RBOT101: Mathematical Foundations of Robotics

Instructor: Dr. Lekan Molu

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

PREAMBLE

Consider this the roadmap for this course. Please read through the syllabus posted on Moodle2 carefully and feel free to share any questions that you may have. Please print a copy of the Syllabus for reference. Some relevant parts of the Syllabus are repeated here but the Moodles reference should serve as your guide throughout the ten weeks of this course.

1.1 Course Description

This course focuses on the algorithmic and mathematical concepts with respect classical and recent methods for solving real-world problems in robotics. While some students may have encountered some of the concepts we will be treating in past courses or avenues of study, we will provide the breadth and depth necessary for equipping students to be world-class roboticists. The topics covered by this course shall include the configuration space, rigid bodies, semi-rigid soft bodies, as well as their motions in $\mathbb{R}^n$, wrenches, homogeneous transformations, optimal algorithms for rigid body rotations, linear systems theory, probability theory, the Kalman filter. The course will begin and end with a self-assessment to allow students to gauge their strengths and weaknesses in these topics. References for further, in-depth study in each topic are provided at the end of this course.

1.2 Course Outcomes

After taking this course, each student will be able to

- Develop mathematical tools for solving fundamental kinematic problems in robot operation;
- Formulate optimal state estimation tools for solving real-time smoothing and filtering operations in robotics;
- Integrate state estimation with rigid and semi-rigid soft bodies to solve real-world automation problems; and 4. Use open-source Python, and C++ tools to solve classical and emerging problems in robotics in our day.

1.3 Prerequisites

An undergraduate-level understanding of linear algebra, analytical mechanics, Python and C++ programming.

1.4 Recommended Texts

- Main Texts

- Simon, Dan. (2007). Optimal state estimation: Kalman, $H - \infty$, and nonlinear approaches. Choice Reviews Online, Vol. 44, pp. 44-3334-44–3334. https://doi.org/10.5860/choice.44-3334
- Murray, R. M., Li, Z., and Sastry, S. S. (1994). A Mathematical Introduction to Robotic Manipulation. Book (Vol. 29). Free PDF preprint downloadable from, Murray's website.
- Theory of Screws: A Study in the Dynamics of a Rigid Body by Robert Stawell Ball, Dublin: Hodges, Foster, and Co., Grafton-Street. a. Textbooks:

- Secondary Text

  - Modern Robotics: Mechanics, Planning, and Control. Free PDF preprint downloadable from Author's Northwestern University Website.

- Auxiliary Text:

  - Theory of Screws: A Study in the Dynamics of a Rigid Body by Robert Stawell Ball, Dublin: Hodges, Foster, and Co., Grafton-Street (Should be downloadable via Interlibrary Loan).

1.5 Recommended Journals

- IEEE Transactions on Robotics.
- The International Journal of Robotics Research.
- The IEEE International Conference on Robotics and Automation (ICRA).
- IEEE/Robotics Society of Japan International Conference on Intelligent Robots and Systems (IROS).
- Robotics and Autonomous Systems, An Elsevier Journal.

1.6 Required Software

- A working knowledge of python and the anaconda environment.
- ROS 1.x Installation Instructions: ros 1.x website.
- ROS 2 installation ros 2.0 website.

1.7 Online Course Content

This course will be conducted completely online using Brandeis' LATTE site. The site contains the course syllabus, assignments, our discussion forums, links/resources to course-related professional organizations and sites, and weekly checklists, objectives, outcomes, topic notes, self-tests, and discussion questions. Access information is emailed to enrolled participants before the start of the course. To begin participating in the course, review the "Welcoming Message" and the "Week 1 Checklist."

1.8 Errata

If in the course of using these notes, you find sentence errors, errata or mistakes in equations, please annotate them and upload it to the discussion forum. Points will awarded, at the discretion of the instructor, for such help.

CHAPTER 2

INTRODUCTION TO MATRIX ANALYSIS.

Our goal here is to introduce the student to the study of matrix theory. Matrices are symbolism of the important transformations in everyday life; these transformations lie at the heart of mathematics and robotics. The contents of this topic are thus positioned toward the aspiration of roboticists, engineers of all stripes and scientists. Specifically, we are concerned with the *theory of symmetric matrices*, which is important for all fields, *matrices and differential equations*, necessary for engineering and robotics, as well as *positive matrices*, necessary for probability theory. Most of the texts im this chapter are drawn from Richard Bellman's Matrix Analysis Book given in the Syllabus.

2.1 Maximization and Minimization

Of importance to us in this section is to ascertain the range of values of *homogeneous quadratic functions* of two variables and how it is connected to the determination of the maximum or minimum of a general function of two variables.

2.1.1 Maximization of Functions of a Variable

Suppose $f(x)$ is a real function of the real variable x for $x \in [a, b]$, and let us suppose that it is a Taylor series of the form

$$f(x) = f(c) + f'(x - c) + f'' \frac{(x - c)^2}{2!} + \dots \tag{2.1.1}$$

around every point in the open interval (a, b). We define a *stationary point* of $f(x)$ to be a point where $f'(x) = 0$ and it is the point that determines if c is a point at which $f(x)$ is a relative maximum, a relative minimum, or a stationary point of a subtle characteristic. If c is a stationary point, we must have

$$f(x) = f(c) + f'' \frac{(x - c)^2}{2!} + \dots \tag{2.1.2}$$

If $f''(c) > 0$, then $f(x)$ has a relative minimum at $x = c$. Otherwise, if $f''(c) < 0$, $f(x)$ has a relative maximum at $x = c$. Whereas, if $f''(c) = 0$, we must needs consider further terms in the expansion.

Quiz 1. Suppose that $f''(c) = 0$, what are the sufficient conditions that c must furnish to be a relative minimum?

2.1.2 Maximization of Functions of Two Variables

Now, suppose that we have two variables x, y as arguments of a function f, defined over the rectangle $a_1 \leq x \leq b_1$, $a_2 \leq y \leq b_2$, and possessing a convergent Taylor series around each point

(c_1, c_2) within the region. Then, for sufficiently small $|x - c_1$ and $|y - c_2|$, we have

$$f(x,y) = f(c_1,c_2) + (x-c_1)\frac{\partial f}{\partial c_1} + (y-c_2)\frac{\partial f}{\partial c_2} + \frac{(x-c_1)^2}{2}\frac{\partial^2 f}{\partial c_1^2} + (x-c_1)(y-c_2)\frac{\partial^2 f}{\partial c_1 \partial c_2} + \frac{(y-c_2)^2}{2}\frac{\partial^2 f}{\partial c_2^2} + \dots \tag{2.1.3}$$

where

$$\begin{aligned} \frac{\partial f}{\partial c_1} &= \frac{\partial f}{\partial x} \text{ at } x = c_1, \qquad y = c_2 \\ \frac{\partial f}{\partial c_2} &= \frac{\partial f}{\partial y} \text{ at } x = c_1, \qquad y = c_2 \text{ e.t.c.} \end{aligned} \tag{2.1.4}$$

As before, the stationary point of $f(x, y)$ is defined to be (c_1, c_2) so that $\frac{\partial f}{\partial c_1} = 0$ and $\frac{\partial f}{\partial c_2} = 0$; and the behavior of $f(x, y)$ in the immediate neighborhood of (c_1, c_2) depends on the nature of the quadratic terms in the expansion of (2.1.3),

$$Q_2(x,y) = a(x-c_1)^2 + 2b(x-c_1)(y-c_2) + c(y-c_2)^2 \tag{2.1.5}$$

where $a = \frac{1}{2}\frac{\partial^2 f}{\partial c_1^2}$, $2b = \frac{\partial^2 f}{\partial c_1 \partial c_2}$, and $c = \frac{1}{2}\frac{\partial^2 f}{\partial c_2^2}$.

Suppose we set $x - c_1 = u$ and $y - c_2 = v$, then we can write a quadratic expression in variables u and v $i.e.$

$$Q(u,v) = au^2 + 2buv + cv^2 \tag{2.1.6}$$

whereupon we are interested in the behavior of $Q(u, v)$ in the vicinity of $u = v = 0$ and the fact that $Q(u, v)$ is homogeneous allows us to examine the range of values of $Q(u, v)$ for the set of values on $u^2 + v^2 = 1$.

If $Q(u, v) > 0$ for all u and v distinct from $u = v = 0$, $f(x, y)$ will have a relative minimum at $x = c_1$, $y = c_2$; and if $Q(u, v) < 0$ for all u and v distinct from $u = v = 0$, $f(x, y)$ will have a relative maximum at $x = c_1$, $y = c_2$; The stationary point is a *saddle point* if $Q(u, v)$ can take on both positive and negative values.

2.1.3 Algebraic Approach

How do we determine which of the three situations described in the foregoing occur for any given quadratic form, $au^2 + 2buv + cv^2$, with real coefficients. To determine the sign of $Q(u, v)$, we complete the square in $au^2 + 2buv$ and write $Q(u, v)$ as

$$Q(u,v) = a\left(u + \frac{bv}{a}\right)^2 + \left(c - \frac{b^2}{a}\right)v^2 \tag{2.1.7}$$

provided that $a \neq 0$.

If $a = c = -$, then $Q(u, v) \equiv 2buv$. If $b \neq 0$, then $Q(u, v)$ can be positive or negative. If however, $b = 0$, the quadratic form is eliminated.

If $a \neq 0$, from (2.1.7), we must have a $Q(u, v) > 0$ for all unique u and v different from the pair $(0, 0)$ provided that $a > 0$ and $c - \frac{b^2}{a} > 0$.

In the same vein, $Q(u, v) < 0$ for all nontrivial u and v, provided that we have the inequalities, $a < 0$ and $c - \frac{b^2}{a} < 0$.

Positivity Requirement

A set of *necessary and sufficient* conditions that $Q(u, v)$ be positive for all nontrivial u and v is that

$$a > 0, \qquad \begin{vmatrix} a & b \\ b & c \end{vmatrix} > 0. \tag{2.1.8}$$

2.1.4 Analytic Approach

To find the range of values of $Q(u, v)$, we can examine the set of values that $Q(u, v)$ occupies on the circle $u^2 + v^2 = 1$. If Q is to be positive for all nontrivial values of u and v, we must have

$$\min_{u^2+v^2=1} Q(u, v) > 0 \tag{2.1.9}$$

and to have $Q(u, v)$ negative for all u and v on the unit circle, we must have

$$\max_{u^2+v^2=1} Q(u, v) < 0. \tag{2.1.10}$$

Introducing a Lagrange multiplier, λ, we can rewrite the problem as

$$R(u, v) = au^2 + 2buv + cv^2 - \lambda(u^2 + v^2). \tag{2.1.11}$$

At the stationary points, we must have $\frac{\partial R}{\partial u} = \frac{\partial R}{\partial v} = 0$ so that

$$\begin{aligned} au + bv - \lambda u &= 0 \\ bu + cv - \lambda v &= 0 \end{aligned} \tag{2.1.12}$$

whereupon, we see that λ satisfies

$$\begin{vmatrix} a - \lambda & b \\ b & c - \lambda \end{vmatrix} = 0 \tag{2.1.13}$$

$$\lambda^2 - (a + c)\lambda + ac - b^2 = 0. \tag{2.1.14}$$

The roots of (2.1.14) are real seeing that the discriminant is non-negative $i.e.$

$$(a + c)^2 - 4\left(ac - b^2\right) = (a - c)^2 + 4b^2, \tag{2.1.15}$$

and as long as $a \neq 0$ and $b \neq 0$, the roots are distinct.

If $b = 0$, the roots of the quadratic in (2.1.14) becomes $\lambda_1 = a$, $\lambda_2 = c$. For $\lambda_1 = a$, the linear set of equations from (2.1.12) becomes

$$(a - \lambda_1)\, u = 0 \qquad (c - \lambda_2)\, v = 0 \tag{2.1.16}$$

which leaves u arbitrary and $v = 0$, if $a \neq c$.

Whereas if $b \neq 0$, we obtain the nontrivial solutions of (2.1.12) by using one equation and discarding the other. Therefore, u and v are connected by the relation

$$(a - \lambda_1)\, u = -bv. \tag{2.1.17}$$

For the exact solution, we can add the normalization requirement that $u^2 + v^2 = 1$ so that the values of u and v are

$$\begin{aligned} u_1 &= -b/\left(b^2 + (a - \lambda_1)^2\right)^{1/2} \\ v_1 &= (a - \lambda_1)/\left(b^2 + (a - \lambda_1)^2\right)^{1/2} \end{aligned} \tag{2.1.18}$$

with another set (u_2, v_2) determined in a similar fashion when λ_2 is used in place of λ_1.

CHAPTER 3

VECTORS AND MATRICES

In the previous chapter, we looked into the problem of the minima and maxima (locally) of a function of a single and two variables. Suppose that we have N variables, and proceed in a similar manner as before, we see that finding basic necessary and sufficient conditions that ensure the positivity of a quadratic form of N variables are of the form

$$Q(x_1, x_2, \ldots, x_N) = \sum_{i,j=1}^{N} a_{ij} x_i x_j \tag{3.0.1}$$

We will thus develop a notation that allows us to solve the problem *analytically* using a minimum of arithmetic or analytical calculation. In this light, we will develop a notation that allows us to study linear transformations such as

$$y_i = \sum_{j=1}^{N} a_{ij} x_j \qquad i = 1, 2, \ldots, N \tag{3.0.2}$$

3.1 Vectors

We shall define a set of N complex-valued numbers as a *vector*, written as

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} \tag{3.1.1}$$

The vector $\boldsymbol{x}$ in (3.1.1) shall be called a *column vector*. If the elements of the vector are stacked horizontally, $i.e.$

$$\boldsymbol{x} = \begin{bmatrix} x_1 & x_2 & \ldots & x_N, \end{bmatrix} \tag{3.1.2}$$

then we shall call it a *row vector*.

Going forward, we shall use the notation of (3.1.1) to represent all forms of vectors we shall be using. When we mean a row vector, we shall use the notation of a transpose of (3.1.1), $i.e.$ $\boldsymbol{x}^T$. Bold font letters such as $\boldsymbol{x}$, or $\boldsymbol{y}$ shall denote vectors and lower-case letters with subscripts i such as x_i, y_i, z_i or p_i, q_i, r_i shall denote the components of a vector. When discussing a particular set of vectors, we shall use the superscripts $\boldsymbol{x}^1, \boldsymbol{x}^2$ e.t.c. N shall denote the dimension of a vector $\boldsymbol{x}$.

One-dimensional vectors are called *scalars* and shall be our quantities of analysis. When we write $\bar{\boldsymbol{x}}$, we shall mean the vector whose components are the complex conjugates of the elements of $\boldsymbol{x}$.

3.1.1 Addition of Vectors

Two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are said to be equal if all of their components, (x_i, y_i) are equal for $i = 1, 2, \ldots, N$. Addition is the simplest of the arithmetic operations on vectors. We shall write the sum of two vectors as $\boldsymbol{x} + \boldsymbol{y}$ so that

$$\boldsymbol{x} + \boldsymbol{y} = \begin{bmatrix} x_1 + y_1 \\ x_2 + y_2 \\ \vdots \\ x_N + y_N, \end{bmatrix} \tag{3.1.3}$$

whereupon we note that the " $+$ " sign connecting $\boldsymbol{x}$ and $\boldsymbol{y}$ is different from the one connecting x_i and y_i.

Homework 1. Prove that we have the *commutativity*, $\boldsymbol{x} + \boldsymbol{y} = \boldsymbol{y} + \boldsymbol{x}$, and the associativity $\boldsymbol{x} + (\boldsymbol{y} + \boldsymbol{z}) = (\boldsymbol{x} + \boldsymbol{y}) + \boldsymbol{z}$

Homework 2. Just as we showed the addition property of two vectors above, show the subtraction property of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$.

3.1.2 Scalar Multiplication

When a vector is multiplied by a scalar, we shall write it out as follows

$$c_1\boldsymbol{x} = \boldsymbol{x}c_1 = \begin{bmatrix} c_1x_1 \\ c_1x_2 \\ \vdots \\ c_1x_N \end{bmatrix} \tag{3.1.4}$$

3.1.3 The Inner Product of Two Vectors

This is a scalar function of two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{N} x_i y_i. \tag{3.1.5}$$

Further to the above, we define the following properties for inner product

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle \tag{3.1.6a}$$

$$\langle \boldsymbol{x} + \boldsymbol{y}, \boldsymbol{u} + \boldsymbol{v} \rangle = \langle \boldsymbol{x}, \boldsymbol{u} \rangle + \langle \boldsymbol{x}, \boldsymbol{v} \rangle + \langle \boldsymbol{y}, \boldsymbol{u} \rangle + \langle \boldsymbol{y}, \boldsymbol{v} \rangle \tag{3.1.6b}$$

$$\langle c_1\boldsymbol{x}, \boldsymbol{y} \rangle = c_1 \langle \boldsymbol{x}, \boldsymbol{y} \rangle \tag{3.1.6c}$$

The above is an easy way to *multiply* two vectors. The inner product is important because $\langle \boldsymbol{x}, \boldsymbol{x} \rangle$ can be considered as the square of the "length" of the real vector $\boldsymbol{x}$.

Homework 3. Prove that $\langle a\boldsymbol{x}+b\boldsymbol{y}, a\boldsymbol{x}+b\boldsymbol{y}\rangle = a^2\langle \boldsymbol{x},\boldsymbol{x}\rangle + 2ab\langle \boldsymbol{x},\boldsymbol{y}\rangle + b^2\langle \boldsymbol{y},\boldsymbol{y}\rangle$ is a non-negative quadratic form in the scalar variables a and b if $\boldsymbol{x}$ and $\boldsymbol{y}$ are real.

Homework 4. Hence, show that for real-valued vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, that the Cauchy-Schwarz Inequality $\langle \boldsymbol{x},\boldsymbol{y}\rangle^2 \leq \langle \boldsymbol{x},\boldsymbol{x}\rangle\langle \boldsymbol{y},\boldsymbol{y}\rangle$ holds.

Homework 5. Using the above result, show that for any two complex vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, $|\langle \boldsymbol{x},\boldsymbol{y}\rangle|^2 \leq \langle \boldsymbol{x},\bar{\boldsymbol{x}}\rangle\,\langle \boldsymbol{y},\bar{\boldsymbol{y}}\rangle$

Homework 6. Show that the *triangle inequality*

$$\langle \boldsymbol{x}+\boldsymbol{y}, \boldsymbol{x}+\boldsymbol{y}\rangle^{\frac{1}{2}} \leq \langle \boldsymbol{x},\boldsymbol{x}\rangle^{\frac{1}{2}} + \langle \boldsymbol{y},\boldsymbol{y}\rangle^{\frac{1}{2}}$$

holds for any two real-valued variables.

3.1.4 Orthogonality

Two vectors are said to be orthogonal if their inner product is 0 $i.e.$

$$\langle \boldsymbol{x},\boldsymbol{y}\rangle = 0 \tag{3.1.7}$$

When the set of real vectors $\{\boldsymbol{x}^i\}$ possess the property that $\langle \boldsymbol{x}^i,\boldsymbol{y}^i\rangle = 1$, then we say they are *orthonormal.*

Homework 7. Show that $\boldsymbol{x}^i$ are mutually orthogonal and normalized $i.e.$ orthonormal for the following N-dimensional Euclidean basis coordinate vectors

$$\boldsymbol{x}^1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \boldsymbol{x}^2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix} \qquad \boldsymbol{x}^N = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{3.1.8}$$

3.2 Matrices

We can write an array of complex numbers in the form

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1N} \\ x_{21} & x_{22} & \dots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NN} \end{bmatrix} \tag{3.2.1}$$

The matrix of (3.2.1) shall be called a *square matrix.* The quantities x_{ij} are the *elements* of the matrix X; the quantities $x_{i1}, x_{i2}, \dots, x_{iN}$ are the ith *rows* of the matrix X and the quantities

$x_{1j}, x_{2j}, \ldots, x_{Nj}$ are the jth columns of X. We denote matrices with upper case letters or the lower-case subscript notations

$$X = (x_{ij}) \tag{3.2.2}$$

while the *determinant* of the array associated with (3.2.1) shall be denoted $|X|$ or $|x_{ij}|$.

Similar to the equality definition between vectors, two matrices are said to be equal if and only if their elements are equal $i.e.$

$$A + B = (a_{ij} + b_{ij}) \tag{3.2.3}$$

Scalar multiplication of a matrix can be expressed as

$$c_1 X = X c_1 = (c_1 x_{ij}) \tag{3.2.4}$$

Lastly, by $\bar{X}$ we shall mean the matrix whose elements are the complex conjugates of X. X is a real matrix if the elements of X are real.

3.2.1 Vector by Matrix Multiplication

Recall the linear transformation

$$y_i = \sum_{j=1}^{N} a_{ij} x_j \qquad i = 1, 2, \ldots, N \tag{3.2.5}$$

where a_{ij} are complex quantities. For two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ related as above, we have

$$\boldsymbol{y} = A\boldsymbol{x} \tag{3.2.6}$$

to describe the multiplication of a vector $\boldsymbol{x}$ by a matrix X.

Homework 8. Consider the identity matrix I, so defined

$$I = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \tag{3.2.7}$$

$i.e.$ $I = (\delta_{ij})$, where δ_{ij} is the Kronecker delta symbol, defined as

$$\delta_{ij} = \begin{cases} 0, & \text{if } i \neq j \\ 1, & \text{if } i = j \end{cases} \tag{3.2.8}$$

Show that

$$\delta_{ij} = \sum_{k=1}^{N} \delta_{ik} \delta_{kj} \tag{3.2.9}$$

Homework 9. Show that

$$\langle A\boldsymbol{x}, A\boldsymbol{x} \rangle = \sum_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij} x_j \right)^2 \tag{3.2.10}$$

3.2.2 Matrix by Matrix Multiplication

Consider (3.2.6). Now, suppose our goal is to generate a second-order linear transformation so defined

$$\boldsymbol{z} = B\boldsymbol{y} \tag{3.2.11}$$

which converts the components of $\boldsymbol{y}$ into components of $\boldsymbol{z}$. To express the components of $\boldsymbol{z}$ in terms of the components of $\boldsymbol{x}$ this, we write

$$z_i = \sum_{k=1}^{N} b_{ik} y_k = \sum_{k=1}^{N} b_{ik} \left(\sum_{j=1}^{N} a_{kj} x_j \right) \tag{3.2.12}$$

$$= \sum_{j=1}^{N} \left(\sum_{k=1}^{N} b_{ik} a_{kj} \right) \boldsymbol{x}_j \tag{3.2.13}$$

Introducing $C = (c_{ij})$ defined as

$$c_{ij} = \sum_{k=1}^{N} b_{ik} a_{kj} \qquad i, j = 1, 2, \ldots, N \tag{3.2.14}$$

we may write

$$\boldsymbol{z} = C\boldsymbol{x} \tag{3.2.15}$$

Since, formally

$$\boldsymbol{z} = B\boldsymbol{y} = B(A\boldsymbol{x}) = B(A\boldsymbol{x}) = (BA)\boldsymbol{x} \tag{3.2.16}$$

so that

$$C = BA \tag{3.2.17}$$

Note the ordering of the matrix product above.

