).
cofactor	The $ij$ -th <b>minor</b> of an $n \times n$ matrix is the determinant of the $(n-1) \times (n-1)$ matrix formed by crossing out the $i$ -th row and $j$ -th column. A <b>cofactor</b> is just a minor with a plus or minus sign affixed, according to whether $(i, j)$ is an even or odd number of steps away from $(1, 1)$ : $C_{ij} = (-1)^{i+j} M_{ij}$
compact	(math) for our purposes, closed and bounded [Tay thm 2-6I p66]. A compact region may comprise multiple (infinite number??) disjoint closed and bounded regions.
congruence	a set of 1D non-intersecting curves that cover every point of a manifold. Equivalently, a foliation of a manifold with 1D curves. Compare to “foliation.”
contrapositive	The contrapositive of the statement “If A then B” is “If not B then not A.” The contrapositive is equivalent to the statement: if the statement is true (or false), the contrapositive is true (or false). If the contrapositive is true (or false), the statement is true (or false).
convergent	approaches a definite limit
converse	The converse of the statement “If A then B” is “If B then A”. In general, if a statement is true, its converse may be either true or false. The converse is the contrapositive of the inverse, and hence the converse and inverse are equivalent statements.

connected	There exists a continuous path between any two points in the set (region). See also: simply connected. [One p178].
coordinate geometry	the use of coordinate systems along with algebra and calculus to study geometry. Aka “analytic geometry”
diffeomorphism	a $C^\infty$ (1-1) map, with a $C^\infty$ inverse, from one manifold <i>onto</i> another. “Onto” implies the mapping covers the whole range manifold. Two diffeomorphic manifolds are topologically identical, but may have different geometries.
divergent	not convergent: a sequence is divergent iff it is not convergent.
domain	of a function: the set of numbers (usually real or complex) on which the function is defined.
entire	A complex function is entire iff it is analytic over the entire complex plane. An entire function is also called an “integral function.”
essential singularity	a “pole” of infinite order, i.e. a singularity around which the function is unbounded, and cannot be made finite by multiplication by any power of $(z - z_0)$ [Det p165].
factor	a number (or more general object) that is <i>multiplied</i> with others. E.g., in “ $(a + b)(x + y)$ ”, there are two factors: “ $(a + b)$ ”, and “ $(x + y)$ ”.
finite	a non-zero number. In other words, not zero, and not infinity.
foliation	a set of non-intersecting submanifolds that cover every point of a manifold. E.g., 3D real space can be foliated into 2D “sheets stacked on top of each other,” or 1D curves packed around each other. Compare to “congruence.”
holomorphic	syn. for analytic. Other synonyms are regular, and differentiable. Also, a “holomorphic map” is just an analytic function.
homomorphic	something from abstract categories that should not be confused with homeomorphism.
homeomorphism	a continuous (1-1) map, with a continuous inverse, from one manifold <i>onto</i> another. “Onto” implies the mapping covers the whole range manifold. A homeomorphism that preserves distance is an <b>isometry</b> .
identify	to establish a 1-1 and <i>onto</i> relationship. If we identify two mathematical things, they are essentially the same thing.
iff	if, and only if,
injection	A mapping from a set A to a set B is an injection if it is 1-1, that is, if given a value of B in the mapping, we can uniquely find the value of A which maps to it. Note that every value of A is included by the definition of “mapping” [CRC 30 <sup>th</sup> ]. The mapping does <i>not</i> have to cover all the elements of B.
integral function	Syn. for “entire function:” a function that is analytic over the entire complex plane.
inverse	The inverse of the statement “If A then B” is “If not A then not B.” In general, if a statement is true, its inverse may be either true or false. The inverse is the contrapositive of the converse, and hence the converse and inverse are equivalent statements.
invertible	A map (or function) from a set A to a set B is invertible iff for every value in B, there exists a unique value in A which maps to it. In other words, a map is invertible iff it is a bijection.
isolated singularity	a singularity at a point, which has a surrounding neighborhood of analyticity [Det p165].
isometry	a homeomorphism that preserves distance, i.e. a continuous, invertible (1-1) map from one manifold <i>onto</i> another that preserves distance (“onto” in the mathematical sense).