Homework 10. Show that

$$f(\theta_1) f(\theta_2) = f(\theta_2) f(\theta_1) = f(\theta_1 + \theta_2) \tag{3.2.18}$$

where

$$f(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{3.2.19}$$

Homework 11. Let

$$A = \begin{bmatrix} a_1 & a_2 \\ -a_2 & a_1 \end{bmatrix} \text{ and } B = \begin{bmatrix} b_1 & b_2 \\ -b_2 & b_1 \end{bmatrix}, \tag{3.2.20}$$

show that

$$\left(a_1^2 + a_2^2\right)\left(b_1^2 + b_2^2\right) = \left(a_2 b_1 + a_1 b_2\right)^2 + \left(a_1 b_1 - a_2 b_2\right)^2 \tag{3.2.21}$$

Hint: $|AB| = |A||B|$,

3.2.3 Non-Commutativity

Matrix multiplication is not commutative, *i.e.* $AB \neq BA$. For an example, consider the following 3×3 matrices

$$A = \begin{bmatrix} 5 & 6 & 9 \\ 2 & 1 & 6 \\ 3 & 6 & 9 \end{bmatrix} \qquad B = \begin{bmatrix} 1 & 4 & 13 \\ 23 & 6 & 24 \\ 8 & 3 & 9 \end{bmatrix} \tag{3.2.22}$$

where

$$AB = \begin{bmatrix} 215 & 83 & 290 \\ 73 & 32 & 104 \\ 213 & 75 & 264 \end{bmatrix} \qquad \text{and } BA = \begin{bmatrix} 52 & 88 & 150 \\ 199 & 288 & 459 \\ 73 & 105 & 171 \end{bmatrix} \tag{3.2.23}$$

so that $AB \neq BA$. If, however, $AB = BA$, we say A and B *commute*. Note that

$$(AB)^{-1} = B^{-1}A^{-1}. \tag{3.2.24}$$

3.2.4 Associativity

Associativity of matrix multiplication gets preserved unlike the commutativity. So for matrices A, B, and C, we have

$$(AB)\,C = A\,(BC) \tag{3.2.25}$$

that is, the product ABC is unambiguously defined without the parentheses. To prove this, we write the ijth element of AB as

$$a_{ik}b_{kj} \tag{3.2.26}$$

so that the definition of multiplication implies that

$$(AB)\,C = [(a_{ik}b_{kl})\,c_{lj}] \tag{3.2.27}$$

$$A\,(BC) = [a_{ik}\,(b_{kl}c_{lj})] \tag{3.2.28}$$

which establishes the equality $(AB)\,C$ and $A\,(BC)$.

3.2.5 Invariant Vectors

The problem of finding the minimum or maximum of $Q = \sum_{i,j=1}^{N} a_{ij}\boldsymbol{x}_i\boldsymbol{x}_j$ for $\boldsymbol{x}_i$ satisfying the relation $\sum_{i=1}^{N} \boldsymbol{x}_i^2 = 1$ can be reduced to the problem of finding the values of the scalar λ that satisfies the set of linear homogeneous equations

$$\sum_{j=1}^{N} a_{ij}\boldsymbol{x}_j = \lambda\boldsymbol{x}_i, \qquad i = 1, 2, \ldots, N \tag{3.2.29}$$

which possesses nontrivial solutions. Vectorizing, we have

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \tag{3.2.30}$$

Here, $\boldsymbol{x}$ signifies the direction indicated by the N direction numbers $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$, and we are searching for the directions that are invariant.

3.2.6 The Matrix Transpose

We define the transpose of the matrix $A = (a_{ij})$ as $A^T = (a_{ji})$ $i.e.$ the rows of A^T are the columns of A and vice versa. An important consequence of this is that the transformation A on the set of a vector $\boldsymbol{x}$ is same as the transformation of the matrix A^T on the set $\boldsymbol{y}$. This is shown in the following

$$\langle A\boldsymbol{x}, \boldsymbol{y}\rangle = y_1 \sum_{j=1}^{N} a_{1j}\boldsymbol{x}_j + y_2 \sum_{j=1}^{N} a_{2j}\boldsymbol{x}_j + \ldots + y_N \sum_{j=1}^{N} a_{Nj}\boldsymbol{x}_j \tag{3.2.31}$$

which becomes upon rearrangement,

$$\langle A\boldsymbol{x}, \boldsymbol{y}\rangle = x_1 \sum_{i=1}^{N} a_{i1}\boldsymbol{y}_i + x_2 \sum_{i=1}^{N} a_{i2}\boldsymbol{y}_i + \ldots + x_N \sum_{i=1}^{N} a_{iN}\boldsymbol{y}_i \tag{3.2.32}$$

$$= \langle \boldsymbol{x}, A^T\boldsymbol{y}\rangle \tag{3.2.33}$$

We can then regard A^T as the *induced or adjoint transformation* of A. An interesting property of the transpose of a matrix product is that

$$(AB)^T = B^T A^T \tag{3.2.34}$$

3.2.7 Symmetric Matrices

Matrices that satisfy the relation

$$A = A^T \tag{3.2.35}$$

play a crucial role in the study of quadratic forms and such matrices are said to be *symmetric*, with the property that

$$a_{ij} = a_{ji} \tag{3.2.36}$$

Homework 12. Prove that $(A^T)^T = A$

Homework 13. Prove that $\langle A\boldsymbol{x}, B\boldsymbol{y}\rangle = \langle \boldsymbol{x}, A^T B\boldsymbol{y}\rangle$

3.2.8 Hermitian Matrices

The scalar function for complex vectors is the expression $\langle \boldsymbol{x}, \bar{\boldsymbol{y}}\rangle$. Suppose we define $\boldsymbol{z} = \bar{A^T}\boldsymbol{y}$, then

$$\langle A\boldsymbol{x}, \bar{y}\rangle = \langle \boldsymbol{x}, \bar{\boldsymbol{z}}\rangle \tag{3.2.37}$$

$i.e.$ the induced transformation is now $\bar{A^T}$, the complex conjugate of A. Matrices for which

$$A = \bar{A^T} \tag{3.2.38}$$

are called Hermitian. Note that in some literature, the Hermitian matrix is often written as $A^\star$.

3.2.9 Orthogonal Matrices

This section has to do with the invariance of distance between matrices, that is, taking the Euclidean measure of distance as the measure of the magnitude of the real-valued vector $\boldsymbol{x}$. The prodding question of interest is to figure out the linear transformation $\boldsymbol{y} = H\boldsymbol{x}$ that leaves the inner product $\langle \boldsymbol{x}, \boldsymbol{z} \rangle$. Mathematically, we express this problem such that

$$\langle \boldsymbol{x}, \boldsymbol{x} \rangle = \langle H\boldsymbol{x}, H\boldsymbol{x} \rangle \tag{3.2.39}$$

is satisfied for *all* $\boldsymbol{x}$. We know that

$$\langle H\boldsymbol{x}, H\boldsymbol{x} \rangle = \langle \boldsymbol{x}, H^T H\boldsymbol{x} \rangle \tag{3.2.40}$$

and that $H^T H$ is symmetric so that (3.2.39), gives

$$H^T H = I. \tag{3.2.41}$$

Orthogonal Matrix

A real matrix H for which $H^T H = I$ is called *orthogonal*.

3.2.10 Unitary Matrices

This is the measure of the distance of a complex vector, akin to the invariance condition of real-valued matrices (3.2.41). We define the unitary property as follows:

$$H^\star H = I. \tag{3.2.42}$$

Matrices defined as in the foregoing play a crucial role in the treatment of Hermitian matrices, such as the role that orthogonal matrices play in symmetric matrices theory.

3.2.11 Matrix Determinant

The determinant of a scalar is same as the scalar while the determinant of a matrix shall be inductively defined for square matrices. Suppose we have an $n \times n$ matrix A, its determinant is defined as

$$|A| = \sum_{j=1}^{N} (-1)^{i+j} a_{(i,j)} |a^{(i,j)}| \tag{3.2.43}$$

for any value of $i \in [1, n]$, where (3.2.43) is called the Laplace expansion of $|A|$ along its ith row. Equation (3.2.43) shows us that the determinant of the square matrix A is found in terms of the determinants of the $(n-1) \times (n-1)$ matrices. Similarly, the determinants of $(n-1) \times (n-1)$

matrices are defined by $(n-2)\times(n-2)$ and so on until we get to the determinant of 1×1 matrices which are scalars. We can also define the determinant of A as

$$|A| = \sum_{i=1}^{N}(-1)^{i+j}a_{(i,j)}|a^{(i,j)}| \tag{3.2.44}$$

for any value of $j \in [1, n]$. This is termed the Laplace expansion of A along its jth column. It follows that

$$|A_{11}| = A_{11} \tag{3.2.45a}$$

$$\det\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = A_{11}A_{22} - A_{12}A_{21} \tag{3.2.45b}$$

and that

$$\det\begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = A_{11}(A_{22}A_{33} - A_{23}A_{32}) - \tag{3.2.45c}$$

$$A_{12}(A_{21}A_{33} - A_{23}A_{31}) + \tag{3.2.45d}$$

$$A_{13}(A_{21}A_{32} - A_{22}A_{31}) \tag{3.2.45e}$$

3.2.12 Properties of the Matrix Determinant

1. $|AB| = |A||B|$, where A and B are assumed to be of equal dimensions.
2. $|A| = \prod_{i=1}^{N}\lambda_i$, where λ_i are the eigenvalues of A.
3. The inverse of A is said to exist if $AA^{-1} = I$. Such a matrix is said to be *non-singular*. Note that A must be a square matrix in order for it to have a determinant. A square matrix whose inverse does not exist is said to be *singular*.

Take for example,

$$\begin{bmatrix} 3 & 0 \\ 2,1 & \end{bmatrix}\begin{bmatrix} 1/3 & 0 \\ -2/3, 1 & \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0, 1 & \end{bmatrix}. \tag{3.2.46}$$

Then we say that the two matrices on the left are inverses of one another. Among other ways of stating the nonsingularity of A are that

- A's rows or columns are linearly independent.
- $|A| \neq 0$.
- $Ax = b$ has a unique solution x for all b.
- The rank of $A = n$.
- 0 is not an eigenvalue of A.
- A^{-1} exists.

3.2.13 The Matrix Trace

The trace of a matrix exists if and only of the matrix is square. It is defined as the sum of its diagonal elements.

$$\text{Tr}(A) = \sum_{i=1} A_{ii} \tag{3.2.47}$$

Also, the trace can be expressed in terms of the sum of the matrix's eigenvalues,

$$\text{Tr}(A) = \sum_{i=1} \lambda_i. \tag{3.2.48}$$

The trace of a matrix product is not dependent in the order of multiplication of the matrices:

$$\text{Tr}(AB) = \text{Tr}(BA). \tag{3.2.49}$$

3.2.14 Eigenvectors and Eigenvalues of a Matrix

A square $n \times n$ matrix A has n eigenvalues and n eigenvectors. If

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \tag{3.2.50}$$

for a scalar λ and an $n \times 1$ vector $\boldsymbol{x}$ then we say the matrix A has eigenvalues λ and eigenvectors $\boldsymbol{x}$. Together, λ and $\boldsymbol{x}$ are called *eigendata*, the *characteristic roots*, *latent roots*, or *proper numbers and vectors* of the matrix.

Homework 14. If A has eigendata $(\lambda, \boldsymbol{x})$, show that A^2 has eigendata $(\lambda^2, \boldsymbol{x})$.

Homework 15. Show that A^{-1} exists if and only if none of the eigenvalues of A are zero.

Homework 16. Show that the eigenvalues of A are real numbers if A is symmetric.

3.2.15 Other Matrix Properties

A *symmetric* $n \times n$ matrix A can be characterized as either positive definite, positive semidefinite, negative definite, negative semidefinite, or indefinite if matrix A is

- *Positive definite* if $\boldsymbol{x}^T A\boldsymbol{x} > 0$ for all nonzero $n \times 1$ vectors $\boldsymbol{x}$. That is, all the eigenvalues of A are positive real numbers. If A is positive definite, then so is A^{-1}.
- *Positive semidefinite* if $\boldsymbol{x}^T A\boldsymbol{x} \geq 0$ for all nonzero $n \times 1$ vectors $\boldsymbol{x}$. That is, all the eigenvalues of A are non-negative real numbers. A positive semidefinite matrices are sometimes called nonnegative definite.
- *Negative definite* if $\boldsymbol{x}^T A\boldsymbol{x} < 0$ for all nonzero $n \times 1$ vectors $\boldsymbol{x}$. That is, all the eigenvalues of A are negative real numbers. If A is negative definite, then so is A^{-1}.

- *Negative semidefinite* if $\boldsymbol{x}^T A\boldsymbol{x} \leq 0$ for all nonzero $n \times 1$ vectors $\boldsymbol{x}$. That is, all the eigenvalues of A are non-negative real numbers. A positive semidefinite matrices are sometimes called non positive definite.
- When some of the eigenvalues of A are positive and some are negative, then the matrix is said to be *indefinite*.

The singular values of matrix A are defined as

$$\begin{aligned} \sigma^2(A) &= \lambda(A^T A) \\ &= \lambda(A^T A) \end{aligned} \tag{3.2.51}$$

For an $n \times n$ matrix A, we have a $\min(n, m)$ singular values. If $n > m$, then AA^T will have the same eigenvalues as $A^T A$ and an additional $n - m$ zeroes. We do not consider the additional zeroes to be singular values of A because A always has $\min(n, m)$ singular values.

Quiz 2. If A is $n \times m$, what are number of eigenvalues of $A^T A$ and AA^T respectively?

3.2.16 The Matrix Inversion Lemma

This is sometimes called the *Woodbury matrix identity*, named after Max A. Woodbury., *Sherman-Morrison formula*, or the *modified matrices formula*. It a tool frequently used in statistics, system identification, state estimation and control theory. Assume we have a blockwise matrix $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$ where A and D are invertible square matrices, and B and C are not necessarily square. We can define the following matrices

$$\begin{aligned} E &= D - CA^{-1}B \\ F &= A - BD^{-1}C. \end{aligned} \tag{3.2.52}$$

If E is invertible, it follows that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} A^{-1} + A^{-1}BE^{-1}CA^{-1} & -A^{-1}BE^{-1} \\ -E^{-1}CA^{-1} & E^{-1} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{3.2.53}$$

Also, if F were invertible, it follows that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} F^{-1} & -A^{-1}BE^{-1} \\ -D^{-1}CF^{-1} & E^{-1} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix}. \tag{3.2.54}$$

It follows that (3.2.53) and (3.2.54) are two expressions for the inverse of $\begin{pmatrix} A & B \\ C & D \end{pmatrix}$. We must therefore have the upper-left partitions of the two matrices equal so that

$$F^{-1} = A^{-1} + A^{-1}BE^{-1}CA^{-1} \tag{3.2.55}$$

and from the definition of F, we have

$$\left(A - BD^{-1}C\right)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1} \tag{3.2.56}$$

An alternative statement of the matrix inversion lemma is

$$\left(A + BD^{-1}C\right)^{-1} = A^{-1} - A^{-1}B\left(D + CA^{-1}B\right)^{-1}CA^{-1} \tag{3.2.57}$$

Quiz 3. Verify the expressions in (3.2.53) and (3.2.54).

Example 1. Suppose that at Brandeis, you took three courses in your Freshman year namely, RBOT 101, RBOT 103, and RBOT 105 where you got 90%, 85%, and 86% respectively. In your Sophomore year, you took RBOT 201, RBOT 203, and RBOT 205, where you got 65%, 68%, and 92% respectively, and in your junior year, you decide to retake your Sophomore classes, where your scores increased by 10%, 5%, on the first two courses and decreased by 8% in the last course. Your GPA each year increased by 4%, 3.5% and 2.5% respectively. Given your analytical prowess, you decide to model each year's GPA changes with the equation $z = au + bv + cw$, where u and v and w are the scores/grades you got as percentages and a, b, and c are unknown constants. To find the unknown constants, you figure you need to invert the matrix

$$A = \begin{bmatrix} 90 & 85 & 86 \\ 65 & 68 & 92 \\ 71.5 & 71.4 & 84.64 \end{bmatrix} \tag{3.2.58}$$

so that

$$A^{-1} = \begin{bmatrix} 23/20 & 155/104 & -145/52 \\ -207/136 & -569/274 & 6725/1768 \\ 5/16 & 205/416 & -175/208 \end{bmatrix} \tag{3.2.59}$$

It follows that the unknown constants are

$$X = A^{-1}\begin{pmatrix} 4 \\ 7/2 \\ 5/2 \end{pmatrix} = \begin{pmatrix} 2959/1040 \\ -2693/700 \\ 725/832 \end{pmatrix} \tag{3.2.60}$$

As a result, you are able to determine a model which allows you to predict future GPA changes based on how hard you work, sleep, engage in social activities. You can better allocate your time resource and improve your grades in the following years.

Suppose that in the aftermath of generating this model, you now realize that your grade in RBOT 201 the second year was 86% rather than 65%, this means that in order to find the constants, you want to invert

$$\bar{A} = \begin{bmatrix} 90 & 85 & 86 \\ 86 & 68 & 92 \\ 71.5 & 71.4 & 84.64 \end{bmatrix}. \tag{3.2.61}$$

Rather than invert all the matrix all over, you decide to apply a mathematical trick leveraging the inversion lemma. You write $\bar{A} = A + BD^{-1}C$, where

$$B = \begin{bmatrix} 0 & 21 & 0 \end{bmatrix}^T, \; C = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, \text{ and } D = 1 \tag{3.2.62a}$$

so that

$$\bar{A}^{-1} = \left(A + BD^{-1}C\right)^{-1} \tag{3.2.63}$$

$$= A^{-1} - A^{-1}B\left(D + CA^{-1}B\right)^{-1}CA^{-1} \tag{3.2.64}$$

The $(D + CA^{-1}B)^{-1}$ term turns out to be a scalar so that

$$\bar{A}^{-1} = \begin{bmatrix} 0.0506 & 0.0656 & -0.1228 \\ 0.0239 & -0.073 & 0.0551 \\ -0.065 & 0.0035 & 0.0741 \end{bmatrix} \tag{3.2.65}$$

$$\begin{aligned} X &= \bar{A}^{-1} \begin{pmatrix} 4 \\ 7/2 \\ 5/2 \end{pmatrix} \\ &= \begin{pmatrix} 51/407 \\ -134/6037 \\ -148/2361 \end{pmatrix} \end{aligned} \tag{3.2.66}$$

Here, the matrix inversion lemma may not be necessary since the size of the matrix is small. However, if the matrix had a larger size, the computational savings of using the matrix inversion lemma becomes appreciated.

Homework 17. Using the same linear model employed in 1, suppose that on a typical weekend, you go to your local Farmers' market and bought tomatoes, bell peppers, and blue berries for $\$35\%$, $\$18\%$, and $\$32\%$ respectively; on your way home, you drove to your local grocery store and found that the prices for each item were actually increased by 10%, 25% for each of tomatoes, and bell peppers, and decreased by 68% for the blue berries. You decide to buy more blueberries that cost a total of $\$50$ at your local grocery; and discovered that $\$5$ worth of tomatoes bought at the Farmers' market was defective and had to be discarded. Can you compute a model that allows you to predict future prices for good tomatoes, blue berries and bell peppers at your local Farmers' market?

CHAPTER 4

REGISTRATION OF OBJECTS IN ROBOTICS.

In this chapter, we are concerned with the problem of optimally aligning two vectors, a model point/shape to a “sensed" or measured point/shape in space e.g. $\nu_1, \nu_2 \in \mathbb{R}^n$ to one another with the minimal amount of errors. To transform between two points in the Cartesian coordinate system is akin to the problem of solving a rigid body motion problem where that yields a rotation and a translation. In addition, the scaling factor may be unknown. For translation, there are three degrees of freedom, while rotation has another three viz., the direction of the axis about which we are rotating, the angle of rotation itself, and the scaling. Three points in either coordinate systems give us nine constraints (with each contributing three coordinates), more than enough to find the seven unknowns. If we discard two of the constraints, we end up with seven equations in seven unknowns that can be developed to allow us to recover the parameters.

There exists many methods of solving this problem. Most of them leverage clever optimization methods and we will be looking into these in this chapter. We could follow the homogeneous transformation scheme we presented in Chapter 1, but we would not have an optimal solution. A popular technique in computer geometry and computer vision is to use the iterative closest point algorithm(ICP), an algorithm by Paul Besl and Neil McKay developed out of General Motors Laboratory in the 1990’s (Besl and McKay, 1992). This is more appropriate for 3D tasks and it describes a generic, representation method for the accurate and computationally efficient registration of three-dimensional (3-D) shapes. The ICP algorithm always converges monotonically to the nearest local minimum of a mean-square distance metric such as an l_2 distance, and this convergence rate is of the order of a few iterations. An important property of the ICP algorithm is that it can register data from unfixtured rigid objects with an ideal geometrical model prior to shape inspection. So, if we want to figure out that two geometric representations are congruent, estimate the motion between them in real-time where the correspondences are not known, ICP tends to be really good for such operations.

Now, suppose our dataset is not a complex geometric primitive[1], but rather a set of two vectors such that we are tasked with the problem of determining the best *unconstrained transformation* between the two sets of coordinates. We can formulate the problem into a constrained optimization problem and thereafter, through clever factorization, turn the problem into a simple one of factorizing the unconstrained transformation into a symmetric and orthogonal matrix by which we may solve for the optimal rotation and translation. The algorithm we shall be looking into will be the one that was invented in crystallography in 1976 and updated in 1978 by Wolfgang Kabsch, today dubbed the Kabsch algorithm (Kabsch, 1978). Kabsch showed that a direct solution was possible, irrespective of the non-linear character of the problem.

While other newer algorithms exist, these are the two popular algorithms that we shall be concerning ourselves with in this chapter.

[1]We shall refer to a geometric primitive as a primitive 3D shape such as a cylinder, square, prism and the likes.

4.1 Preliminaries

We will denote the real line by $\mathbb{R}$. An example of a **metric space** is the **Euclidean n-space** $\mathbb{R}^n$, which consists of $n-$tuples $x = (x_1, x_2, \ldots, x_n)$ where each $x_i \in \mathbb{R}$. We shall mean an $\mathbb{R}^n$ metric space to have the metric

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}. \tag{4.1.1}$$

If $n = 0$, then $\mathbb{R}^0$ is taken to be a single point $0 \in \mathbb{R}$.

A manifold is "locally" similar to one of the example metric spaces $\mathbb{R}^n$. Precisely, a **manifold** is a metric space $\boldsymbol{M}$ with the property that, *if $x \in M$., then there is some neighborhood U of x and some integer $n \geq 0$ such that U is homeomorphic*[2] *to $\mathbb{R}^n$.*

A simple example of a manifold is $\mathbb{R}^n$: for each $x \in \mathbb{R}^n$ we can take U to be everything in $\mathbb{R}^n$.

Quiz 4. Suppose we supply $\mathbb{R}^n$ with an equivalent metric, which makes it homeomorphic to $\mathbb{R}^n$, would it also be a manifold?

Another example of a metric space is an open ball in $\mathbb{R}^n$, wherein one can take U to be the entire open ball since an open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$. Similarly, an open subset V of $\mathbb{R}^n$ is a manifold, *i.e.* for each $x \in V$ we can choose U to be some open ball with $x \in U \subset V$.

The **Euclidean distance** $d(\boldsymbol{r}_1, \boldsymbol{r}_2)$ between two points $\boldsymbol{r}_1 = (x_1, y_1, z_1)$ and $\boldsymbol{r}_2 = (x_2, y_2, z_2)$ is given by

$$d(\boldsymbol{r}_1, \boldsymbol{r}_2) = \|\boldsymbol{r}_1 - \boldsymbol{r}_2\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}. \tag{4.1.2}$$

Suppose that P is a point set with N_p points denoted as $\boldsymbol{p}_i : P = \{p_i\}$ for $i = 1, \ldots, N_p$. The distance between the point $\boldsymbol{q}$ and the point set P is

$$d(\boldsymbol{q}, P) = \min_{i \in \{1, \ldots, N_p\}} d(\boldsymbol{q}, \boldsymbol{p}_i). \tag{4.1.3}$$

We find that the closest point $\boldsymbol{p}_j$ of P satisfies $d(\boldsymbol{q}, \boldsymbol{p}_j) = d(\boldsymbol{q}, P)$.

Suppose that we have a **line segment** that connects the points, $(\boldsymbol{r}_1, \boldsymbol{r}_2)$, the distance between the point $\boldsymbol{r}$ and the line segment l is

$$d(\boldsymbol{p}, l) = \min_{x+y=1} \|x\boldsymbol{r}_1 + y\boldsymbol{r}_2 - \boldsymbol{p}\| \tag{4.1.4}$$

where $x, y \in [0, 1]$.

[2] A homeomorphic mapping means intrinsic topological equivalence between e.g. objects. Two objects are homeomorphic if they can be deformed into each other by a continuous, invertible mapping. Such a homeomorphism ignores the space in which surfaces are embedded, so the deformation can be completed in a higher dimensional space than the surface was originally embedded. Mirror images are homeomorphic, as are Möbius strip with an even number of half-twists, and Möbius strip with an odd number of half-twists (Weisstein, Weisstein).

Homework 18. Find a closed-form expression for the solution to (4.1.4).

Now, if instead of a line segment, suppose we have a set of N_l line segments denoted l_i, and let $L = \{l_i\}$ for $i = 1, \ldots, N_l$. The **distance between the point $\boldsymbol{p}$ and the line segment set** L is

$$d(\boldsymbol{p}, L) = \min_{i \in \{1,\ldots,N_l\}} d(\boldsymbol{p}, l_i). \tag{4.1.5}$$

The closest point y_j on the line segment set L satisfies $d(\boldsymbol{p}, y_j) = d(\boldsymbol{p}, L)$. Let g be a triangle with the following coordinates $\boldsymbol{r} = (x_1, y_1, z_1), \boldsymbol{r}_2 = (x_2, y_2, z_3)$, and $\boldsymbol{r}_3 = (x_3, y_3, z_3)$. The **distance between the point $\boldsymbol{p}$ and the triangle** g is

$$d(\boldsymbol{p}, g) = \min_{x+y+z=1} \|x\boldsymbol{r}_1 + y\boldsymbol{r}_2 + z\boldsymbol{r}_3 - \boldsymbol{p}\| \tag{4.1.6}$$

where $x \in [0, 1],\ y \in [0, 1]$, and $z \in [0, 1]$.

Homework 19. Find a closed-form expression for the problem in (4.1.6).

Now, if we have a collection of N_g triangles G, denoted by g_i such that $G = \{g_i\}$ for $i = 1, \ldots, N_g$. The **distance between the point $\boldsymbol{p}$ and the triangle set** G is

$$d(\boldsymbol{p}, G) = \min_{i \in \{1,\ldots,N_g\}} d(\boldsymbol{p}, g_i), \tag{4.1.7}$$

and the closest point y_j on the triangle set G satisfies the equality $d(\boldsymbol{p}, y_j) = d(\boldsymbol{p}, G)$.

4.1.1 Distance between a Point and a Parameterized Entity

We define a parametric curve and a parametric surface as single parametric entities $\boldsymbol{r}(\boldsymbol{u})$, where $\boldsymbol{u} = u \in \mathbb{R}^1$ denotes a parameterized curve, and $\boldsymbol{u} = (u, v) \in \mathbb{R}^2$ denotes parametric surfaces. We will evaluate a curve within an interval domain e.g. $[x, y]$ while the evaluation domain of a surface can be an arbitrarily closely-connected region in a plane.

We will take the distance from a given point $\boldsymbol{p}$ to a parametric entity E to be

$$d(\boldsymbol{p}, E) = \min_{\boldsymbol{r}(\boldsymbol{u}) \in E} d(\boldsymbol{p}, \boldsymbol{r}(\boldsymbol{u})) \tag{4.1.8}$$

To compute the point-to-curve and point-to-surface distances, let F be the set of N_e parametric entities denoted by E_i, and let $F = \{E_i\}$ for $i = 1, N_e$> The distance between a point $\boldsymbol{p}$ and the parametric entity set F is

$$d(\boldsymbol{p}, F) = \min_{i \in \{1,\ldots,N_e\}} d(\boldsymbol{p}, \boldsymbol{E}_i). \tag{4.1.9}$$

To find the distance from a point to a parametric entity, we can create a simplex-based approximation for e.g. a line segment or triangle. For a parametric space curve $C = \{\boldsymbol{r}(u)\}$, we can compute a polyline $L(C, \delta)$ such that the piecewise-linear approximation never deviates from the space curve by more than a prespecified distance δ. If we tag every point of the polyline with a

corresponding u argument values of the parametric curve, we can obtain an estimate of the closest point from the line segment set.

In a similar vein, for a parametric surface $S = \{\boldsymbol{r}(u, v)\}$, one can compute a triangle set $G(S, \delta)$ such that the piecewise triangular approximation never deviates from the surface by more than a prespecified distance δ. If we tag each truangle vertex with the corresponding (u, v) argument values of the parametric surface, we can find the $U_a, v_a)$ of the argument values of the closest point from the triangle set. The initial value of $\boldsymbol{u}_a$ is assumed to be available such that $\boldsymbol{r}(\boldsymbol{u}_a)$ is very close to the closest point on the parametric entity.

We can employ a Newtonian minimization approach for solving the point to parametric entity problem when a reliable starting point $\boldsymbol{u}_a$ is available. The scalar objective function to be minimized is

$$f(\boldsymbol{u}) = \|\boldsymbol{r}(\boldsymbol{u}) - \boldsymbol{p}\|^2. \tag{4.1.10}$$

Suppose $\Delta = [\partial/\partial \boldsymbol{u}]^T$ is the vector differential gradient operator, the minimum of f must occur at $\Delta f = 0$. If we have a surface, then we must have $\Delta f = [f_u, f_v]^T$, with the 2-D Hessian matrix is given by

$$\Delta\Delta^T(f) = \begin{bmatrix} f_{uu} & f_{uv} \\ f_{uv} & f_{vv} \end{bmatrix} \tag{4.1.11}$$

where the partial derivatives of the objective function is

$$f_u(\boldsymbol{u}) = 2\boldsymbol{r}_u^T(\boldsymbol{u})(\boldsymbol{r}(\boldsymbol{u}) - \boldsymbol{p}) \tag{4.1.12a}$$
$$f_v(\boldsymbol{u}) = 2\boldsymbol{r}_v^T(\boldsymbol{u})(\boldsymbol{r}(\boldsymbol{u}) - \boldsymbol{p}) \tag{4.1.12b}$$
$$f_{uu}(\boldsymbol{u}) = 2\boldsymbol{r}_{uu}^T(\boldsymbol{u})(\boldsymbol{r}(\boldsymbol{u}) - \boldsymbol{p}) + 2\boldsymbol{r}_u^T(\boldsymbol{u})\boldsymbol{r}_u(\boldsymbol{u}) \tag{4.1.12c}$$
$$f_{vv}(\boldsymbol{u}) = 2\boldsymbol{r}_{vv}^T(\boldsymbol{u})(\boldsymbol{r}(\boldsymbol{u}) - \boldsymbol{p}) + 2\boldsymbol{r}_v^T(\boldsymbol{u})\boldsymbol{r}_v(\boldsymbol{u}) \tag{4.1.12d}$$
$$f_{uv}(\boldsymbol{u}) = 2\boldsymbol{r}_{uv}^T(\boldsymbol{u})(\boldsymbol{r}(\boldsymbol{u}) - \boldsymbol{p}) + 2\boldsymbol{r}_u^T(\boldsymbol{u})\boldsymbol{r}_v(\boldsymbol{u}). \tag{4.1.12e}$$

And the update relation for the curve and surface case is

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k - \left[\Delta\Delta^T(f)(\boldsymbol{u}_k)\right]^{-1} \Delta f(\boldsymbol{u}_k) \tag{4.1.13}$$

where $\boldsymbol{u}_0 = \boldsymbol{u}_a$.

4.1.2 Distance between a Point and an Implicit Entity

An implicit geometric entity is the zero set of a possibly vector-valued multivariate function $\boldsymbol{g}(\boldsymbol{r}) = 0$. Examples of this distance could be a point-to-curve or point-to-surface distance. The important thing to bear in mind is that the distance metric for an individual entity, once defined, makes the sets of implicit entities straightforward to implement. The distance from a given point $\boldsymbol{p}$ to an implicit entity I is given by

$$d(\boldsymbol{p}, I) = \min_{\boldsymbol{g}(\boldsymbol{r})=0} d(\boldsymbol{p}, \boldsymbol{r}) = \min_{\boldsymbol{g}(\boldsymbol{r})=0} \|\boldsymbol{r} - \boldsymbol{p}\|. \tag{4.1.14}$$

It is helpful to note that when computing the implicit entity distance from a point, the solution is never closed-form and are usually involved. Suppose that J is the set of N_I parametric entities, represented by I_k and $J = \{I_k\}$ for $k = 1, N_I$. The distance between a point $\boldsymbol{p}$ and the implicit entity set J is given by

$$d(\boldsymbol{p}, J) = \min_{k \in \{1,\ldots,N_I\}} d(\boldsymbol{p}, \boldsymbol{I}_k), \tag{4.1.15}$$

and the closest point $\boldsymbol{y}_j$ on the implicit entity I_j satisfies the equality $d(\boldsymbol{p}, \boldsymbol{y}_j) = d(\boldsymbol{p}, J)$. In order to compute the distance from a point to an implicit entity, we can create a simplex-based approximation such as line segments or triangles. The point-to-line or point-to-triangle set distance yields an approximate closest point $\boldsymbol{r}_a$ which can be used to compute the exact distance.

Typically, we must solve a constrained optimization problem when finding the closest point on an implicit entity, say $\boldsymbol{g}(\boldsymbol{r}) = 0$ to a point $\boldsymbol{p}$ in order to minimize a quadratic objective function that is subject to a nonlinear constraint

$$\min f(\boldsymbol{r}) = \|\boldsymbol{r} - \boldsymbol{p}\|^2 \tag{4.1.16}$$

where $\boldsymbol{g}(\boldsymbol{r}) = 0$ We can form the augmented Lagrange multiplier system of equations to solve the above, $i.e.$

$$\begin{aligned} \Delta f(\boldsymbol{r}) + \boldsymbol{\lambda}^T \Delta \boldsymbol{g}(\boldsymbol{r}) &= 0 \\ \boldsymbol{g}(\boldsymbol{r}) &= 0 \end{aligned} \tag{4.1.17}$$

where $\Delta - [\partial/\partial \boldsymbol{r}]^T$.

4.1.3 Quaternions

The unit quaternion is a four vector $\boldsymbol{q}_R = [q_0, q_1, q_2, q_3]^T$, where $q_0 \geq 0$, and $q_0^2 + q_1^2 + q_2^2 + q_3 = 1$, used to parameterize a rotation matrix. The 3×3 rotation matrix generated by a unit rotation quaternion is given by

$$R = \boldsymbol{q}_R^T \boldsymbol{q}_R = \begin{bmatrix} q_0^2 + q_1^2 - q_2^2 - q_3 & 2(q_1 q_2 - q_0 q_3) & 2\,(q_1 q_3 + q_0 q_2) \\ 2(q_1 q_2 + q_0 q_3) & q_0^2 + q_1^2 - q_2^2 - q_3 & 2(q_2 q_3 - q_0 q_1) \\ 2(q_1 q_3 + q_0 q_2) & 2(q_2 q_3 + q_0 q_1) & q_0^2 + q_3^2 - q_1^2 - q_2 \end{bmatrix} \tag{4.1.18}$$

For more on unit quaternions, see **??**.

4.2 Closed-form Solution using Least Sum of Squares Errors

As we will see from our sensors, measurements are often inexact, which means we need a way to enforce greater accuracy when determining the transformation parameters. Therefore, we will need more than three points. One approach is to minimize the sum of squares of residual errors using various empirical, graphical, and numerical procedures. Because these are iterative in nature, they

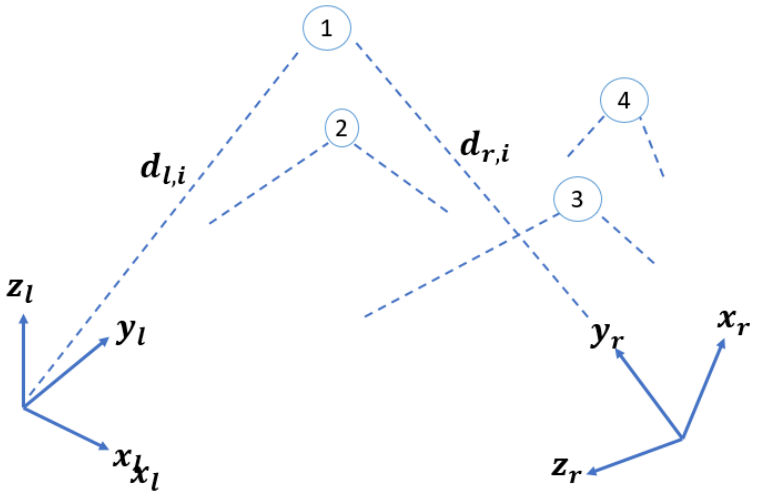


Figure 4.1: Given two coordinate systems, we measure a number of points in the two different coordinate systems. The goal is to find the transformation between the two points.

lead to an approximate solution and while the answer is better, it is imperfect.Iterative methods are repeatedly applied until the residual error is negligible.

There are closed-form solutions which present the absolute orientation in a single step with the best possible transformation given the measurements of the points in the two coordinate systems (Horn, 1987; Kabsch, 1978). With these closed-form least sum of squares methods, we do not need to find an initial good guess as is the case for iterative methods.

4.2.1 Kabsch Algorithm

Suppose we have two sets of vectors $\boldsymbol{x_n}$ and $\boldsymbol{y_n}$ where $n = 1, \ldots, N$, and weight w_n that corresponds to each pair $\boldsymbol{x_n}$ and $\boldsymbol{y_n}$. Our goal is to find an orthogonal matrix $\cup = (u_{ij})$ which minimizes the cost function

$$C = \frac{1}{2}\Sigma_n w_n \left(\cup \boldsymbol{x_n} - \boldsymbol{y_n}\right)^2 \tag{4.2.1}$$

subject to

$$\sum_k u_{ki}u_{kj} - \delta_{ij} = 0 \tag{4.2.2}$$

where δ_{ij} are the elements of a unit matrix. When there is a translation, we can find the centroid of the vector sets to the origin.

In order to solve the problem, we may introduce a symmetric Lagrangian matrix of multipliers, $L = (l_{ij})$ and an auxiliary function as follows

$$D = \frac{1}{2}\Sigma_{i,j} l_{ij} \left(\Sigma_k u_{kl} u_{kj} - \delta_{ij}\right) \tag{4.2.3}$$

so that we can form the Lagrangian, $E = C + D$. For each condition in eq. 4.2.2, we have an independent number l_{ij} so that the constrained minimum of C is part of the free minima of D. A free minimum of D can occur if

$$\frac{\partial E}{\partial u_{ij}} = \sum_k u_{ik} \left(\Sigma_n w_n x_{nk} x_{nj} + l_{k,j}\right) - \sum_n w_n y_{nl} x_{nj} = 0 \tag{4.2.4}$$

and

$$\frac{\partial^2 E}{\partial u_{mk} \partial u_{ij}} = \delta_{mi} \left(\Sigma_n w_n x_{nk} x_{nj} + l_{kj}\right) \tag{4.2.5}$$

are elements of a positive definite matrix x_{nk} and y_{nk} are the kth elements of $\boldsymbol{x_n}$ and $\boldsymbol{y_n}$. Now, suppose we have a matrix $R = (r_{ij})$ and a symmetric matrix $\boldsymbol{S} = (s_{ij})$, such that

$$r_{ij} = \sum_n w_n y_{ni} x_{nj} \tag{4.2.6}$$

and

$$s_{ij} = \sum_n w_n x_{ni} x_{nj}. \tag{4.2.7}$$

If the matrix (4.2.5) has 1 along its diagonal, we must have the minimum of the Lagrangian E to mean that $S + L$ is positive definite, and (4.2.4) translates to

$$U.\,(S + L) = R. \tag{4.2.8}$$

Our goal would be to find a matrix L of Lagrange multipliers so that $\cup$ is orthogonal. We can do this by multiplying both sides of (4.2.8) by their transposed matrices so that we can get rid of matrix $\cup$ as follows:

$$\begin{aligned} U(S + L)^T (S + L) &= (S + L)^T U^T U (S + L) \\ &= (S + L)(S + L) = R^T R. \end{aligned} \tag{4.2.9}$$

Now, we know that $R^T R$ is a symmetric positive definite matrix so that we can find the eigenvalues λ_k and eigenvectors $\boldsymbol{v}_k$ using standard procedures e.g. single value decomposition. Thus, since $S + L$ is symmetric and positive definite, it must have normalized eigenvectors, $\boldsymbol{v}_k$ and positive eigenvalues $\sqrt{\lambda_k}$ so that the Lagrange multipliers are

$$l_{ij} = \Sigma_k \sqrt{\lambda_k}; \qquad \boldsymbol{v}_{ki} \boldsymbol{v}_{ki} - s_{ij} \tag{4.2.10}$$

where $\boldsymbol{v}_{ki}$ signifies the ith component of $\boldsymbol{v}_k$ and the effect of the orthogonal matrix U on these eigenvectors $\boldsymbol{a}_k$ is determined from (4.2.8) which defines the unit vectors $\boldsymbol{q}_k$ as

$$\boldsymbol{q}_k = U.\boldsymbol{v}_k = \frac{1}{\sqrt{\lambda_k}} U(S+L)\boldsymbol{v}_k = \frac{1}{\sqrt{\lambda_k}} R\boldsymbol{v}_k. \tag{4.2.11}$$

The solution to find the constraint minimum of the minimum of the proposed cost function in (4.2.1) is then given by,

Kabsch's Optimal Rotation

$$u_{ij} = \Sigma_k b_{kl} a_{kj}. \tag{4.2.12}$$

4.2.2 Examples

There are clever ways of solving the optimal rotation between two vectors.

There is a jupyter notebook at the following link: Kabsch Algorithm and Implementation. For your convenience, it is included as a pdf file below.

Kabsch

December 15, 2020

0.1 Overview

Throughout this course, we will be leveraging Google's Colab Notebooks to reinforce the concepts we have been learning in class. For an introduction into how to use collab, in case you are not already familiar with it, have a go at this overview of Colaboratory features.

0.1.1 Kabsch's Algorithm

As stated in the course notes, the Kabsch algorithm is a very versatile tool for optimally aligning two vectors to one another. In this example, we are provided with two point sets - a model set and a point (measured) set, and our goal would be to compute the optimal rotation matrix U that allows us to efficiently rotate the point set into the model set.

0.1.2 Load the Measured Point Set

For the example we are interested in, we have measured the position of an object in 3D space using a Northern Digital Inc's Polaris Camera. The points are collected as a set of three-dimensional (3D) points in space, arranged in rows of `(x,y,z)` tuples and they are as given by the `measured_points_full` function below:

```
In [9]: # Here, we are importing all the libraries we will be using in these notebook
        import os
        import numpy as np
        from os.path import join, expanduser
        import scipy.linalg as LA
```

```
In [8]: def measured_points_full():
            # these are the (x,y,z) tuples
            pre_calib = {
                '0,0.0': [-369.88531494140625, 101.30087280273438, -1960.3780517578125],
                '200,0': [-369.8937683105469, 101.32111358642578, -1960.302734375],
                '0,0.1': [-369.8780212402344, 101.32646942138672, -1960.353271484375],
                '220,0': [-369.8780212402344, 101.32646942138672, -1960.353271484375],
                '0,0.2': [-367.74957275390625, 101.65080261230469, -1953.7960205078125],
                '240,0': [-370.8532409667969, 101.074951171875, -1942.255126953125],
                '0,0.3': [-366.7646484375, 101.17594909667969, -1949.628173828125],
                '255,0': [-381.33837890625, 97.10205078125, -1920.667236328125],
                '0,0.4': [-368.0609436035156, 100.83153533935547, -1953.857177734375],
                '0,220': [-382.8047790527344, 100.34918975830078, -1944.807373046875],
```

```
    '0,0.5': [-369.7981262207031, 100.01362609863281, -1958.6396484375],
    '0,240': [-382.71600341796875, 99.87244415283203, -1945.184326171875],
    '0,0.6': [-370.24237060546875, 98.66026306152344, -1957.2281494140625],
    '0,255': [-382.71600341796875, 99.87244415283203, -1945.184326171875],
    '0,0.7': [-370.1295166015625, 98.33242797851562, -1956.1732177734375],
    }
def exp(x):
    'This function expands the array along the second dimension so that '
    return np.expand_dims(x, 1)

# sort pre-recorded points in the order.
measured_calib = np.array(([[
                            pre_calib['0,0.0'],
                            pre_calib['0,220'],
                            pre_calib['0,0.1'],
                            pre_calib['0,240'],
                            pre_calib['0,0.2'],
                            pre_calib['0,255'],
                            pre_calib['0,0.3'],
                            pre_calib['200,0'],
                            pre_calib['0,0.4'],
                            pre_calib['220,0'],
                            pre_calib['0,0.5'],
                            pre_calib['240,0'],
                            pre_calib['0,0.6'],
                            pre_calib['255,0'],
                            pre_calib['0,0.7'],
                        ]]))
"""
 As it is currently, our array has 3 dimensions. We need to reduce the size of the
 array alongthe singleton dimention for efficient matrix manipulations, hence why we
 are squeezing the matrix
"""
measured_calib_zero_centered = np.array(([0, 0, 0]))
for i in range(len(measured_calib)):
    ' find the centroid of the points '
    centered = measured_calib[i] - np.min(measured_calib, 0)
    measured_calib_zero_centered = np.vstack((measured_calib_zero_centered, centered
measured_calib_zero_centered = measured_calib_zero_centered[1:]

return measured_calib_zero_centered
```

0.1.3 Load the model set

It now behooves us to load the model set so we can begin our Kabsch computation. For this, we have them saved in a numpy array. Therefore, we will import numpy as well as associated and needed libraries necessary for our computation.

```
In [12]: model_points = np.array((
                                  [[-1755.87720294,   866.87898685,   283.0353811 ],
                                   [-1755.76266696,   866.8540598,    282.9782946 ],
                                   [-1758.9453555,    857.8363267,    296.13326449],
                                   [-1759.02853104,   865.92774874,   283.52951211],
                                   [-1777.42772925,   826.8692224,    293.38292356],
                                   [-1784.34737705,   836.7521396,    281.74652354],
                                   [-1777.77335781,   826.96331701,   292.88727602],
                                   [-1783.56649649,   836.45510137,   281.56390533],
                                   [-1783.53245361,   836.46510174,   281.54257437],
                                   [-1783.6947516,    836.55773878,   281.52364873],
                                   [-1783.58522171,   836.46979064,   281.55684051],
                                   [-1783.66230977,   836.54098015,   281.50709046],
                                   [-1783.52724697,   836.44927943,   281.56064662],
                                   [-1783.59681243,   836.52118858,   281.52347799],
                                   [-1783.44129296,   836.40624764,   281.5671847 ]]
         ))
```

0.1.4 Get the point set from the function above.

```
In [13]: point_set = measured_points_full()
```

0.2 Now, let us calculate the transformation as we described in our notes

```
In [23]: def Kabsch(P=None, Q=None, augment_Q=True, center=True):
             '''P and Q must be nX3. This rotation is accurate.
              Rotates points in P optimally to measured reference points in Q

              Params
              ======
              Q: Points to be rotated into
              augment_Q: Whether Q was recorded without the zero/home points embedded between su

              '''
             if not isinstance(P, np.ndarray) or not isinstance(Q, np.ndarray):
                 P, Q = prepro()

             # calculate the centroids
             if center:
                 'This only for computed old points'
                 q0 = np.mean(Q, 1)
                 p0 = np.mean(P, 1)

                 Q_ctr = Q - np.expand_dims(q0, 1)
                 P_ctr = P - np.expand_dims(p0, 1)
             else:
                 Q_ctr, P_ctr = Q, P
```

```
    # add the zero points to preconputed control points
    if augment_Q:
        'This only for computed old points'
        Q_aug = np.array(([[0,0,0]]))
        for i in range(Q_ctr.shape[0]-1):
                Q_aug = np.append(Q_aug, np.expand_dims(Q_ctr[i+1], 0),0)
                Q_aug=  np.append(Q_aug, np.expand_dims(Q_ctr[0], 0),0)
        Q_ctr = Q_aug[1:]


    Hmat = P_ctr.T@Q_ctr
    U, S, V = LA.svd(Hmat)
    d = np.sign(np.linalg.det(V@U.T))
    M = np.eye(3); M[-1][-1] =d
    opt_rot = V@M@U.T
    opt_trans = Q_ctr.T- opt_rot@P.T

    return opt_rot, np.mean(opt_trans, 1)
```

0.2.1 Test the algorithm

Remember that we are rotating the points in `point_set` into `model_points`. So we would go ahead and call the Kabsch function above as follows:

```
In [24]: Rot, Trans = Kabsch(model_points, point_set, augment_Q=False, center=False)

In [25]: print(Rot)

[[1. 0. 0.]
 [0. 1. 0.]
 [0. 0. 1.]]


In [26]: print(Trans)

[1775.85125374 -842.66314863 -284.40256961]
```

Homework 20. For the following model points P and measured poinyts Q, compute the optimal rotation matrices for moving points Q into point P. For the three assignments below, report your results within a colab notebook, download the colab notebook as a pdf and upload on Latte.