isomorphic	“same structure.” A widely used general term, with no single precise definition.
limit point	of a domain is a boundary of a region of the domain: for example, the open interval (0, 1) on the number line and the closed interval [0, 1] both have limit points of 0 and 1. In this case, the open interval excludes its limit points; the closed interval includes them (definition of “closed”). Some definitions define all points in the domain as also limit points. Formally, a point $p$ is a limit point of domain $D$ iff every open subset containing $p$ also contains a point in $D$ other than $p$ .
mapping	syn. “function.” A mapping from a set $A$ to a set $B$ defines a value of $B$ for <i>every</i> value of $A$ [CRC 30 <sup>th</sup> ].
meromorphic	A function is meromorphic on a domain iff it is analytic except at a set of isolated <i>poles</i> of finite order (i.e., non-essential poles). Note that branch points are nonanalytic points, but some are not poles (such as $\sqrt{z}$ at zero), so a function including such a branch point is <i>not</i> meromorphic.
minor	The $ij$ -th minor of an $n \times n$ matrix is the determinant of the $(n-1) \times (n-1)$ matrix formed by crossing out the $i$ -th row and $j$ -th column, i.e., the minor matrix: $M_{ij} = \det(\mathbf{A} \text{ deleting row } i \text{ and column } j)$ . See also “cofactor.”
$\mathbb{N}$	the set of natural numbers (positive integers).
noise	<i>unpredictable</i> variations in a quantity.
oblique	non-orthogonal and not parallel.
one-to-one	see “1-1”.
onto	covering every possible value. A mapping from a set $A$ <i>onto</i> the set $B$ covers every possible value of $B$ , i.e. the mapping is a surjection.
open	(math) An region is open iff every point in the region has a finite neighborhood of points around it that are also all in the region. In other words, every point is an interior point. Note that open is <i>not</i> “not closed;” a region can be both open and closed.
pole	a singularity near which a function is unbounded, but which becomes finite by multiplication by $(z - z_0)^k$ for some finite $k$ [Det p165]. The value $k$ is called the <b>order</b> of the pole.
positive definite	a matrix or operator which is $> 0$ for all <i>non-zero</i> operands. It may be 0 when acting on a “zero” operand, such as the zero vector. This implies that all eigenvalues $> 0$ .
positive semidefinite	a matrix or operator which is $\geq 0$ for all <i>non-zero</i> operands. It may be 0 when acting on a non-zero operands. This implies that all eigenvalues $\geq 0$ .
predictor	in regression: a variable put into a model to predict another value, e.g. $y_{\text{mod}}(x_1, x_2)$ is a model (function) of the predictors $x_1$ and $x_2$ .
PT	perturbation theory.
$\mathbf{Q}$ or $\mathbb{Q}$	the set of rational numbers. $\mathbf{Q}^+ \equiv$ the set of positive rationals.
$\mathbf{R}$ or $\mathbb{R}$	the set of real numbers.
RMS	root-mean-square.
RV	random variable.
removable singularity	an isolated singularity that can be made analytic by simply defining a value for the function at that point. For example, $f(x) = \sin(x)/x$ has a singularity at $x = 0$ . You can remove it by defining $f(0) = 1$ . Then $f$ is everywhere analytic. [Det p165]
residue	The residue of a complex function at a complex point $z_0$ is the $a_{-1}$ coefficient of the Laurent expansion about the point $z_0$ .