1.

$$P = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \qquad Q = \begin{bmatrix} 0 & -1 & -1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix} \tag{4.2.13}$$

2.

$$P = \begin{bmatrix} 3172.79468418 & 727.52462347 & 7122.70450243 \\ 165.28953155 & -3552.32467068 & -2045.15346584 \\ 5292.45250241 & -1748.52037006 & -6181.40300009 \\ 1893.07584225 & 5897.19719625 & 3130.41287776 \end{bmatrix}, \tag{4.2.14}$$

$$Q = \begin{bmatrix} 1774.11606309 & -4241.11341178 & 5259.04277742 \\ 6079.70499031 & -98.14197972 & -3442.0914569 \\ 813.07069876 & 3334.26289147 & -6112.55652513 \\ 1856.72080823 & 2328.86927901 & 6322.16611888 \end{bmatrix}$$

3. For a toy problem, measure the coordinates of an object in the world using your favorite measuring instrument (a 3D camera sensor, iPhone app (e.g. ArkIt), android app e.t.c.). Be sure to record the position of the object at multiple points in world coordinates and make sure that the physical locations of these points are known (these are your model points). Then compute the optimal rotation and translation between the model and measured points.

4.2.3 Corresponding Point Set Registration with Quaternions

While the Kabsch algorithm does yield an optimal solution for the rotation of two sets of points that correspond to one another, it leverages the orthonormal rotation matrix with positive determinant in its computations. This suffers from the non-uniqueness of solutions that arise from reflections. Using matrices straightforward is problematic because we need six nonlinear constraints to guarantee the orthonormality of the rotation matrix. To yield a least squares rotation, and translation, we will generally avoid singular value decomposition (SVD) methods in two and three dimensions since we generally do not want reflections. For $n > 3$ in any n-dimensional application, the SVD approach, based on the cross-covariance matrix of two point distributions, does generalize easily to n dimensions.

Let $\boldsymbol{t} = [t_x, t_y, t_z]^T$ denote the translation vector and $\boldsymbol{q}_R = [q_0, q_1, q_2, q_3]^T$ denote the unit quaternion. Suppose further that the complete registration set of vectors is $\boldsymbol{H} = [\boldsymbol{q}_R|\boldsymbol{t}]^T$. Now, let $D_l = \{\boldsymbol{d}_{l_i}\}$ be the measured set of points which we want to align with the model point set

$D_r = \{\boldsymbol{d}_{r_i}\}$, where the cardinality, N_l of D_l is same as that of D_r, N_r, and where each point $\boldsymbol{d}_{l_i}$ corresponds to point $\boldsymbol{d}_{r_i}$ with the same index. We are looking for a transformation of the form

$$D_r = a\boldsymbol{R}(D_l) + \boldsymbol{t} \tag{4.2.15}$$

from the left to the right coordinate system as shown in **??**, where a is a scale factor, and $\boldsymbol{t}$ is the translation vector offset. $R(D_l)$ denotes the rotated version of D_l. Since we do not expect to have a perfect data, it will be difficult to find a scale factor, a translation and a rotation so that the transformation equation is satisfied for every point. Thus, there will be a residual error,

$$\boldsymbol{e}_i = \boldsymbol{d}_{r,i} - a\boldsymbol{R}(\boldsymbol{d}_{l,i}) - \boldsymbol{t} \tag{4.2.16}$$

and the cost function will minimize the sum of squares is given as,

$$f(\boldsymbol{s}) = \min \|\boldsymbol{e}_i\|^2. \tag{4.2.17}$$

Finding Translation

We can find the translation, scale and finally rotation by systematically varying the total error.

Consider the centroids of the measured and point sets,

$$\bar{D}_l = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}_{l,i}, \qquad \bar{D}_r = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{d}_{r,i}, \tag{4.2.18}$$

so that the new coordinates are

$$\boldsymbol{d}'_{l,i} = \boldsymbol{d}_{l,i} - \bar{\boldsymbol{d}}_l, \qquad \boldsymbol{d}'_{r,i} = \boldsymbol{d}_{r,i} - \bar{\boldsymbol{d}}_r. \tag{4.2.19}$$

If we write $\boldsymbol{t}' = \boldsymbol{t} - \bar{\boldsymbol{t}} + a\boldsymbol{R}(\boldsymbol{d}_l)$, it follows that we can write the error as

$$\boldsymbol{e}_i = \boldsymbol{d}'_{r,i} - a\boldsymbol{R}(\boldsymbol{d}'_{l,i}) - \boldsymbol{t}' \tag{4.2.20}$$

and the sum of squares of errors becomes

$$\sum_{i=1}^{n} \|\boldsymbol{d}'_{r,i} - a\boldsymbol{R}(\boldsymbol{d}'_{l,i}) - \boldsymbol{t}'\|^2 \equiv \sum_{i=1}^{n} \|\boldsymbol{d}'_{r,i} - a\boldsymbol{R}(\boldsymbol{d}'_{l,i})\|^2 - 2\boldsymbol{t}'.\sum_{i=1}^{n}\left[\boldsymbol{d}'_{r,i} - a\boldsymbol{R}(\boldsymbol{d}'_{l,i})\right] + n\|\boldsymbol{t}'\|^2. \tag{4.2.21}$$

The middle term on the right hand side vanishes since the measurements are referred to the centroid and we are left with the first and the third terms. The first term is independent of $\boldsymbol{t}'$ and the last term cannot be negative given the squared norm. Thus, the total error to be minimized with $\boldsymbol{t}' = 0$ is

Optimal Translation

$$\boldsymbol{t} = \bar{\boldsymbol{d}}_r - a\ R(\bar{\boldsymbol{d}}_l) \tag{4.2.22}$$

In other words, *the translation is the difference between the right centroid and the scaled and rotated left centroid.*

We can now rewrite the error term from (4.2.20) as

$$\boldsymbol{e}_i = \boldsymbol{d}'_{r,i} - a\boldsymbol{R}(\boldsymbol{d}'_{l,i}) \tag{4.2.23}$$

since $\boldsymbol{t}' = 0$. So the total error to be minimized is

$$\sum_{i=1}^{n} \|\boldsymbol{d}'_{r,i} - a\boldsymbol{R}(\boldsymbol{d}'_{l,i})\|^2. \tag{4.2.24}$$

Finding Scale

Expanding (4.2.24), we find that

$$\sum_{i=1}^{n} \|\boldsymbol{d}'_{r,i}\|^2 - 2a\sum_{i=1}^{n} \boldsymbol{d}'_{r,i}.R(\boldsymbol{d}'_{l,i}) + s^2\sum_{i=1}^{n} \|\boldsymbol{d}'_{l,i}\|^2, \tag{4.2.25}$$

and since rotation preserves distances, $\|R(\boldsymbol{d}'_{l,i})\|^2 = \|\boldsymbol{d}'_{l,i}\|^2$, we can write the foregoing as $S_r - 2sD + s^2S_l$, where S_r and S_l are the sums of the squares of the measurement vectors (relative to their centroids), while D is the sum of the dot products of corresponding coordinates in the right system with the rotated coordinates in the left system. Completing the square in s, we find that

$$\left(a\sqrt{S_l} - D/\sqrt{S_l}\right)^2 + \left(S_rS_l - D^2\right)/S_l. \tag{4.2.26}$$

If we minimize with respect to scale a when the first term is 0 or $a = D/S_l$, we find that

$$s = \frac{\sum_{i=1}^{n} \boldsymbol{d}'_{r,i} \cdot R(\boldsymbol{d}'_{l,i})}{\sum_{i=1}^{n} \|\boldsymbol{d}'_{l,i}.\|^2.} \tag{4.2.27}$$

Finding rotation

To find the optimal rotation, we note that the cross-covariance matrix Σ_{lr} between the sets D_l and D_r is given by

$$\Sigma_{lr} = \frac{1}{N_l}\sum_{i=1}^{N_l} \left[(\boldsymbol{d}_{l,i} - \bar{\boldsymbol{d}}_l)(\boldsymbol{d}_{r,i} - \bar{\boldsymbol{d}}_r)^T\right] \tag{4.2.28}$$

$$= \frac{1}{N_l}\sum_{i=1}^{N_l} \left[\boldsymbol{d}_{l,i}\boldsymbol{d}_{r,i}^T\right] - \bar{\boldsymbol{d}}_l\bar{\boldsymbol{d}}_r^T. \tag{4.2.29}$$

The cyclic components of the skew symmetric matrix $Q_{ij} = \left(\Sigma_{lr} - \Sigma_{lr}^T\right)_{ij}$ are used to construct the column vector $\Delta = [Q_{23} \quad Q_{31} \quad Q_{12}]^T$, so that the vector is then used to form the symmetric matrix

$$Q(\Sigma_{lr}) = \begin{bmatrix} \text{tr}(\Sigma_{lr}) & \Delta^T \\ \Delta & \Sigma_{lr} + \Sigma_{lr}^T - \text{tr}(\Sigma_{lr})\boldsymbol{I}_3 \end{bmatrix} \tag{4.2.30}$$

where $\boldsymbol{I}_3$ is the 3×3 identity matrix and the unit eigenvector $\boldsymbol{q}_R = \begin{bmatrix} q_0 & q_1 & q_2 & q_3 \end{bmatrix}^T$ that corresponds to the maximum eigenvalue of $Q(\Sigma_{lr})$ is chosen as the optimal rotation.

4.3 Iterative Closest Point

The Iterative Closest Point (ICP) algorithm applies to the following sets of problems (i) sets of points, (ii) sets of line segments, (iii) sets of parametric curves, (iv) sets of implicit curves, (v) sets of triangles, (vi) sets of parametric surfaces, and (vii) sets of implicit surfaces. To properly describe the algorithm, we choose a data, P, which is to be moved or registered/positioned to best align with a "model" data X. It is best if the data and model shape are decomposed into a point set if they are not already in point set form. For triangles and line segments, we use their vertices and endpoints respectively; while for curves and surfaces, an approximation to the vertices and endpoints of triangles and lines are used. Suppose we denote, as before, the number of points in the data shape as N_p and N_x as the number of points, line segments, or triangles in the model shape. The distance metric d between an individual data point $\boldsymbol{p}$ and a model shape X will be denoted

$$d(\boldsymbol{p}, X) = \min_{\boldsymbol{x}(X)} \|\boldsymbol{x} - \boldsymbol{p}\|. \tag{4.3.1}$$

The closest point in X that yields the minimum distance is denoted $\boldsymbol{y}$ such that $d(\boldsymbol{p}, \boldsymbol{y}) = d(\boldsymbol{p}, X)$, where $\boldsymbol{y} \in X$.

Quiz 5. 1. What is the worst case asymptotic computation for the closest point in X and why?

2. What is the expected worst case computation time?

When the closest point computation from $\boldsymbol{p}$ to X is performed for each point P, that process is worst case $O(N_p, N_x)$. Let Y denote the resulting set of closest points, and $\mathcal{C}$ the closest point operator, $i.e.$

$$Y = \mathcal{C}(P, X). \tag{4.3.2}$$

For the resultant corresponding point set Y, the least squares registration can be computed as

$$(\boldsymbol{q}, d) = \mathcal{Q}\left(P, Y\right). \tag{4.3.3}$$

and the positions of the data shape point set are] then updated via $P = \boldsymbol{q}(P)$.

Algorithm 1 ICP Algorithm

1: Given point set P with N_p points $\{\boldsymbol{p}_i\}$ from the data shape and the model shape X with N_x supporting geometric primitives: points, lines, or triangles
2: Start the iteration with P_0 set to P, $\boldsymbol{q}_0 = [1, 0, 0, 0, 0, 0, 0]^T$ and $k = 0$ and define the registration vector relative to the initial data set P_0 so that the final registration denotes the complete transformation.
3: Given a mean-square error with preset threshold $\tau > 0$, and a desired registration accuracy, d
4: **while** $\tau > d_k - d_{k+1}$ **do**
5: Compute the closest points, $Y_k = \mathcal{C}(P_k, X)$ (cost: $O(N_o, N_x)$, worst-case: $O(N_p \log N_x)$ average).
6: Compute the registration: $(\boldsymbol{q}_k, d_k) = \mathcal{Q}(P_0, Y_k)$ (cost: $O(N_p)$).
7: Apply the registration, $P_{k+1} = \boldsymbol{q}(P_0)$ (cost: $O(N_p)$).
8: **end while**

CHAPTER 5

STATE ESTIMATION

The next few topics in this course shall involve the quantification of uncertainty in order to enable a robot navigate, move, or understand its environment via visual or audio sensors. In order to do justice to this topic, we shall soon find out that the concept of putting a value or percentage on how sure we are about a robot's environment shall be very helpful in effective control of our robots. Thus the concept of probability shall greatly aid us in quantifying uncertainty. Even so, we introduce the concept of states, grounded in a mathematical theory that allows the engineer to implement a state through discrete-time systems (since we assume that most implementations shall be done on digital computers). By the *state* of a system, we shall loosely mean "those variables that provide a complete representation of the internal condition or status of the system at a given time instant." In this sentiment, the states of a motor system may mean currents that flow through the inductive coils, the position and speed of its motor shaft, or the voltage across the coils of a solenoid valve. The states of a military power may include the number of its aircraft carriers, the size and horsepower of its nuclear submarines, the number of enlisted servicemen in its forces e.t.c. For a biological system, the states might include blood sugar levels, heart and respiration rates, or body temperature.

Robot systems may include mobile platforms for extraterrestrial navigation, robotics arms in assembly lines, autonomous cars, or actuated surgical devices that assist surgeons. Our goal is to treat uncertainty. Uncertainty occurs if the robot lacks important information that hinders it from carrying out assigned tasks. We may classify this uncertainty into five different factors, viz.,

1. **Environments.** The physical world is inherently unpredictable. While the degree of uncertainty in well-structured environments such as assembly lines is small, environments such as highways and private homes are highly dynamic and unpredictable.

2. **Sensors.** Most sensors have limitations in their perceptual ability arising from noise and the range and the resolution of the sensors. For example, environmental disturbances, weather, lighting conditions limit the information that can be extracted from sensors. Secondly, as to range and resolution, cameras cannot see through walls despite the perceptual range that the spatial resolution of the camera is limited.

3. **Models.** In general, models are at best an approximation or a mathematical representation or abstraction of the physical world. As such, model errors are a source of uncertainty that need to be incorporated in modeling robotics problems.

4. **Computation.** Being real-time systems, robots require a lot of computation in order to be able to achieve timely-response through sacrificing accuracy.

We will estimate states as they shall represent latent or underlying variables that influence the physical or chemical or financial properties of the system. And in motivating the study of a system's state, we can resolve to many weapons in our estimation arsenal which may include linear state filtering (the simple Kalman filter), nonlinear state filters (the extended Kalman filter,

unscented Kalman filter e.t.c.), Bayesian estimation, and *frequentist/classical* estimation approaches. In general, state estimation is an important topic to the engineer because:

- We may need to implement a feedback controller in order to regulate a system's behavior. If the application was for a surgeon to regulate blood pH levels, we may need to estimate the system's state. Or if the challenge is to adequately position a patient's head to a position in 3D space during cancer stereotactic radiosurgery, we may need to estimate the position and orientation of the patient's head and neck in the inertial frame.
- If the states in question are curious enough, we may want to measure these states to understand the faults tolerance of the system in order to perform a good fault identification and prognosis. For example, we might want to estimate the internal states of an aircraft system in flight such that if an aircraft engine fails during flight, we can safely monitor system states in real-time in order to determine how long we can continue flying the aircraft or if we should quickly find a near-by airport where we could land the aircraft for maintenance.

In our treatment, therefore, we shall give a brief introduction to linear systems theory, touch upon standard linear filters and then proceed to treat probability theory before we treat nonlinear systems, and decision-making.

5.1 Linear Systems

State-space systems are very important in engineering systems because they allow us (i) to gain insight into the characteristics of the system, (ii) be able to predict future behaviors of the system, (iii) identify the controllable and observable states of the system. The mathematical model of the process allows us to infer the information about the process. State-space models can be classified into linear and nonlinear systems. While most real-world systems are nonlinear, the tools that exist for analyzing and synthesizing nonlinear systems are well-developed and sophisticated that most nonlinear systems can be approximated by linear systems in order to exercise good control and estimation for real-world applications.

A continuous-time, deterministic linear system can be described by the equations

$$\begin{aligned} \dot{x} &= Ax + Bu \\ y &= Cx \end{aligned} \tag{5.1.1}$$

where x is the *state vector* in $\mathbb{R}^n \times 1$, u is the *control vector* in $\mathbb{R}^p \times 1$, and y is an $\mathbb{R}^n \times 1$ vector. Matrices A, B, and C are respectively $n \times n$, $n \times p$ and $n \times 1$ in dimension. The matrix A is often called the system matrix, B the input or control matrix, while C is often called the output matrix. A, B, and C can be time-varying matrices, in which case the system is linear. Otherwise, the solution to the linear system of equations above is

$$x(t) = e^{A(t-t_0)}x(t_0) + \int_{t_0}^{t} e^{A(t-\tau)}Bu(\tau)d\tau \tag{5.1.2}$$

$$y(t) = Cx(t) \tag{5.1.3}$$

where t_0 is the initial time of the system. If the input control law is zero, then we have a *non-autonomous system* i.e.

$$x(t) = e^{A(t-t_0)}x(t_0) \tag{5.1.4}$$

and because of this, e^{AT} is called the state-transition matrix $i.e.$ it describes how the state moves between transitions at different times regardless of external inputs. At $t = t_0$, we have that

$$e^{A\,0} = I, \tag{5.1.5}$$

which is similar to the scalar exponential of a zero. What happens if x is an n-element vector? The solution in (5.1.3) still remains valid but we must note that the exponential of the matrix becomes interpreted as

$$\begin{aligned} e^{At} &= \sum_{j=0}^{\infty} \frac{(At)^j}{j!} \\ &= \mathcal{L}^{-1}\left[sI - A\right]^{-1} = Qe^{\hat{A}t}Q^{-1} \end{aligned} \tag{5.1.6}$$

where the symbol $\mathcal{L}^{-1}$ is the symbol for the inverse Laplace transform and "s" is the Laplace operator. We see that A must be square in order for e^{At} to exist. Q contains the eigenvectors of A and $\hat{A}$ are the Jordan form of A.

Quiz 6. Write a note about the Jordan form. Also, explain how it can be determined from (5.1.6).

Quiz 7. Does the matrix A commute with its exponential i.e. does $A\,e^{At} = e^{At}\,A$?

The matrix $\hat{A}$ is often diagonal, so that case $e^{\hat{A}t}$ can be computed as

$$\hat{A} = \begin{bmatrix} \hat{A}_{11} & 0 & \dots & 0 \\ 0 & \hat{A}_{22} & \dots 0 & \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & \hat{A}_{nn} \end{bmatrix} \quad e^{\hat{A}t} = \begin{bmatrix} e^{\hat{A}_{11}} & 0 & \dots & 0 \\ 0 & e^{\hat{A}_{22}} & \dots 0 & \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & \dots & e^{\hat{A}_{nn}} \end{bmatrix} \tag{5.1.7}$$

From (5.1.6), we can write

$$\left[e^{At}\right]^{-1} = e^{-At} = Qe^{-\hat{A}t}Q^{-1} \tag{5.1.8}$$

Since A and $-A$ have eigenvalues that are negative of each other, e^{At} is always invertible.

Example 2. Suppose we are controlling angular heading of a mobile robot (for example, using voltage applied to its wheels' rotor windings in order to generate command velocity along the x, y, and z heading, i.e. θ, ω and α respectively). The derivative of the angular velocity vector can be written as

$$\begin{aligned} \dot{\theta} &= \omega + \alpha + 3.5\omega_1 + 6\theta_2 \\ \dot{\omega} &= u + 0.1\theta + 2.5\alpha + \omega_1 + \omega_2^2 \\ \dot{\alpha} &= \theta_1 + 2u \end{aligned} \tag{5.1.9}$$

The scalars ω_1, ω_2, θ_1 and θ_2 are acceleration noise terms such as gear backlash, friction, and modeling errors. If our measurement consists of the θ and ω states, it follows that we can write the state space equation as

$$\begin{bmatrix} \dot{\theta} \\ \dot{\omega} \\ \dot{\alpha} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 2.5 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix} u + \begin{bmatrix} 3.5\omega_1 + 6\theta_2 \\ \omega_1 + \omega_2^2 \\ \theta_1 \end{bmatrix}$$
$$y = \begin{bmatrix} 1 & 1 & 0 \end{bmatrix} + \begin{bmatrix} \theta \\ \omega \\ \alpha \end{bmatrix} + \begin{bmatrix} v_x \\ v_y \\ v_z \end{bmatrix} \tag{5.1.10}$$

where $v = [v_x, v_y, v_z]^T$ is the linear velocity vector for the robot.

Example 3. Suppose that

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \tag{5.1.11}$$

It follows that

$$\begin{aligned} e^{At} &= \sum_{j=0}^{\infty} \frac{(At)^j}{j!} \\ &= (At)^0 + (At)^1 + \frac{(At)^2}{2!} + \frac{(At)^3}{3!} + \dots \\ &= I + At \end{aligned} \tag{5.1.12}$$

where the last term follows from the fact that $A^k = 0$ for $k > 1$ so that

$$e^{At} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \tag{5.1.13}$$

Using the expression for the inverse Laplace transform earlier, we have

$$\begin{aligned} e^{At} &= \mathcal{L}^{-1}\left[(sI - A)^{-1}\right] \\ &= \mathcal{L}^{-1}\left(\begin{bmatrix} s & -1 \\ 0 & s \end{bmatrix}^{-1}\right) \\ &= \mathcal{L}^{-1}\begin{bmatrix} 1/s & 1/s^2 \\ 0 & 1/s \end{bmatrix} \\ &= \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} \end{aligned} \tag{5.1.14}$$

Homework 21. Find the eigendata of the matrix A in (5.1.14). Then determine the following terms using the eigenvector and eigenvalue that you may find: $\hat{A}, Q$ and e^{At}.

Homework 22. Produce a one-page report on a control system transfer function.

5.2 State Space Standard Forms

For a linear system, there are many possible state space models that can result in the same *transfer function dynamics*. Therefore, standardizing state space model structures is relevant for solving problems in a conformal way. For consider the following input-output system's *linear difference equation*[1]

$$y_n + a_1 y_{n-1} + \ldots + a_{n-1} y_1 + a_n y = b_0 u_n + b_1 u_{n-1} + \ldots + b_{n-1} u_1 + b_n u \tag{5.2.1}$$

with u and y serving respectively as the input and output, and y_n serving as the nth derivative of y with respect to time. If we take the Laplace transform of both sides, we have

$$Y(s)\left(s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n\right) = U(s)\left(b_0 s^n + b_1 s^{n-1} + \ldots + b_{n-1} s + b_n\right) \tag{5.2.2}$$

so that the transfer function from the input u to the output y can be written as

$$\frac{Y(s)}{U(s)} = \frac{b_0 s^n + b_1 s^{n-1} + \ldots + b_{n-1} s + b_n}{s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n} \tag{5.2.3}$$

5.2.1 Companion form

In *companion form* representation, the coefficients of the transfer function in (5.2.3) are arranged along its far rows or columns. An example would be

$$\begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ 0 & 0 & 1 & \cdots & 0 & -a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix} \tag{5.2.4}$$

or

$$\begin{bmatrix} -a_{n-1} & -a_{n-2} & -a_{n-3} & \cdots & -a_1 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{5.2.5}$$

In general, we use the convenient *observable* and *controllable* canonical forms in control theory. They are exactly the transpose of one another and using either for control design simplifies the system structure so that it can be readily manipulated for a desired control.

---

[1]

Quiz 8. What is the difference between a *linear difference equation* and a *linear ordinary differential equation?*

5.2.2 Modal Form

The modal form is the dual to the companion form. In the modal form, the state matrix is a diagonal matrix with non-repeating eigenvalues such that the control has a unitary influence on each eigenspace, and the output is a linear combination of the contributions from the eigenspaces. That is,

$$A = \begin{bmatrix} -p_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & -p_2 & 0 & \cdots & 0 & 0 \\ 0 & 0 & -p_3 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & -p_n \end{bmatrix} \tag{5.2.6a}$$

$$B = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \quad C = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \tag{5.2.6b}$$

Homework 23. Write out the solution to eq. 24 in modal form.

5.2.3 Controllable Canonical Form

When we want to design a controller that leverages the full state of the system (assuming this is known), often the *controllable canonical form* will come in handy. It is expressed as follows:

$$A = \begin{bmatrix} -a_1 & -a_2 & -a_3 & \cdots & -a_{n-1} & -a_n \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \end{bmatrix} \tag{5.2.7a}$$

$$B = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad C = \begin{bmatrix} b_1 & b_2 & b_3 & \cdots & b_n \end{bmatrix} \quad D = \begin{bmatrix} b_0 \end{bmatrix} \tag{5.2.7b}$$

Example 4. For the system

$$\frac{Y(s)}{U(s)} = \frac{5s^2 - s + 8}{s^2 + 4s - 2} \tag{5.2.8}$$

we can realize the state space representation in canonical form as follows:

1. Observe that n from (5.2.3) is 3, *i.e.* the highest s exponent in the given transfer function.

2. It follows that we have $a_0 = -2$ and $a_1 = 4$; and $b_0 = 5,\ b_1 = -1,\ \ b_2 = 8$, so that we can write the state space model as

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 2 \\ 1 & -4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix} \begin{bmatrix} u_1 & u_2 \end{bmatrix}$$
$$y = \begin{bmatrix} -1 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \boldsymbol{u} \tag{5.2.9}$$

Homework 24. Derive the companion form for the system:

$$\frac{Y(s)}{U(s)} = \frac{3s^2 - 2s + 1}{s^2 - 8s + 5} \tag{5.2.10}$$

The controllable canonical form is helpful in when using the pole placement method for controller design. However, the system's transformation to companion form is based on the controllability matrix which is almost always numerically singular for mid-range orders. It should be avoided for computation when possible.

5.2.4 Observable Canonical Form

In observable canonical form, the transfer function coefficients of (5.2.3) are written in the rightmost column of the A matrix similar to the companion canonical form but the B matrix takes a different form. It is given as follows:

$$A = \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & 0 & \cdots & 0 & -a_1 \\ 0 & 1 & 0 & \cdots & 0 & -a_2 \\ 0 & 0 & 1 & \cdots & 0 & -a_3 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & -a_{n-1} \end{bmatrix} \tag{5.2.11}$$

$$B = \begin{bmatrix} b_n - a_n b_0 \\ b_{n-1} - a_{n-1} b_0 \\ b_{n-2} - a_{n-2} b_0 \\ \vdots \\ b_1 - a_1 b_0 \end{bmatrix} \quad C = \begin{bmatrix} 0 & 0 & \dots & 1 \end{bmatrix}, \quad D = b_0 \tag{5.2.12}$$

This observable canonical form is ill-conditioned for most state-space computation. It should be avoided for computation when possible as its controllability matrix is almost always numerically singular for mid-range orders.

The observable and controllable canonical forms' matrices are respectively transposes of one another.

Homework 25. Transform the exercise of 24 to observable canonical form.

5.3 Nonlinear Systems

All the world is a nonlinear system. He linearized to the right. He linearized to the left. Till nothing was right. And nothing was left. – Stephen Billings.

Our treatment of dynamical so far has involved linear systems. These are optimistic models of the real world as in the reality, nothing is really linear. In general, a nonlinear system is a system which is not linear $i.e.$, does not satisfy the *principle of superposition*. Even a simple resistor exhibits nonlinearity. However, we utilize Ohm's law in approximating the dynamics of a resistor. This is because the equation is valid over a wide enough operating range. In this light, while we may say linear systems do not exist in real life, linear systems are a useful tool for describing nonlinear systems. We will write a general nonlinear system with the equation

$$\begin{aligned} \dot{x} &= f(x, u, w) \\ y &= h(x, v) \end{aligned} \tag{5.3.1}$$

where $f(\cdot)$ and $h(\cdot)$ are arbitrary vector valued functions, w denotes the process noise, and v denotes the measurement noise. We have a *time-varying* system if $f(\cdot)$ and $h(\cdot)$ are explicit functions of t, otherwise, the system is termed *time-invariant*. Suppose that

$$f(x, u, w) = Ax + Bu + w; \text{ and} \tag{5.3.2}$$

$$h(x, v) = Hx + v, \tag{5.3.3}$$

then the system is linear. Otherwise, the system is nonlinear.

Often, we will need to linearize a nonlinear system in order to properly analyze its stability properties or synthesize its parameters for a particular control application. Suppose we have a nonlinear vector function $f(\cdot)$ of a scalar x, we can expand $f(x)$ in a Taylor series around some nominal operating point, $x = \bar{x}$ i.e.

$$f(x) = f(\bar{x}) + \frac{\partial f}{\partial x}\bigg|_{\bar{x}} \tilde{x} + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2}\bigg|_{\bar{x}} \tilde{x}^2 + \frac{1}{3!}\frac{\partial^3 f}{\partial x^3}\bigg|_{\bar{x}} \tilde{x}^3 + \dots \tag{5.3.4}$$

where $\tilde{x} = x - \bar{x}$. For a 2×1 vector x, we can write $f(x)$ as follows:

$$\begin{aligned} f(x) = f(\bar{x}) + \frac{\partial f}{\partial x_1}\bigg|_{\bar{x}} \tilde{x}_1 + \frac{\partial f}{\partial x_2}\bigg|_{\bar{x}} \tilde{x}_2 + \frac{1}{2!}\left(\frac{\partial^2 f}{\partial x_1^2}\bigg|_{\bar{x}} \tilde{x}_1^2 + \frac{\partial^2 f}{\partial x_2^2}\bigg|_{\bar{x}} \tilde{x}_2^2 + 2\frac{\partial^2 f}{\partial x_1 x_2}\bigg|_{\bar{x}} \tilde{x}_1\tilde{x}_2\right) + \\ \frac{1}{3!}\left(\frac{\partial^3 f}{\partial x_1^3}\bigg|_{\bar{x}} \tilde{x}_1^3 + \frac{\partial^3 f}{\partial x_2^3}\bigg|_{\bar{x}} \tilde{x}_2^3 + 3\frac{\partial^3 f}{\partial x_1^2 x_2}\bigg|_{\bar{x}} \tilde{x}_1^2\tilde{x}_2 + 3\frac{\partial^3 f}{\partial x_1 x_2^2}\bigg|_{\bar{x}} \tilde{x}_1\tilde{x}_2^2\right) + \dots \end{aligned} \tag{5.3.5}$$

which can be compactly written as

$$\begin{aligned} f(x) = f(\tilde{x}) + \left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2}\right) f\bigg|_{\bar{x}} + \frac{1}{2!}\left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2}\right)^2 f\bigg|_{\bar{x}} + \\ \frac{1}{3!}\left(\tilde{x}_1\frac{\partial}{\partial x_1} + \tilde{x}_2\frac{\partial}{\partial x_2}\right)^3 f\bigg|_{\bar{x}} + \dots \end{aligned} \tag{5.3.6}$$

And when n is an $n \times 1$ vector, the vector $f(x)$, expanded in a Taylor series becomes

$$f(x) = f(\tilde{x}) + \left(\tilde{x}_1 \frac{\partial}{\partial x_1} + \ldots + \tilde{x}_n \frac{\partial}{\partial x_n}\right) f\Big|_{\bar{x}} + \frac{1}{2!}\left(\tilde{x}_1 \frac{\partial}{\partial x_1} + \ldots + \tilde{x}_n \frac{\partial}{\partial x_n}\right)^2 f\Big|_{\bar{x}} + \frac{1}{3!}\left(\tilde{x}_1 \frac{\partial}{\partial x_1} + \ldots + \tilde{x}_n \frac{\partial}{\partial x_n}\right)^3 f\Big|_{\bar{x}} + \ldots \tag{5.3.7}$$

Suppose we define the operation $D^k_{\tilde{x}} f$ as

$$D^k_{\tilde{x}} = \left(\sum_{i=1}^{n} \tilde{x}_i \frac{\partial}{\partial x_i}\right)^k f(x)\Big|_{\bar{x}} \tag{5.3.8}$$

so that we can define $f(x)$ in Taylor series form as

$$f(x) = f(\bar{x}) + D_{\tilde{x}} f + \frac{1}{2!} D^2_{\tilde{x}} f + \frac{1}{3!} D^3_{\tilde{x}} f + \ldots \tag{5.3.9}$$

$$= f(\bar{x}) + D_{\tilde{x}} f + o(\delta). \tag{5.3.10}$$

If $f(x)$ is "sufficiently smooth", it is not far fetched to see that the above equation turns to

$$f(x) \approx f(\bar{x}) + D_{\tilde{x}} f \approx f(\bar{x}) + \frac{\partial f}{\partial x}|_{\bar{x}} \tilde{x} \approx f(\bar{x}) + A\tilde{x}. \tag{5.3.11}$$

since $o(\delta)$ implies that higher order terms satisfy $\lim_{\delta \to 0} \frac{o(\delta)}{\delta} = 0$, and $A = \frac{\partial f}{\partial x}\Big|_{\bar{x}}$.

Recall (5.3.1), if we choose a nominal operating point $(\bar{x}, \bar{u}, \bar{w})$ and carry out a Taylor series expansion about this nominal point of the nonlinear system of equations, for the state part, we have

$$\begin{aligned}
\dot{x} &= f(x, u, w) \\
&\approx f(\bar{x}, \bar{u}, \bar{w}) + \frac{\partial f}{\partial x}\Big|_{(\bar{x},\bar{u},\bar{w})} (x - \bar{x}) + \frac{\partial f}{\partial u}\Big|_{(\bar{x},\bar{u},\bar{w})} (u - \bar{u}) + \frac{\partial f}{\partial w}\Big|_{(\bar{x},\bar{u},\bar{w})} (w - \bar{w}) + o(\delta) \qquad (5.3.12) \\
&= f(\bar{x}, \bar{u}, \bar{w}) + \frac{\partial f}{\partial x}\Big|_{(\bar{x},\bar{u},\bar{w})} \tilde{x} + \frac{\partial f}{\partial u}\Big|_{(\bar{x},\bar{u},\bar{w})} \tilde{u} + \frac{\partial f}{\partial w}\Big|_{(\bar{x},\bar{u},\bar{w})} \tilde{w} + o(\delta) \\
&= \dot{\bar{x}} + A\tilde{x} + B\tilde{u} + L\tilde{w}
\end{aligned}$$

Since $\tilde{w}$ is a noise term, it suffices that $\tilde{w} = \bar{w} = w$ so that we can write

$$\dot{x} - \dot{\bar{x}} = A\tilde{x} + B\tilde{u} + Lw \qquad \text{or}$$

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u} + Lw. \tag{5.3.13}$$

In other words, we have a linear equation for the deviations of the nonlinear system from the nominal system. It is therefore reason that as long as the deviations are minute enough, the linearization will be valid and the linear equation of (5.3.13) will describe the nonlinear system (5.3.1) well enough.

In a similar vein, the measurement equation from (5.3.1) will be approximated with the Taylor series expansion about the nominal operating point $(\bar{x}, \bar{u})$ as follows:

$$\begin{aligned} y &= h(x,u) \\ &\approx h(\bar{x},\bar{u}) + \left.\frac{\partial h}{\partial x}\right|_{(\bar{x},\bar{\nu})} \tilde{x} + \left.\frac{\partial h}{\partial \nu}\right|_{(\bar{x},\bar{\nu})} \tilde{\nu} + o(\delta) \qquad (5.3.14) \\ &= \bar{y} + C\tilde{x} + D\tilde{\nu} \qquad \text{or} \\ \tilde{y} &= C\tilde{x} + D\tilde{\nu}. \qquad (5.3.15) \end{aligned}$$

It follows that we can "solve" a nonlinear control problem by finding linear operating regions whereby we can solve the control problem, after which we can obtain locally linear solutions for the nonlinear control problem.

Example 5. Consider the longitudinal flight control of a hypersonic aircraft cruising at a Mach number of 15 at an altitude of $110,000ft$. The dynamic equations are (Wang and Stengel, 2000)

$$\dot{V} = (T\cos\alpha - D)/m - \mu\sin\gamma/r^2 \qquad (5.3.16a)$$
$$\dot{\gamma} = (L + T\sin\alpha)/mV - \left[(\mu - V^2 r)\cos\gamma\right]/(Vr^2) \qquad (5.3.16b)$$
$$\dot{h} = V\sin\gamma \qquad (5.3.16c)$$
$$\dot{\alpha} = q - \dot{\gamma} \qquad (5.3.16d)$$
$$\dot{q} = M_{yy}/I_{yy} \qquad (5.3.16e)$$

where

$$L = \frac{1}{2}\rho V^2 S C_L \qquad (5.3.17a)$$
$$D = \frac{1}{2}\rho V^2 S C_D \qquad (5.3.17b)$$
$$T = \frac{1}{2}\rho V^2 S C_T. \qquad (5.3.17c)$$

Here, α is the angle of attack, γ is the flight path angle, rad, r is the radial distance from the center of the Earth, $20,903,500$ft, C_T is the thrust coefficient, C_D is the drag coefficient, C_L is the lift coefficient, L is the lift, D is the drag in lbf, h is the altitude, T is the thrust in lbf, V is the velocity in ft/sec, m is the mass, 9375 slugs, q is the pitch rate in rad/sec, S is the reference area, $3603ft^2$, I_{yy} is the moment of inertia, 7×10^6 slug-ft^2, M_{yy} is the pitching moment in lbf-ft, μ is the gravitational constant, $1.39 \times 10^16 ft^3/s^2$.

We can write the state space vector of the dynamics as follows:

$$\dot{x} = \begin{bmatrix} \dot{x}_1 & \dot{x}_2 & \dot{x}_3 & \dot{x}_4 & \dot{x}_5 \end{bmatrix} = \begin{bmatrix} \dot{V} & \dot{\gamma} & \dot{h} & \dot{\alpha} & \dot{q} \end{bmatrix}$$

.

So that the *nonlinear* dynamics of the hypersonic aircraft at the specified cruising altitude of $110,000$ ft and Mach number 15 becomes

$$\dot{x}_1 = \frac{1}{2m}\rho S\left(C_T\cos x_4 - C_D\right)x_1^2 - \frac{\mu}{r^2}\sin x_2 \tag{5.3.18a}$$

$$\dot{x}_2 = \left(\frac{1}{2m}\rho S C_L - \frac{\mu}{x_1^2 r^2}\right)x_1 + \frac{x_1}{r}\cos x_2 + \frac{1}{2m}\rho S C_D \sin x_4 \tag{5.3.18b}$$

$$\dot{x}_3 = x_1 \sin x_2 \tag{5.3.18c}$$

$$\dot{x}_4 = -\left(\frac{1}{2m}\rho S C_L - \frac{\mu}{x_1^2 r^2}\right)x_1 - \frac{x_1}{r}\cos x_2 - \frac{1}{2m}\rho S C_D \sin x_4 + x_5 \tag{5.3.18d}$$

$$\dot{x}_5 = M_{yy}/I_{yy} \tag{5.3.18e}$$

Following our earlier argument, we proceed to linearize the dynamics by first finding the Jacobian with respect to the state transition matrix, x:

$$\begin{aligned} A &= \frac{\partial f}{\partial x} \\ &= \begin{bmatrix} \frac{1}{m}\rho S\left(C_T\cos x_4 - C_D\right)x_1 & -\frac{\mu}{r^2}\cos x_2 & 0 & -\frac{1}{2m}\rho S C_T \sin x_4 x_1^2 & 0 \\ \frac{1}{2m}\rho S C_L - \frac{\mu}{r^2 x_1^2} + \frac{1}{r}\cos x_2 & -\frac{x_1}{r}\sin x_2 & 0 & \frac{1}{2m}\rho S C_D \cos x_4 & 0 \\ \sin x_2 & x_1\cos x_2 & 0 & 0 & 0 \\ -\frac{1}{2m}\rho S C_L - \frac{1}{r}\cos x_2 - \frac{\mu}{r^2 x_1^2} & \frac{x_1}{r}\sin x_2 & 0 & -\frac{1}{2m}\rho S C_D \cos x_4 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}. \end{aligned} \tag{5.3.19}$$

Similarly, the input matrix can be obtained by finding the Jacobian with respect to the lift, L, drag D, and thrust, T, are

$$B = \frac{\partial f}{\partial u} \tag{5.3.20}$$

$$= \begin{bmatrix} 0 & -1/m & 0 \\ 1/mV & 0 & \frac{\sin\alpha}{mV} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \tag{5.3.21}$$

so that the linear system

$$\dot{\tilde{x}} = A\tilde{x} + B\tilde{u} \tag{5.3.22}$$

approximately describes the nonlinear hypersonic aircraft's deviation from its nominal value $\bar{x}$. We can simulate the lift, drag and thrust with the following nominal control values: $L = \sin 2\pi t$, $D = \cos 2\pi t$, $T = 2L - D$ to find the nominal state trajectory $\bar{x}$.

REFERENCES

Besl, P. J. and N. D. McKay (1992). Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, Volume 1611, pp. 586–606. International Society for Optics and Photonics. 21

Horn, B. K. (1987). Closed-form solution of absolute orientation using unit quaternions. *Josa a 4*(4), 629–642. 26

Kabsch, W. (1978). A discussion of the solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A 34*(5), 827–828. 21, 26

Wang, Q. and R. F. Stengel (2000). Robust nonlinear control of a hypersonic aircraft. *Journal of guidance, control, and dynamics 23*(4), 577–585. 47

Weisstein, E. W. Homeomorphic. 22

Probability Theory

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

My goal is to give you theory foundations and practical tools for your research

I'll give lots of definitions, but the underlying concepts are typically simple

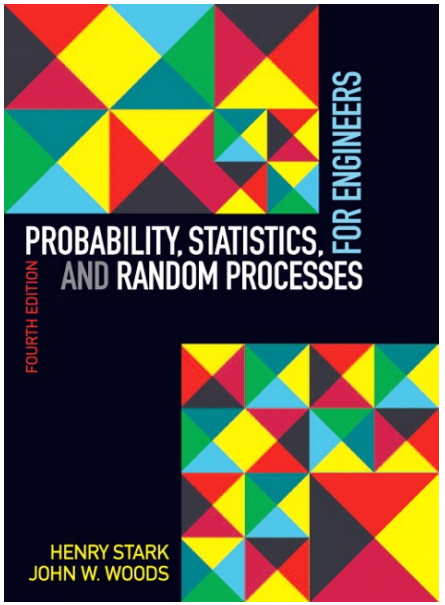
Do the exercises to check your understanding

All referenced Python code is in the `probability_theory` folder

I'm only giving you a small taste of this rich field - take further courses and study on your own!

I will cover material from

- **Stark & Wood's textbook "Probability, Statistics, and Random Processes for Engineers" [1]**
- Assorted other textbooks
- My own experience



Outline

1. What is probability?
2. Boolean and set algebra
3. Axiomatic definition of probability
4. Basic rules of probability

What is probability?

"Probability is a mathematical model to help us study physical systems in an average sense. We have to be able to repeat the experiment many times under the same conditions. Probability then tells us how often to expect the various outcomes." [1]

Why study and use probabilistic models?

"We are forced to use probabilistic models in the real world because we do not know, cannot calculate, or cannot measure all the causes contributing to an effect. The causes may be too complicated, too numerous, or too faint." [1]

Interpretations of probability

Generic
"Probability" means the chance of something

Frequentist
"Probability" means the relative frequency of events

Bayesian
"Probability" means the degree to which we believe something to be true

Axiomatic
"Probability" is a mathematical construct that follows a set of rules

- No interpretation needed - conclusions follow logically from premises
- Be prepared for **counter-intuitive** conclusions

Preliminaries

Set

A **set** is a collection of individual **elements**.

Sets are denoted by braces, with the elements e_i contained inside

$$S = \{e_1, e_2, e_3, \ldots\} \tag{1}$$

Often constructed via set-builder notation

$$S = \{e_i \mid \text{predicate}(e_i)\} \tag{2}$$

"the set of all elements ee-eye such that the predicate holds for ee-eye"

An element e is "in" a set S if S contains e, denoted as $e \in S$.

The **cardinality** of a set is the number of elements in the set.

Examples of Sets

- The set of people reading this slide right now
- The set of hairs on your head
- The **empty set**, denoted $\emptyset$, the set containing nothing at all
  - $\emptyset$ is the only set with cardinality zero
- The set containing the empty set $\{\emptyset\}$
  - This set is not itself empty - it has cardinality one
- The **universal set**, denoted $\mathbb{U}$, the set containing every possible element
- The set of **whole numbers**, denoted $\mathbb{W} = \{0, 1, 2, 3, \ldots\}$
  - It has cardinality $\aleph_0$, a countable infinity
- The set of **real numbers**, denoted $\mathbb{R}$
  - It has cardinality $\mathfrak{c} = 2^{\aleph_0} > \aleph_0$, an uncountable infinity
    - See Cantor's diagonal argument from 1891

Boolean algebra

Basic mathematical operations that apply to **truth/false statements**

- Just like "standard" math operations that apply to numbers like addition, multiplication, etc.

Let x and y be two truth values

Operation	Notation	Definition
Disjunction	$x \vee y$	x is true **or** y is true
Conjunction	$x \wedge y$	x is true **and** y is true
Negation	$\neg x$	x is **not** true
Equivalence	$x \leftrightarrow y$	x is true **if and only if** y is true

Set algebra

Basic mathematical operations that apply to **sets**

- Defined with Boolean algebra applied to set membership

Let E and F be two sets

Operation	Notation	Definition
Union	$E \cup F$	Set of all elements in E **or** in F
Intersection	$E \cap F$	Set of all elements in E **and** F
Complement	E^c	Set of all elements **not** in E
Difference	$E - F$	Set of all elements in E **and not** in F
Exclusive Union	$E \oplus F$	Set of all elements in E **or** F **and not in both**
Subset	$E \subset F$	Every element in E is also in F
Superset	$E \supset F$	Every element in F is also in E
Equality	$E = C$	Every element in E is also in F and vice versa.