- simply connected There are no holes in the set (region), not even point holes. I.e., you can shrink any closed curve in the region down to a point, the curve staying always within the region (including at the point).
- singularity of a function: a point on a boundary (i.e. a limit point) of the domain of analyticity, but where the function is not analytic. [Det def 4.5.2 p156]. Note that the function may be defined at the singularity, but it is not analytic there. E.g.,  $\sqrt{z}$  is continuous at 0, but not differentiable.
- smooth for most references, “smooth” means infinitely differentiable, i.e.  $C^\infty$ . For some, though, “smooth” means at least one continuous derivative, i.e.  $C^1$ , meaning first derivative continuous. This latter definition looks “smooth” to our eye (no kinks, or sharp points).
- surjection A mapping from a set A “onto” the set B, i.e. that covers every possible value of B. Note that every value of A is included by the definition of “mapping” [CRC 30<sup>th</sup>], however multiple values of A may map to the same value of B.
- term a number (or more general object) that is *added* to others. E.g., in “ax + by + cz”, there are three terms: “ax”, “by”, and “cz”.
- trace the trace of a square matrix is the sum of its diagonal elements.
- uniform convergence a series of functions  $f_n(z)$  is uniformly convergent in an open (or partly open) region iff its error  $\epsilon$  after the  $N^{\text{th}}$  function can be made arbitrarily small with a single value of  $N$  (dependent *only* on  $\epsilon$ ) for every point in the region. I.e. given  $\epsilon$ , a single  $N$  works for all points  $z$  in the region [Chu p156].
- voila French for “see there!”
- WLOG or WOLOG without loss of generality
- $\mathbf{Z}$  or  $\mathbb{Z}$  the set of integers.  $\mathbf{Z}^+$  or  $\mathbb{N} \equiv$  the set of positive integers (natural numbers).

**Formulas**

completing the square:  $ax^2 + bx = a\left(x + \frac{b}{2a}\right)^2 - \frac{b^2}{4a}$  (x-shift =  $-b/2a$ )

**Integrals**

$$\int_{-\infty}^{\infty} dx e^{-ax^2} = \sqrt{\frac{\pi}{a}} \qquad \int_{-\infty}^{\infty} dx x^2 e^{-ax^2} = \frac{1}{2} \sqrt{\frac{\pi}{a^3}} \qquad \int_0^{\infty} dr r^3 e^{-ar^2} = \frac{1}{2a^2}$$

**Statistical distributions**

$\chi^2_v$ :  $avg = \nu$        $\sigma^2 = 2\nu$   
 exponential:  $avg = \tau$        $\sigma^2 = \tau^2$

error function [A&S]:  $erf(x) \equiv \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt$

gaussian included probability between  $-z$  and  $+z$ :

$$P_{gaussian}(z) \equiv \int_{-z}^{+z} pdf_{gaussian}(u) du = \frac{1}{\sqrt{2\pi}} \int_{-z}^{+z} e^{-u^2/2} du \qquad \text{Let } u^2/2 \equiv t^2, du = \sqrt{2}dt$$

$$= \frac{2}{\sqrt{2\pi}} \int_0^{+z/\sqrt{2}} e^{-t^2} \sqrt{2}dt = erf\left(z/\sqrt{2}\right)$$

**Special Functions**

$$\Gamma(n) = (n-1)! \quad \Gamma(a) \equiv \int_0^\infty dx x^{a-1} e^{-x} \quad \Gamma(a) = (a-1)\Gamma(a-1) \quad \Gamma(1/2) = \sqrt{\pi}$$

The functions below use the Condon-Shortley phase:

$$Y_{lm}(\theta, \phi) \equiv \begin{cases} (-1)^m \sqrt{\frac{(2l+1)(l-m)!}{2(l+m)!}} P_m(\cos\theta) \frac{e^{im\phi}}{\sqrt{2\pi}}, & m \geq 0, \\ \sqrt{\frac{(2l+1)(l-|m|)!}{2(l+|m|)!}} P_{l|m|}(\cos\theta) \frac{e^{im\phi}}{\sqrt{2\pi}}, & m < 0, \end{cases}$$

$P_m(x)$  is the associated Legendre function,

$$l = 0, 1, 2, \dots, \quad m = -l, -l+1, \dots, l-1, l. \quad [\text{Wyl 3.6.5 p96}]$$

**Index**

The index is not yet developed, so go to the web page on the front cover, and text-search in this document.