Set algebra

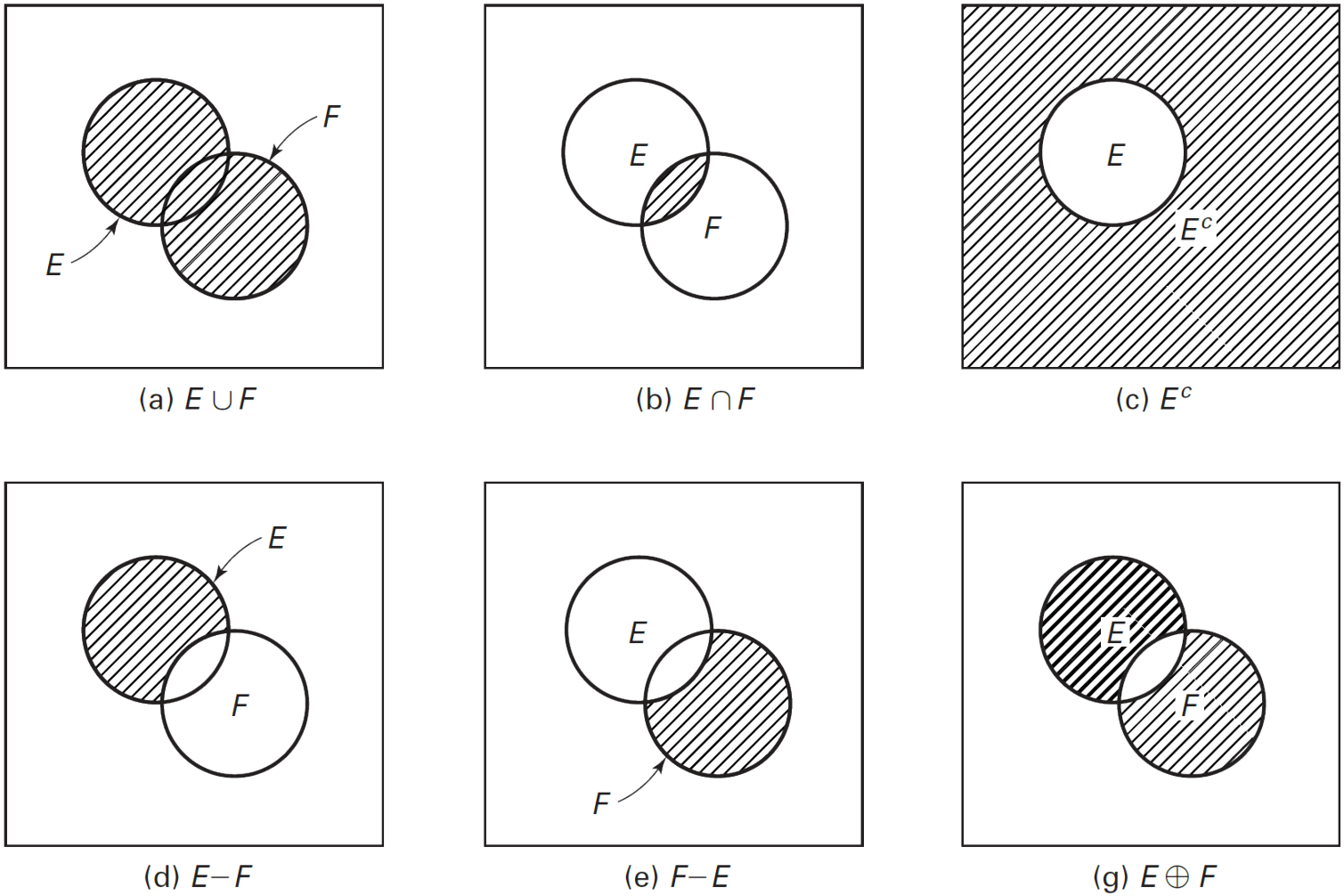


Figure 1: Set operations: (a) Union (b) Intersection (c) Complement (d) Difference (e) Difference (f) Exclusive Union

Set terminology

Let $\{E_i\}$ be a collection of sets

Let A be another set (if unspecified, the universal set $A = \mathbb{U}$ is implied)

- $\{E_i\}$ is **disjoint** or **mutually exclusive** if no elements are shared between any two different sets
- $\{E_i\}$ **collectively exhausts** A if the union of $\{E_i\}$ is A
- $\{E_i\}$ **partitions** A if $\{E_i\}$ is disjoint and collectively exhausts A

Set algebra laws

Set operations are related by simple laws, can be proved using Boolean logic (e.g. truth tables) and definitions

Examples:

- $E = F \quad \leftrightarrow \quad (E \subset F) \wedge (E \supset F)$
- $E \cap E^c = \emptyset$
- $E \cup E^c = \mathbb{U}$
- $E - F = E \cap F^c$
- $E \oplus F = (E - F) \cup (F - E) = (E \cup F) \cap (E \cap F)^c$

More set algebra laws

De Morgan's laws

- $\left[\bigcup_{i=1}^{n} E_i\right]^c = \bigcap_{i=1}^{n} E_i^c$
- $\left[\bigcap_{i=1}^{n} E_i\right]^c = \bigcup_{i=1}^{n} E_i^c$

Associative laws

- $A \cup (B \cup C) = (A \cup B) \cup C$
- $A \cap (B \cap C) = (A \cap B) \cap C$

Distributive laws

- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$

Events

Outcome

A **random experiment** results in **individual outcomes**, denoted as ζ.

Sample space

The **sample space** of a random experiment is the set of all possible outcomes of the experiment, denoted as Ω.

Event

An **event** is a subset of the sample space i.e. a set of outcomes.

In probability

- The sample space plays Ω the role of the universal set $\mathbb{U}$, and is called the **certain event**.
- The empty set $\emptyset$ is called the **null event**.
- Any individual outcome ζ is an element of Ω.

Sigma fields (sigma algebras)

Field

The collection of events $\mathcal{F} = \{E_i\}$ is a **field** if

1. $\emptyset \in \mathcal{F}$ and $\Omega \in \mathcal{F}$
2. If $E_i \in \mathcal{F}$ for all $i = 1, \ldots, n$, then $\bigcup_{i=1}^{n} E_i \in \mathcal{F}$ and $\bigcap_{i=1}^{n} E_i \in \mathcal{F}$
   - "Closed under **finite** union and intersection"
3. If $E \in \mathcal{F}$, then $E^c \in \mathcal{F}$
   - "Closed under complement"

If condition 2 further holds with n countably infinite i.e. "closed under **countably infinite** union and intersection", then $\mathcal{F}$ is a **sigma (σ) field**.

Ensures any union, intersection, and complement of any set of events is well-defined (by construction).

Sigma fields (sigma algebras)

If Ω is continuous and thus uncountable, e.g. $\Omega = \mathbb{R}$, we can generate a sigma field from the set of all open and closed intervals in Ω.

- In this case the sigma field is called the **Borel field**.

We can compute sigma fields of finite and discrete Ω using combinatorics

- See `sigma_field.py`

Axiomatic definition of probability

Probability is a function that maps events to real numbers $P[\cdot] : \mathcal{F} \to [0, 1]$ that satisfies three axioms

1. $P[E] \geq 0$
2. $P[\Omega] = 1$
3. $P[E \cup F] = P[E] + P[F]$ if $P[EF] = \emptyset$

From the axioms we can establish the additional properties

4. $P[\emptyset] = 0$
5. $P[E - F] = P[E] - P[E \cap F]$
6. $P[E^c] = 1 - P[E]$
7. $P[E \cup F] = P[E] + P[F] - P[EF]$

Probability formulation example

Example: Single coin flip

- Sample space is $\Omega = \{H, T\}$ where $H =$ heads, $T =$ tails
- There are 2^2 possible events, $\emptyset$, H, T, Ω
  - Consider events H and T with equal probability
- σ-field is $\mathcal{F} = \{\emptyset, H, T, \Omega\}$

Example: Die roll

- Sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$
- There are 2^6 possible events, each one containing, or not, each of the 6 possible outcomes
  - Consider events $\{1, 3\}$ and $\{2, 3, 4\}$
  - Consider each singleton event equally probable i.e. $P[\{i\}] = 1/6$
- σ-field is...tedious - see Example 1.4-9 [1]

Probability of a union of disjoint events

Let $\{E_i\}_{i=1}^n$ be a set of mutually disjoint events, i.e. $E_i \cap E_j = \phi$ for all $i \neq j$.
Then

$$P\left[\bigcup_{i=1}^{n} E_i\right] = \sum_{i=1}^{n} P\left[E_i\right]. \tag{3}$$

Proof: Use mathematical induction with Axiom 3.

Union bound (Boole's inequality)

Let $\{E_i\}_{i=1}^n$ be a set of events.
Then

$$P\left[\bigcup_{i=1}^{n} E_i\right] \leq \sum_{i=1}^{n} P\left[E_i\right]. \tag{4}$$

Proof: Use mathematical induction with Axiom 7.

Note: The only difference vs the previous result is that the events E_i are not assumed disjoint - the union bound always applies!

Bonferroni inequality

Let $\{E_i\}_{i=1}^n$ be a set of events. Define the sums

$$S_m = \sum_{1 \leq i_1 < i_2 < \cdots < i_m \leq n} P\left[\bigcap_{j=1}^{m} E_{i_j}\right] \tag{5}$$

Then for any $k \in \{1, \ldots, n\}$

$$P\left[\bigcup_{i=1}^{n} E_i\right] \quad \begin{cases} \leq & \text{if } k \text{ odd} \\ \geq & \text{if } k \text{ even} \\ = & \text{if } k = n \end{cases} \quad \sum_{j=1}^{k} (-1)^{j-1} S_j \tag{6}$$

Proof: Use mathematical induction, see Theorem 1.5-1 in [1].

Note: Bonferroni is more tedious, but gives tighter bounds than Boole

Let A and B be two events with nonzero probability.

Joint probability

The **joint probability** of events A and B is the probability of their intersection $P[A \cap B]$.

Intuitively, it is the probability that both A and B will occur.

Conditional probability

The **conditional probability** of event A given B is the ratio

$$P[A|B] = \frac{P[A \cap B]}{P[B]}. \tag{7}$$

Intuitively, it is the probability that event A will occur, given the knowledge that event B already occurred.

Product Rule for events

The joint probability of events A and B can be computed as

$$P[A \cap B] = P[B|A]P[A] \tag{8}$$

When the events are independent we recover the

Proof: Follows by rearranging the definition of conditional probability.

Sum Rule for events

Suppose the events $\{A_i\}_{i=1}^n$ are disjoint and collectively exhaustive, i.e.

- $A_i \cap A_j = \emptyset$ for any $i \neq j$
- $\bigcup_{i=1}^n A_i = \Omega$

Then the **total probability** of event B can be computed as

$$P[B] = \sum_{i=1}^{n} P[B|A_i]P[A_i] = \sum_{i=1}^{n} P[B \cap A_i] \tag{9}$$

Proof: Follows by the product rule and the assumptions on the A_i's.

The sum rule is useful when the conditional probabilities or intersection probabilities are readily available but the total probability is not.

The sum rule is also known as the **law of total probability**.

The total probability is also known as the **marginal probability**, since we are *marginalizing out* the other events A_i.

Total probability - example

Microchip factories

Given information:

1. Factory A makes 4000 chips/day with defect rate of 5%
2. Factory B makes 2000 chips/day with defect rate of 2%
3. Chips from both factories are mixed together at the end of each day then sent to a lab for testing

Question:
What is the probability of getting a defective chip at the lab?

Total probability - example

Solution:
Denote the following events:

- D: Chip is defective
- A: Chip is from factory A
- B: Chip is from factory B

First compute base probabilities from frequency of occurrence:

$$P[A] = \frac{4000}{4000 + 2000} = 66.7\% \tag{10}$$

$$P[B] = \frac{2000}{4000 + 2000} = 33.3\% \tag{11}$$

Now use the law of total probability:

$$P[D] = P[D|A]P[A] + P[D|B]P[B] \tag{12}$$

$$= (5\%)(66.7\%) + (2\%)(33.3\%) \tag{13}$$

$$= \boxed{4\%} \tag{14}$$

Statistical independence

Two events A and B are **statistically independent** if and only if

$$P[A \cap B] = P[A]P[B]. \tag{15}$$

Equivalently, the conditional and unconditional probabilities of A and B are equal:

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A]P[B]}{P[B]} = P[A] \tag{16}$$

$$P[B|A] = \frac{P[B \cap A]}{P[A]} = \frac{P[B]P[A]}{P[A]} = P[B] \tag{17}$$

Intuitively, the outcome B has no effect on the chance of A occurring, and vice versa.

What if there are more than 2 events?

Joint statistical independence

The events $\{A_i\}_{i=1}^n$ are **jointly statistically independent** if and only if for all $k \in \{1, 2, \ldots, n\}$

$$P\left[\bigcap_{1 \leq i_1 < i_2 < \cdots \leq i_k} A_{i_k}\right] = \prod_{1 \leq i_1 < i_2 < \cdots \leq i_k} P\left[A_{i_k}\right] \tag{18}$$

Note: pairwise independence does not suffice!

- See e.g. this note `http://faculty.washington.edu/fm1/394/Materials/2-3indep.pdf`

Pit-stop to build your intuition

Question: Can two disjoint events A and B with $P[A] > 0$, $P[B] > 0$ be statistically independent?

Think about it for a moment

Claim: No, A and B **must be dependent**

Explanation:

1. A, B disjoint means $A \cap B = \emptyset$ which implies $P[A \cap B] = 0$
2. $P[A] > 0$, $P[B] > 0$ implies $P[A]P[B] > 0$
3. Therefore $P[A \cap B] \neq P[A]P[B]$ and the claim follows

Intuition: If we know we flipped heads on a coin, that tells us we did not flip tails.

Derivation from definition of conditional probabilities:

$$P[A|B] = \frac{P[A \cap B]}{P[B]}, \tag{19}$$

$$P[B|A] = \frac{P[A \cap B]}{P[A]} \tag{20}$$

Notice the numerators of the right sides are the same!

Rearrange first line into

$$P[A \cap B] = P[A|B]P[B] \tag{21}$$

and put it into the second line to get Bayes' theorem

$$\boxed{P[B|A] = \frac{P[A|B]P[B]}{P[A]}} \tag{22}$$

Intuition: Lets us reason about conditional probability of "flipped" events

Cancer test

Denote the events

- A: test says patient has cancer
- B: patient actually has cancer

Given information:

- Test has an accuracy of 95%
  - 95% of the time when the test says the patient has cancer, they actually do
  - 95% of the time when the test says the patient does not have cancer, they actually do not
- The cancer rate in the population is 0.5%

Question: The patient being tested for cancer cares about the chance they actually have cancer given the test says they do.
What is this probability?

Bayes' theorem - example

Solution:
Translate given information into math:

$$P[A|B] = P[A^c|B^c] = 95\%, \qquad P[B] = 0.5\% \tag{23}$$

Use the law of total probability to find $P[A]$, the probability of the test saying a patient has cancer:

$$\begin{aligned} P[A] &= P[A|B]P[B] + P[A|B^c]P[B^c] && (24)\\ &= (95\%)(0.5\%) + (100\% - 95\%)(100\% - 0.5\%) && (25)\\ &= 5.45\% && (26) \end{aligned}$$

Now use Bayes' theorem:

$$P[B|A] = \frac{P[A|B]P[B]}{P[A]} = \frac{(95\%)(0.5\%)}{5.45\%} \approx \boxed{8.72\%} \tag{27}$$

How do we resolve this counter-intuitive result?

Even though the test is highly accurate (95%), the chance of actually having cancer is low (8.72%), despite a positive test result. This is because the base rate of cancer is very small, only 0.5%.

On the other hand, conditioning on a positive test result makes the chance of cancer increase dramatically in a relative sense from 0.5% to 8.72%.

From the standpoint of the designer of the cancer test, the smaller the base rate of cancer, the more accurate the test has to be to yield the same probability of a patient actually having cancer.

Homework P1-1:
Consider the previous example. Compute the probability that a patient has cancer, given a negative test result.

Homework P1-2: (1.33 in [1])
A large class in probability theory is taking a multiple-choice test. For a particular question on the test, the fraction of examinees who know the answer is p; $1-p$ is the fraction that will guess. The probability of answering a question correctly is unity for an examinee who knows the answer and $1/m$ for a guessee; m is the number of multiple-choice alternatives.

1. Compute the probability that an examinee knew the answer to a question given that he or she has correctly answered it in terms of m and p.
2. Then evaluate this probability for the specific choice $m = 4$ and $p = 50\%$.

Homework P1-3: (1.35 in [1])
Assume there are three machines A, B, and C in a semiconductor manufacturing facility that make chips. They manufacture, respectively, 25, 35, and 40 percent of the total semiconductor chips there. Of their outputs, respectively, 5, 4, and 2 percent of the chips are defective. A chip is drawn randomly from the combined output of the three machines and is found defective. What is the probability that this defective chip was manufactured by machine A? by machine B? by machine C?

Homework P1-4: (1.55 in [1])
An automatic breathing apparatus (B) used in anesthesia fails with probability P_B. A failure means death to the patient unless a monitor system (M) detects the failure and alerts the physician. The monitor system fails with probability P_M. The failures of the system components are independent events. Professor X, an M.D. at Hevardi Medical School, argues that if $P_M > P_B$ installation of M is useless. Compute the probability of a patient dying with and without the monitor system in place. Take $P_M = 0.1 = 2P_B$. Is Professor X correct in his assessment?

Bibliography I

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

Random Variables

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Random variables

Random variable

A **random variable (RV)** X is a function that maps the sample space Ω to real numbers $\mathbb{R}$ i.e. $X : \Omega \to \mathbb{R}$ that satisfies the following properties:

1. For every Borel set of numbers B, the set $E_B = \{\zeta \in \Omega, X(\zeta) \in B\}$ is an event.
2. $P[X = \infty] = P[X = -\infty] = 0$

Realizations

Upon outcome ζ, a random variable produces a **realization** / **observation** $X(\zeta)$, which is simply a number.

- Think of a realization "popping into being" upon some trigger.
- As shorthand we often refer to the realizations by the same name/variable as the RV.
- We can only observe realizations of the random variable, but not the random variable itself.
- Qualities of the random variable must either be
  1. Assumed before-hand (model)
  2. Inferred from realizations (data)

Random variables - examples

Flip a coin:
X is one or zero for heads or tails respectively

Roll a die:
X is 1, 2, 3, 4, 5, 6, corresponding to the number of dots on the die face

Spin a wheel:
X is the angle at which it lands between 0 and 360 degrees

Cumulative distribution function (cdf)

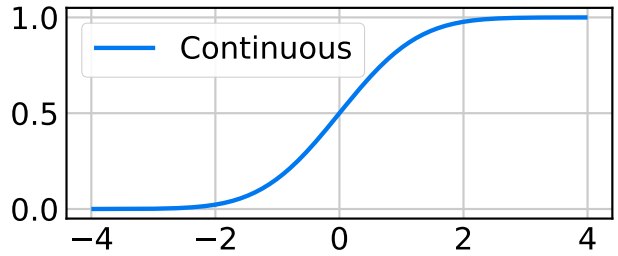
The **cumulative distribution function (cdf)** is defined as

$$F_X(x) = P[\{\zeta | X(\zeta) \leq x\}] \tag{1}$$

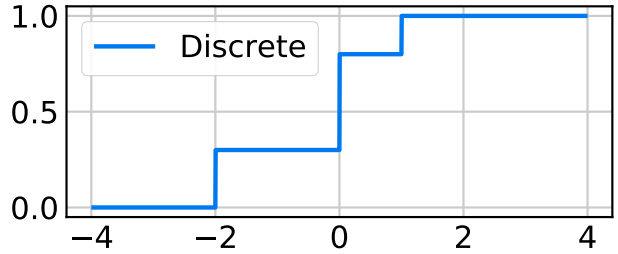
Notation: From here we will usually drop the notation of ζ related to the underlying probability space, so $P[\{\zeta | X(\zeta) \leq x\}]$ becomes $P[X \leq x]$.

Continuous and discrete random variables

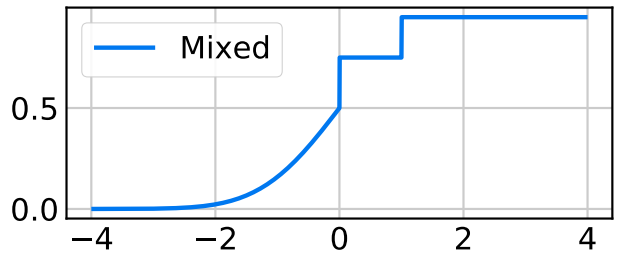
If the cdf $F_X(x)$ is everywhere continuous and differentiable, then X is a **continuous random variable**.



If the cdf $F_X(x)$ is piecewise constant (stairstep shape), then X is a **discrete random variable**.



If neither holds, then X is a **mixed random variable**.



See `mixed.py`

Probability mass function (pmf)

The **probability mass function (pmf)** of a discrete random variable is defined as

$$P_X(x) = P[X = x] \tag{2}$$
$$= P[X \leq x] - P[X < x] \tag{3}$$

Probability density function (pdf)

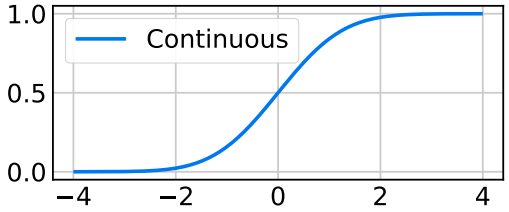
The **probability density function (pdf)** of a continuous random variable* is defined as

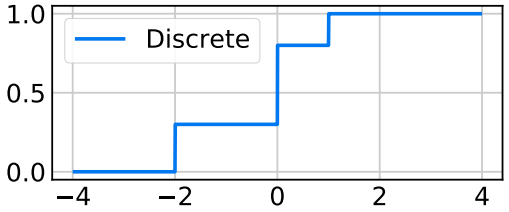
$$f_X(x) = \frac{d}{dx} F_X(x) \tag{4}$$

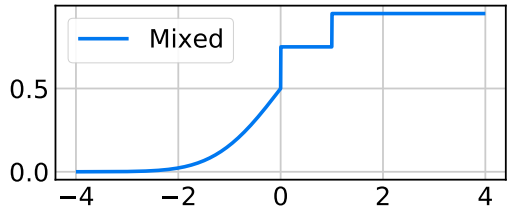
* By introducing Dirac delta functions, the pdf can be defined for discrete and mixed random variables.

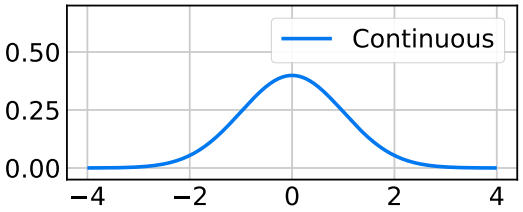
Continuous and discrete random variables

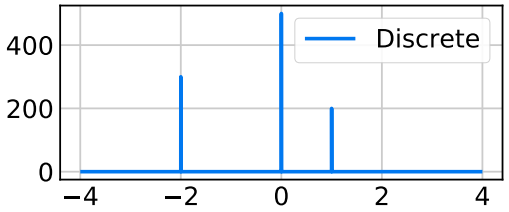
cdfs on top row, pdfs on bottom row

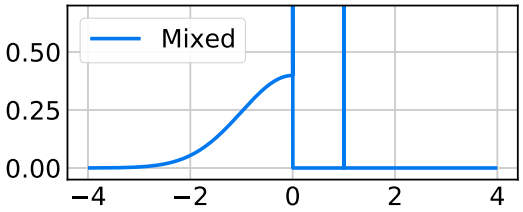












See `mixed.py`

Properties of the cdf

1. $F_X(\infty) = 1$, $F_X(-\infty) = 0$
2. $F_X(x)$ is nondecreasing in x,
   i.e. $X_1 \leq x_2$ implies $F_X(x_1) \leq F_X(x_2)$
3. $F_X(x)$ is continuous from the right,
   i.e. $F_X(x) = \lim_{\epsilon \to 0^+} F_X(x + \epsilon)$

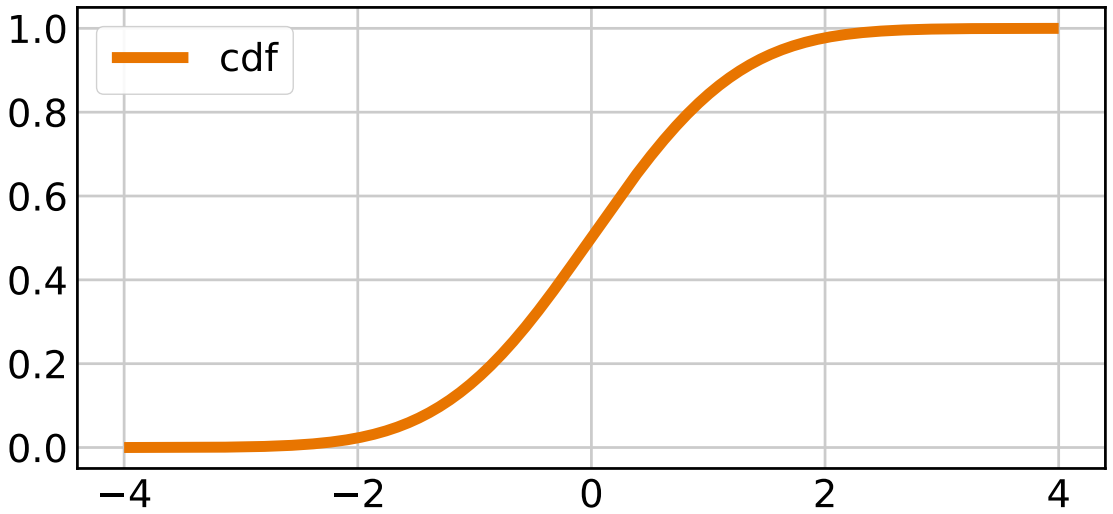


Figure 1: Plot of a typical cdf (std normal)

Properties of the pdf

1. $f_X(x) \geq 0$
2. $\int_{-\infty}^{\infty} f_X(\xi)d\xi = F_X(\infty) - F_X(-\infty) = 1$
3. $F_X(x) = \int_{-\infty}^{x} f_X(\xi)d\xi = P[X \leq x]$
4. $F_X(x_2) - F_X(x_1) = \int_{x_1}^{x_2} f_X(\xi)d\xi = P[x_1 < X \leq x_2]$

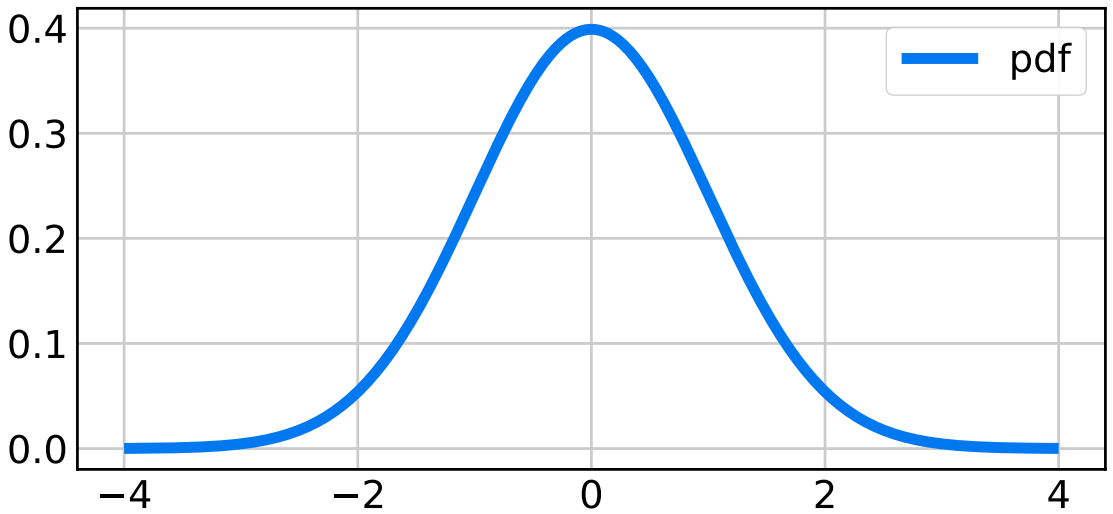


Figure 2: Plot of a typical pdf (std normal)

Knowledge of either the pdf or cdf is sufficient to compute the other, via integration or differentiation.

When we refer to a "distribution," we mean anything that fully specifies a random variable:

- pdf / pmf
- cdf
- Moment generating function (see Ch. 4.5 of [1])
- Characteristic function (see Ch. 4.7 of [1])

Let's introduce a couple of quick concepts before we survey various distributions

Support of a distribution

The **support** of a distribution is the set of values that the random variable X can take with nonzero probability density, i.e.

$$\text{supp}(X) = \{x \mid f_X(x) > 0\}. \tag{5}$$

The distinction between the support and the sample space only comes into effect when the sample space is bigger than required by X

- Sometimes convenient when working with different random variables on a shared sample space
- Example: Two dice with faces $\{1, 1, 1, 2, 3, 3\}$ and $\{2, 3, 4, 5, 6, 6\}$ have different supports $\{1, 2, 3\}$ and $\{2, 3, 4, 5, 6\}$, but we might want a sample space $\{1, 2, 3, 4, 5, 6\}$ to accommodate every possible outcome from either of dice

Mixture distribution

A **mixture distribution** is the distribution of a **mixture random variable** Y formed as a composite of other component random variables $X_1, X_2, \ldots, X_N$ by selecting among them at random according to weights $w_1, w_2, \ldots, w_N$.

If the component pdfs are $f_{X_1}, f_{X_2}, \ldots f_{X_N}$, then the **mixture pdf** is simply the weighted average

$$f_Y(Y) = \sum_{i=1}^{N} w_i f_{X_i} \tag{6}$$

Discrete distributions

Trivial distribution

What happens if we treat a non-random, fixed, constant number as a random variable? (w.l.o.g. set $X = 0$)

Trivial distribution

A **trivial** random variable has the pmf

$$P[X = x] = \begin{cases} 1 & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Accordingly, the pdf is the Dirac delta function

$$f_X(x) = \delta(x) \tag{8}$$

and the cdf is the Heaviside step function

$$F_X(x) = H(x) \tag{9}$$

All discrete distributions can be "built" from mixtures of this distribution.

- Follows by definition of pmf

Bernoulli distribution

A **Bernoulli** random variable has the pmf

$$P[X = x] = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = 0, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

If p is not specified, then assume $p = 1/2$.

Example: A coin flip is Bernoulli where heads $= 1$ and tails $= 0$.

Rademacher distribution

A **Rademacher** random variable has the pmf

$$P[X = x] = \begin{cases} p & \text{if } x = 1, \\ 1 - p & \text{if } x = -1, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Basically just the symmetric version of Bernoulli (which is asymmetric)

- Use whichever is most convenient for the task at hand

Example: A coin flip is Rademacher where heads $= 1$ and tails $= -1$.

Homework P2-1: If X is a Bernoulli random variable, write down a function g such that $Y = g(X)$ is a Rademacher random variable. Also, write down an inverse function $h = g^{-1}$ such that $X = h(Y)$ recovers a Bernoulli distribution. Prove that your functions are correct by directly evaluating the pmfs of $g(X)$ and $h(Y)$.

Binomial distribution

Consider the binomial experiment with n independent success/fail trials, each governed by a Bernoulli RV.

The number of ways to choose k elements from a population of size n (irrespective of their ordering) is called the number of **combinations** and is determined by the **binomial coefficient**

$$\binom{n}{k} = \frac{n!}{(n-k)!k!} \tag{12}$$

The probability of an experiment with k successes and $n-k$ failures is

$$p^k(1-p)^{n-k} \tag{13}$$

Since there are $\binom{n}{k}$ ways in which the experiment could end like this, the probability of seeing an experiment with k successes and $n-k$ failures is

$$\binom{n}{k}p^k(1-p)^{n-k} \tag{14}$$

Binomial distribution

A random variable X follows a **binomial distribution** if it represents getting exactly k successes out of the n trials, whose pmf is

$$P[X = k] = \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k = 0, 1, \ldots, n, \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

where $p \in [0, 1]$ is a parameter representing the success probability of each trial.

Homework P2-2: (1.56 in [1])
In a particular communication network, the server broadcasts a packet of data to N receivers. The server then waits to receive an acknowledgment message from each of the N receivers before proceeding to broadcast the next packet. If the server does not receive all the acknowledgments within a certain time period, it will rebroadcast (retransmit) the same packet. The server is then said to be in the "retransmission mode." It will continue retransmitting the packet until all N acknowledgments are received. Then it will proceed to broadcast the next packet.

Let $p := P$[successful transmission of a single packet to a single receiver along with successful acknowledgment]. Assume that these events are independent for different receivers and separate transmission attempts. Due to random impairments in the transmission media and the variable condition of the receivers, we have that $p < 1$.

(continued on next slide)

Homework P2-2 (cont.):

(a) In a fixed protocol of method of operation, we require that all N of the acknowledgments be received in response to a given transmission attempt for that packet transmission to be declared successful. Let the event $S(m)$ be defined as follows: $S(m) := \{$ a successful transmission of one packet to all N receivers in m or fewer attempts $\}$.

Find the probability

$$P(m) := P[S(m)]$$

Hint: Consider the complement of the event $S(m)$.

(continued on next slide)

Homework P2-2 (cont.):

(b) An improved system operates according to a dynamic protocol as follows. Here we relax the acknowledgment requirement on retransmission attempts, so as to only require acknowledgments from those receivers that have not yet been heard from on previous attempts to transmit the current packet. Let $S_D(m)$ be the same event as in part (a) but using the dynamic protocol. Find the probability

$$P_D(m) := P[S_D(m)]$$

Hint: First consider the probability of the event $S_D(m)$ for an individual receiver, and then generalize to the N receivers.

(continued on next slide)

Homework P2-2 (cont.):

(c) Compare the performance of the two protocols from parts (a) and (b) by comparing $P(m)$ and $P_D(m)$ for $N = 5$ receivers, $m = 2$ transmission attempts, and success probability $p = 0.9$.

Continuous distributions

Uniform distribution

A random variable is **uniform** if the pdf is constant over a finite interval $[a, b]$, i.e. of the form

$$f_X(x) = \begin{cases} \dfrac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases} \tag{16}$$

Tail behavior: density drops to zero instantly outside $[a, b]$

- Log density decays "infinitely" fast

Homework P2-3: Derive an expression for the cdf of a uniform random variable.

Gaussian distribution

A random variable X is **Gaussian** or **normal** if it has a pdf of the form

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right) \tag{17}$$

where μ and σ^2 are parameters (we will define and see later they are the mean and variance).

Notation: $X \sim \mathcal{N}(\mu, \sigma^2)$ is read as "X is distributed according to a normal distribution with mean mu and variance sigma-squared."

Special case: If $\mu = 0$ and $\sigma^2 = 1$, then the distribution is called the **standard normal**.

Tail behavior: log density decays quadratically

See `gaussian.py`

Exponential distribution

A random variable X is **exponential** if it has a pdf of the form

$$f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{if } x \geq 0, \\ 0 & \text{otherwise,} \end{cases} \tag{18}$$

where $\lambda > 0$ is a parameter.

Homework P2-4: Derive an expression for the cdf of an exponential random variable.

Laplace distribution

A random variable X is **Laplace** or **double exponential** if it has a pdf

$$f_X(x) = \frac{1}{2\beta} \exp\left(-\frac{|x-\mu|}{\beta}\right) \tag{19}$$

where μ and β are location and scale parameters.

Notice how similar the Laplace distribution is to a Gaussian

Tail behavior: log density decays linearly - heavier than a Gaussian!

Cauchy distribution

A random variable X is **Cauchy** if it has a pdf of the form

$$f_X(x) = \frac{1}{\pi\gamma}\left(\frac{\gamma^2}{(x-x_0)^2+\gamma^2}\right) \tag{20}$$

where x_0 and γ are location and scale parameters.

Example: The ratio of two independent normal variables $X = Z_1/Z_2$ is Cauchy

The Cauchy distribution is very bizarre pathological distribution

- It actually has an undefined mean and variance! (discussed later)
- Makes parameter estimation tricky

Tail behavior: log density decays logarithmically - heavier than a Laplace!

Other distributions

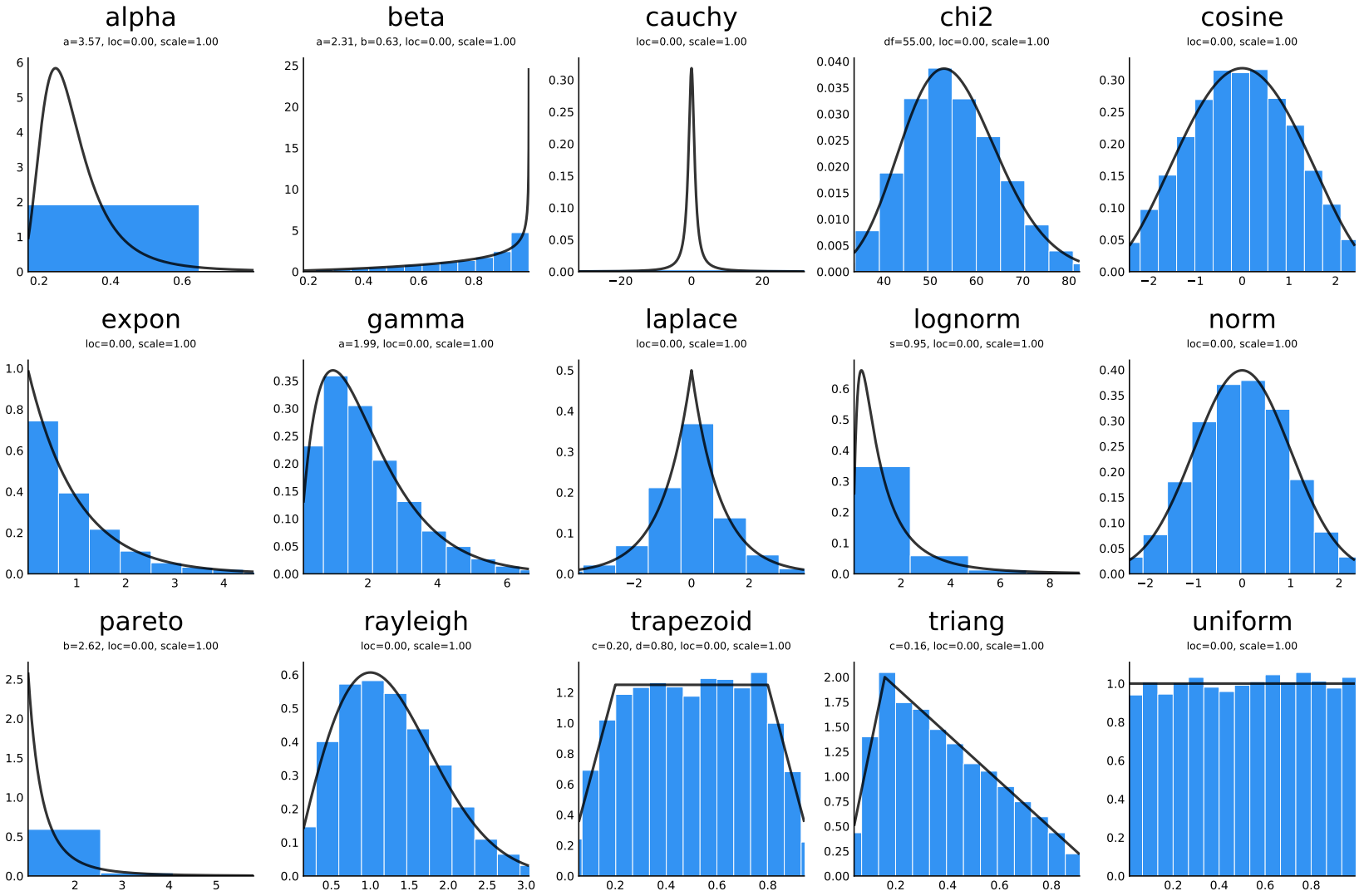


Figure 3: Plot of various pdfs available in SciPy - see `distributions.py`

We can condition random variables on random events

Conditional distribution function

The **conditional distribution function of X given event B** is

$$F_X(x|B) = \frac{P[X \leq x \text{ and } B]}{P[B]} \tag{21}$$

Conditional density function

The **conditional density function of X given event B** is

$$f_X(x|B) = \frac{d}{dx} F_X(x|B) \tag{22}$$

Just as we had the joint probability of two events, we have the joint distribution of two random variables

Joint distribution function

The **joint (cumulative) distribution function** of X and Y is

$$F_{XY}(x, y) = P[X \leq x \text{ and } Y \leq y] \tag{23}$$

Joint probability mass function

The **joint probability mass function** of X and Y is

$$P_{XY}(x, y) = P[X = x, Y = y] \tag{24}$$

Joint density function

The **joint density function** of X and Y is

$$f_{XY}(x, y) = \frac{\partial}{\partial x}\frac{\partial}{\partial y}F_{XY}(x, y) \tag{25}$$

Multinomial distribution

Here is an example of a joint distribution

Idea: Generalize the binomial distribution to trials with more than two outcomes

Consider the multinomial experiment with n independent trials with m outcomes, with each trial governed by a discrete RV with success probabilities $\{p_i\}_{i=1}^m$.

The number of times each outcome happens throughout the entire experiment is a discrete RV X_i for $i = 1, \ldots, m$.

We are interested in the probability that the ith outcome appears exactly k_i times i.e. the joint distribution of the X_i.

The **multinomial coefficient** is the number of ways that the ith outcome appears exactly k_i times (irrespective of their ordering):

$$\frac{n!}{k_1!k_2!\cdots k_m!} \tag{26}$$

The probability of an experiment with the ith outcome appearing exactly k_i times (irrespective of their ordering) is

$$\prod_{i=1}^{m} p_i^{k_i} \tag{27}$$

Since there are $\frac{n!}{k_1!k_2!\cdots k_m!}$ ways in which the experiment could end with the ith outcome appearing exactly k_i times, the probability of seeing such an experiment is

$$\frac{n!}{k_1!k_2!\cdots k_m!}\prod_{i=1}^{m} p_i^{k_i} \tag{28}$$

Multinomial distribution

A collection of RVs $\{X_i\}_{i=1}^m$ follows a **multinomial distribution** if it represents the multinomial experiment, whose joint pmf is

$$P[X_1 = k_1, X_2 = k_2, \ldots, X_m = k_m] \tag{29}$$

$$= \begin{cases} \dfrac{n!}{k_1! k_2! \cdots k_m!} \displaystyle\prod_{i=1}^m p_i^{k_i} & \text{if } \displaystyle\sum_{i=1}^m k_i = n, \\ 0 & \text{otherwise} \end{cases} \tag{30}$$

where $\{p_i\}_{i=1}^m$ is a set of parameters representing the success probabilities, and must satisfy $\sum_{i=1}^m p_i = 1$.

Exercise: As a special case, how can we recover the binomial distribution from the multinomial distribution?

Marginal distributions

If we have a joint distribution in hand, we can get the distribution of each of the components by integrating ("marginalizing")

Marginal density function

The **marginal density functions** are

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y)dy \tag{31}$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y)dx \tag{32}$$

Marginal distribution function

The **marginal distribution functions** are

$$F_X(x) = F_{XY}(x, \infty) = \int_{-\infty}^{x} f_X(\xi)d\xi \tag{33}$$

$$F_Y(y) = F_{XY}(\infty, y) = \int_{-\infty}^{y} f_Y(\eta)d\eta \tag{34}$$

We can also condition random variables on other random variables

Conditional density function

The **conditional density function of X given Y** is

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} \tag{35}$$

Conditional distribution function

The **conditional distribution function of X given Y** is

$$F_{X|Y}(x|y) = P[X \leq x|Y \leq y] = \int_{-\infty}^{x} f_{X|Y}(\xi|y)d\xi \tag{36}$$

Notice that $F_{X|Y}(x|y) \neq \frac{F_{XY}(x,y)}{F_Y(y)}$ (unlike the conditional pdf)

Sum Rule for random variables

Let X and Y be two discrete random variables.

The probability that X takes the value x_i, irrespective of the value of Y, is the **total probability** of $X = x_i$, written as $P[X = x_i]$.

Sum Rule for random variables

The total probability of X can be computed as

$$P[X = x_i] = \sum_j P[X = x_i | Y = y_j] P[Y = y_j] \tag{37}$$

$$= \sum_j P[X = x_i, Y = y_j]. \tag{38}$$

This follows from the law of total probability for the event $X = x_i$ and the fact that all the events $Y = y_j$ partition the sample space of Y.

The **total probability** is also referred to as the **marginal probability**, *since we are marginalizing out* the other variable, Y.

Let X and Y be two discrete random variables.

Conditional probability (Again!)

For only the instances for which $A = a_i$, the fraction of such instances for which $B = b_j$ is $P[B = b_j | A = a_i]$ and are called the **conditional probability of $B = b_j$ given $A = a_i$**.

Product Rule for random variables

The joint pmf of X and Y can be computed as

$$P[X = x_i, Y = y_j] = P[Y = y_j | X = x_i] P[X = x_i] \tag{39}$$

Bayes' theorem for random variables

RV conditioned on RV

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \tag{40}$$

Event conditioned on RV

$$P[A|X=x] = \frac{f_{X|A}(x)P[A]}{f_X(x)} \tag{41}$$

RV conditioned on event

$$f_{Y|A}(y) = \frac{P[A|Y=y]f_Y(y)}{P[A]} \tag{42}$$

Independent random variables

Two random variables X and Y are **statistically independent** if the two events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent for any pair (x, y).

Equivalently,

$$F_{XY}(x, y) = F_X(x)F_Y(y) \tag{43}$$

or

$$f_{XY}(x, y) = f_X(x)f_Y(y) \tag{44}$$

You can imagine the generalization to more than two RV's - joint distribution is equal to product of the marginals

It is nice when RV's are independent because it makes computing their joint distribution trivial - just multiply the marginals!

Functions of random variables

Core problem:
What is the distribution of a function of a random variable?

Math:
Given $f_X(x)$ and $Y = g(X)$, what is $f_Y(y)$?

"Indirect" procedure:

1. Find the point set C_y such that $\{Y \leq y\} = \{X \in C_y\}$
2. Find the cdf of Y as

$$F_Y(y) = P[Y \leq y] = P[g(X) \leq y] = P[X \in C_y] \tag{45}$$

3. Find the pdf of Y as

$$f_Y(y) = \frac{d}{dy} F_Y(y) \tag{46}$$

Example: affine function of a random variable

Suppose g is affine, i.e. $Y = g(X) = aX + b$.

Case 1: $a > 0$

Step 1: Find the point set

$$\{Y \leq y\} = \{aX + b \leq y\} \tag{47}$$

$$= \left\{X \leq \frac{y-b}{a}\right\} = \{X \in C_y\} \tag{48}$$

Step 2: Find the cdf

$$F_Y(y) = P[Y \leq y] = P[aX + b \leq y] \tag{49}$$

$$= P\left[X \leq \frac{y-b}{a}\right] = F_X\left(\frac{y-b}{a}\right) \tag{50}$$

Step 3: Differentiate cdf to get pdf
Use the change of variables $z = \frac{y-b}{a}$ so

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X\left(\frac{y-b}{a}\right) \tag{51}$$

$$= \frac{dF_X(z)}{dz} \cdot \frac{dz}{dy} \quad \text{(chain rule)}$$

$$= f_X(z) \cdot \frac{1}{a} \tag{52}$$

Optional Exercise: Work out Case 2: $a < 0$

After doing that, you will find the solution is

$$\boxed{f_Y(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right) \quad \text{if } a \neq 0} \tag{53}$$

Optional Exercise: Work out Case 3: $a = 0$ (degenerate case)
Hint: The solution is trivial: $f_Y(y) = \delta(y-b)$, a Dirac delta at b.

Example: nonlinear function of RV

Example 3.2-8 in [1]

Consider the vertical coordinate of a spinner with uniform random angle

$$g(X) = \sin(X) \quad \text{(sine map)} \tag{54}$$

$$f_X(x) = \begin{cases} \dfrac{1}{2\pi} & \text{if } -\pi \leq X \leq \pi \\ 0 & \text{else} \end{cases} \quad \text{(uniform distribution)} \tag{55}$$

Example: nonlinear function of RV

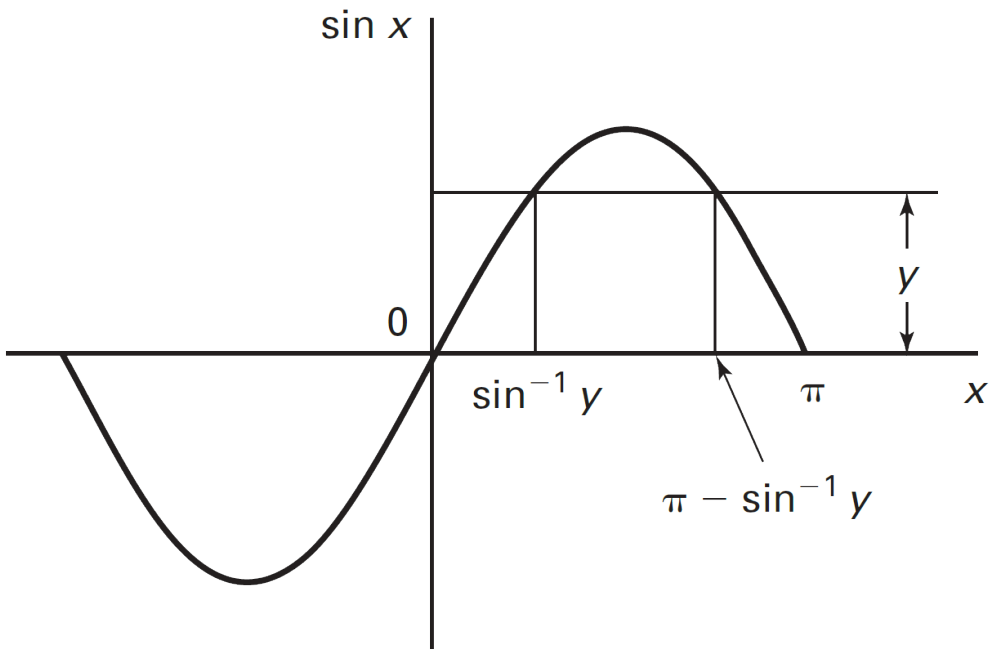
Case 1: $0 \leq y < 1$

Step 1: Find the point set (this time it's trickier)

$$\{Y \leq y\} = \{\sin(X) \leq y\} \tag{56}$$

$$= \left\{-\pi < X \leq \sin^{-1}(y)\right\} \cup \left\{\pi - \sin^{-1}(y) < X \leq \pi\right\} \tag{57}$$

$$= \{X \in C_y\} \tag{58}$$



Example: nonlinear function of RV

Step 2: Find the cdf

$$
\begin{aligned}
F_Y(y) &= P[Y \le y] && (59)\\
&= P\left[\left\{-\pi < X \le \sin^{-1}(y)\right\} \cup \left\{\pi - \sin^{-1}(y) < X \le \pi\right\}\right] && (60)\\
&= P\left[-\pi < X \le \sin^{-1}(y)\right] + P[\pi - \sin^{-1}(y) < X \le \pi] && (61)\\
&= \left[F_X(\sin^{-1}(y)) - F_X(-\pi)\right] + \left[F_X(\pi) - F_X(\pi - \sin^{-1}(y))\right] && (62)
\end{aligned}
$$

Step 3: Differentiate cdf to get pdf

$$
\begin{aligned}
f_Y(y) &= \frac{d}{dy} F_Y(y) && (63)\\
&= f_X\left(\pi - \sin^{-1} y\right) \frac{1}{\sqrt{1-y^2}} + f_X\left(\sin^{-1} y\right) \frac{1}{\sqrt{1-y^2}} && (64)\\
&= \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-y^2}} \quad \text{for } 0 \le y < 1 && (65)
\end{aligned}
$$

Optional Exercise: Work out Case 2: $-1 < y \leq 0$
Hint: You should find the pdf is the same as for $0 \leq y < 1$

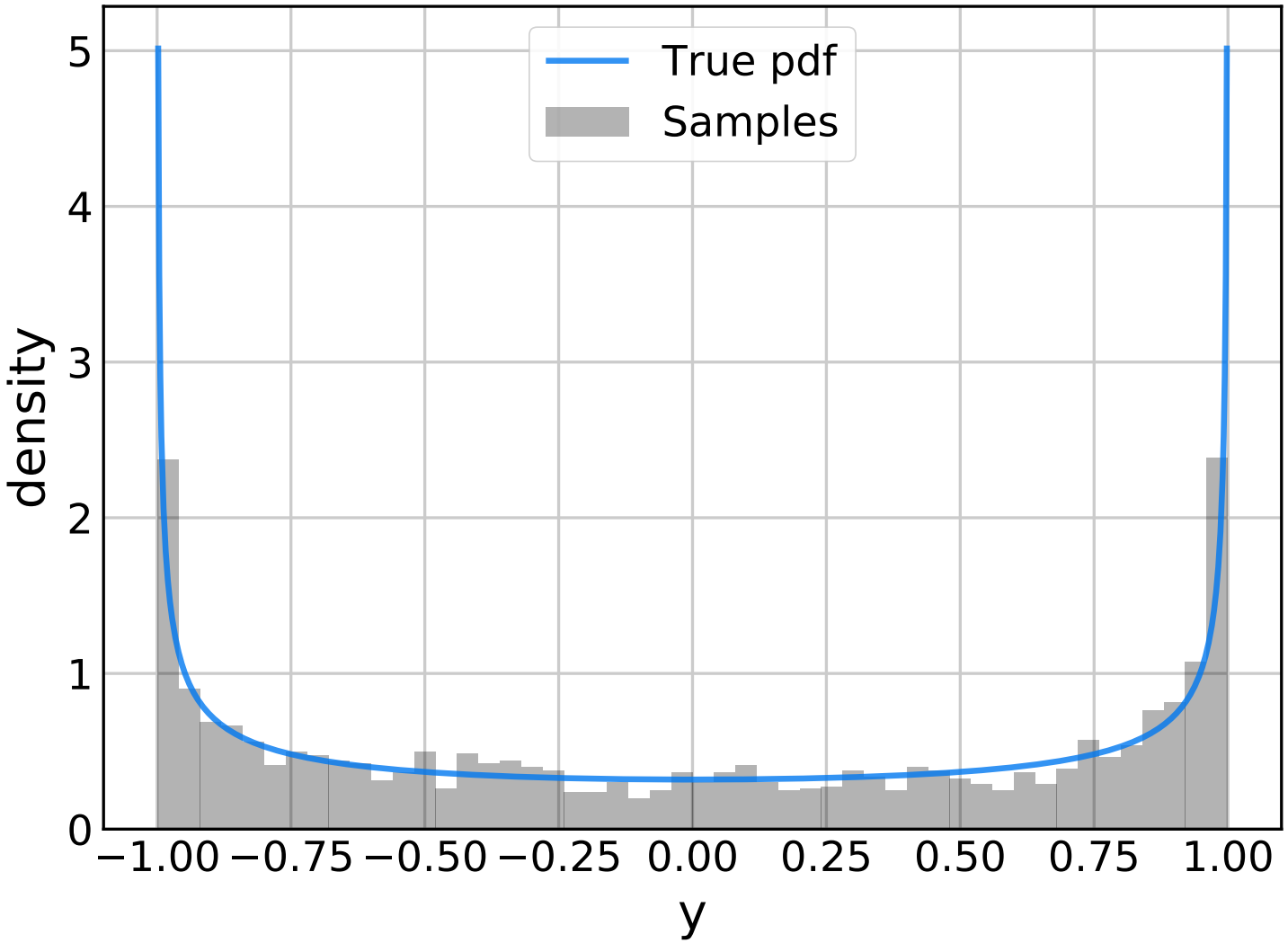
Optional Exercise: Work out Case 3: $|y| >= 1$
Hint: You should find the cdf is constant with respect to y *(either* $F_Y(y) = 0$ *or* $F_Y(y) = 1$*) and therefore the pdf is zero.*

Therefore, the complete solution is

$$\boxed{f_Y(y) = \begin{cases} \frac{1}{\pi} \cdot \frac{1}{\sqrt{1-y^2}} & \text{if } |y| < 1 \\ 0 & \text{else} \end{cases}} \tag{66}$$

Example: nonlinear function of RV

We can check our solution against a histogram of empirical samples - see `function_of_rv.py`



Can we go directly from pdf of X to pdf of $Y = g(X)$ (without finding intermediate cdf)?

"Direct" procedure:

1. Find the root functions $x_i = x_i(y)$ that satisfy $y - g(x_i) = 0$ for any fixed y
2. Compute derivative $g'(x)$
3. Evaluate $|g'(x_i)|$ check $|g'(x_i)| \neq 0$
4. Compute the pdf directly as

$$f_Y(y) = \sum_i \frac{f_X(x_i)}{|g'(x_i)|} \tag{67}$$

Note: Throughout keep in mind that $x_i = x_i(y)$ are functions!

Example revisited: nonlinear function of RV

Example 3.2-9 in [1]

Consider again the problem

$$g(X) = \sin(X) \quad \text{(sine map)} \tag{68}$$

$$f_X(x) = \begin{cases} \frac{1}{2\pi} & \text{if } -\pi \leq X \leq \pi \\ 0 & \text{else} \end{cases} \quad \text{(uniform distribution)} \tag{69}$$

Case 1: $0 \leq y < 1$

Step 1:
For any $0 \leq y < 1$ we have the roots of

$$y - g(x) = y - \sin(x) = 0 \tag{70}$$

are

$$x_1 = \sin^{-1}(y) \quad \text{and} \quad x_2 = \pi - \sin^{-1}(y) \tag{71}$$

Example revisited: nonlinear function of RV

Step 2:
We have the derivative

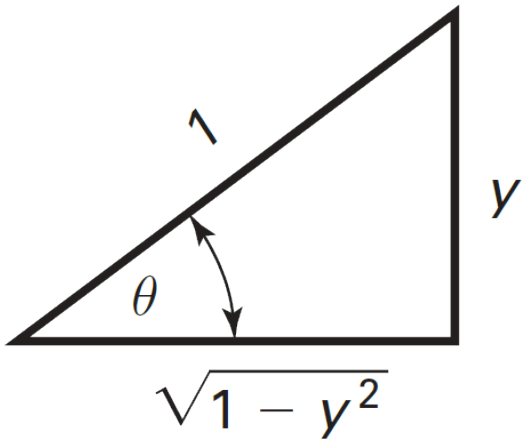
$$\frac{dg}{dx} = \cos(x) \tag{72}$$

Step 3:
Evaluated at the roots, the derivative is

$$\left.\frac{dg}{dx}\right|_{x_1} = \cos(\sin^{-1}(y)), \qquad \left.\frac{dg}{dx}\right|_{x_2} = -\cos(\sin^{-1}(y)) \tag{73}$$

Example revisited: nonlinear function of RV

When you see the **composition of trig and inverse trig**, there is usually a nice simplification to make - use triangle diagram to help



$$\sin(\theta) = \frac{y}{1} \tag{74}$$

$$\theta = \sin^{-1}(y) \tag{75}$$

$$\cos(\theta) = \frac{\sqrt{1-y^2}}{1} \tag{76}$$

$$\cos(\sin^{-1}(y)) = \sqrt{1-y^2} \tag{77}$$

We have the absolute values

$$\left|\frac{dg}{dx}\right|_{x_1} = \left|\frac{dg}{dx}\right|_{x_2} = \sqrt{1-y^2} \neq 0 \text{ for } 0 \leq y < 1 \tag{78}$$

Step 4:
Compute the pdf

$$f_Y(y) = \sum_i \frac{f_X(x_i)}{|g'(x_i)|} \tag{79}$$

$$= \frac{\frac{1}{2\pi}}{\sqrt{1-y^2}} + \frac{\frac{1}{2\pi}}{\sqrt{1-y^2}} \tag{80}$$

$$= \frac{1}{\pi}\sqrt{1-y^2} \quad \text{for } 0 \leq y < 1 \tag{81}$$

which is the same result as we got using the "indirect" method.

Optional Exercise: Repeat the procedure for Case 2: $-1 < y \leq 0$

Optional Exercise: Repeat the procedure for Case 3: $|y| \geq 1$

Core problem:
What is the distribution of a function of a random variable?

Math:
Given $f_{XY}(x,y)$ and $Z = g(X,Y)$, what is $f_Z(z)$?

"Indirect" procedure:

1. Find the point set C_z such that $\{Z \leq z\} = \{(X,Y) \in C_z\}$
2. Find the cdf of Z as

$$F_Z(z) = \iint_{(x,y)\in C_z} f_{XY}(x,y)dxdy \tag{82}$$

3. Find the pdf of Z as

$$f_Z(z) = \frac{d}{dz}F_Z(z) \tag{83}$$

Optional Exercise: Find $f_Z(z)$ where $Z = XY$
Hint: See Example 3.3-1 in [1]
Solution:

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|y|} f_{XY}(z/y, y) dy \tag{84}$$

Optional Exercise: Find $f_Z(z)$ where $Z = X + Y$ Eqs. (3.3-13), (3.3-14) in [1]

Solution:

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(z - y, y)dy \tag{85}$$

If X and Y are independent

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z - x)dx = \int_{-\infty}^{\infty} f_X(z - y)f_Y(y)dy \tag{86}$$

which is a **convolution integral**

Evaluate by reversing one function and sliding it

See Examples 3.3-4, 3.3-5, 3.3-6, 3.3-7, 3.3-8 in [1]

Homework: Max of two RVs

Homework P2-4: Find $f_Z(z)$ where $Z = \max(X, Y)$ and X, Y are independent.
Hint: See Example 3.3-2 in [1]

Bibliography I

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

Expectation and Moments

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

Expectation and moments

Expectation

The **expectation** or **mean** of a random variable X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx \tag{1}$$

The expectation of a function of a random variable $g(X)$ is

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx \tag{2}$$

If the RV is discrete, these integrals become simple sums:

$$\mathbb{E}[X] = \sum_i x_i P_X(x_i) \tag{3}$$

$$\mathbb{E}[g(X)] = \sum_i g(x_i) P_X(x_i) \tag{4}$$

Linearity of expectation

Expectation is a **linear operator** - follows from linearity of integration

$$\mathbb{E}[X+Y] \tag{5}$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x+y)f_{XY}(x,y)dxdy \tag{6}$$

$$= \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}xf_{XY}(x,y)dxdy + \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}yf_{XY}(x,y)dxdy \tag{7}$$

$$= \int_{-\infty}^{+\infty}x\left(\int_{-\infty}^{+\infty}f_{XY}(x,y)dy\right)dx + \int_{-\infty}^{+\infty}y\left(\int_{-\infty}^{+\infty}f_{XY}(x,y)dx\right)dy \tag{8}$$

$$= \int_{-\infty}^{+\infty}xf_X(x)dx + \int_{-\infty}^{+\infty}yf_Y(x)dy \tag{9}$$

$$= \mathbb{E}[X] + \mathbb{E}[Y] \tag{10}$$

Use induction to conclude the linearity property

$$\mathbb{E}\left[\sum_{i=1}^{N}X_i\right] = \sum_{i=1}^{N}\mathbb{E}\left[X_i\right] \tag{11}$$

Expectation of a Gaussian

Recall the Gaussian random variable $X \sim \mathcal{N}(\mu, \sigma^2)$.

Let's show the mean is μ using the change of variable $z = \frac{x-\mu}{\sigma}$

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot f_X(x) dx \tag{12}$$

$$= \int_{-\infty}^{\infty} x \cdot \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) dx \tag{13}$$

$$= \int_{-\infty}^{\infty} (\sigma z + \mu) \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz \tag{14}$$

$$= \frac{\sigma}{\sqrt{2\pi}} \underbrace{\int_{-\infty}^{\infty} z \cdot \exp\left(-\frac{1}{2}z^2\right) dz}_{=0 \text{ because integrand odd}} + \mu \underbrace{\left[\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) dz\right]}_{=1 \text{ because } P[Z \leq \infty] = 1} \tag{15}$$

$$= \mu \tag{16}$$

Conditional expectation

The **conditional expectation** of random variable Y given event B has occurred is

$$\mathbb{E}[Y|B] = \int_{-\infty}^{\infty} y f_{Y|B}(y|B) dy \tag{17}$$

The **conditional expectation** of random variable Y conditioned on random variable X is

$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \tag{18}$$

We have a **law of total expectation** (like law of total probability)

$$\mathbb{E}[Y] = \int_{-\infty}^{\infty} \mathbb{E}[Y|X = x] f_X(x) dx \tag{19}$$

Moments are expectations of monomials of (shifted and scaled) RVs

Moments

The k^{th} **(raw) moment** of X is

$$m_k = \mathbb{E}[X^k] \tag{20}$$

The k^{th} **central moment** of X is

$$c_k = \mathbb{E}[(X - \mathbb{E}[X])^k] \tag{21}$$

The k^{th} **standardized moment** of X is

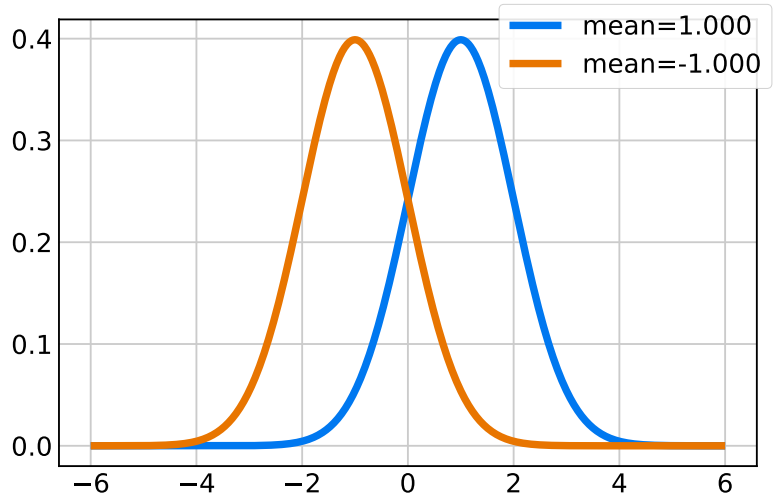
$$s_k = \frac{\mathbb{E}[(X - \mathbb{E}[X])^k]}{\mathbb{E}[(X - \mathbb{E}[X])^2]^{k/2}} = \frac{c_k}{c_2^{k/2}} \tag{22}$$

Moments summarize different aspects of the **shape** of a distribution

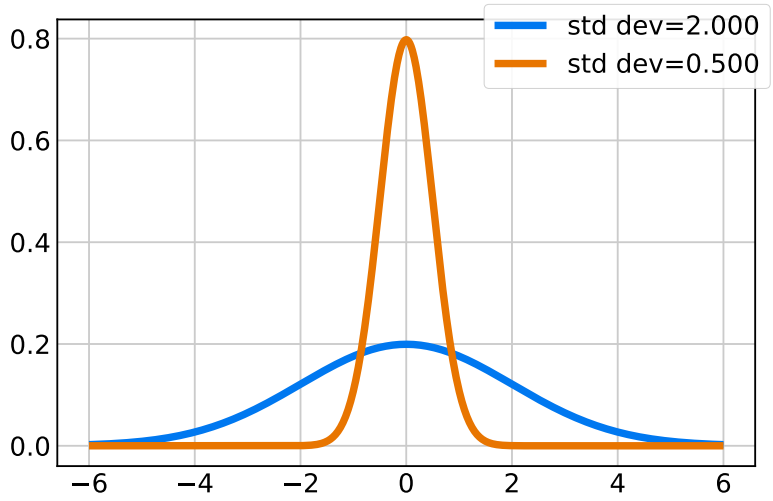
Name	Definition	Intuition
Mean	$\mu = m_1$	Location or center
Variance	$\sigma^2 = c_2$	Dispersion or spread
Std deviation	$\sigma = \sqrt{\sigma^2}$	Dispersion or spread
Skewness	s_3	Asymmetry or tilt
Kurtosis	s_4	Heaviness of tails

See `moments.py`

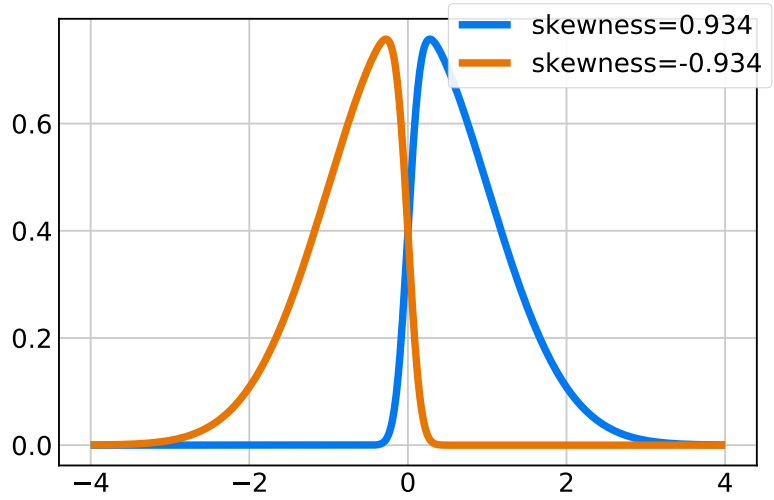
Comparison of pdfs with different moments



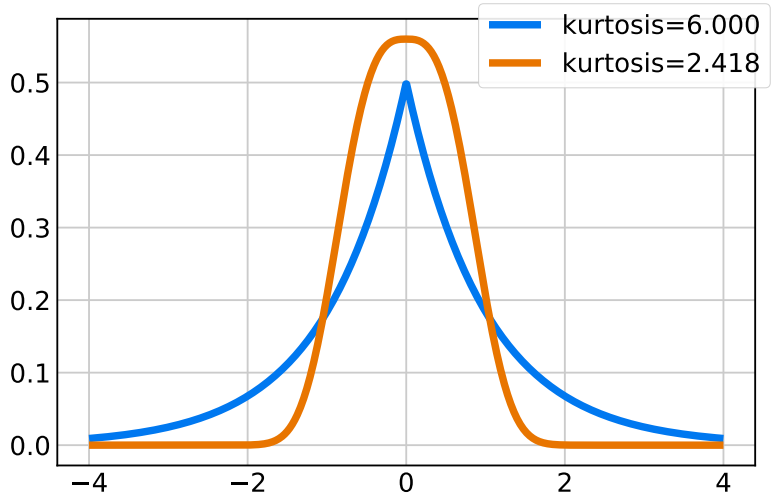
(a) Mean



(b) Standard deviation



(c) Skewness



(d) Kurtosis

We can convert between raw and central moments

Example: Second moment

$$
\begin{aligned}
c_2 &= \mathbb{E}[(X - \mathbb{E}[X])^2] && (23)\\
&= \mathbb{E}[X^2 - 2\mathbb{E}[X]X + \mathbb{E}[X]^2] && (24)\\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && \text{(linearity of } \mathbb{E}[\cdot]) \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2 && (25)\\
&= m_2 - m_1^2 && (26)
\end{aligned}
$$

This relation generalizes to higher-order moments as

$$
c_k = \sum_{i=0}^{k} \binom{k}{i} (-1)^i \mu^i m_{k-i} \tag{27}
$$

Homework: Moments of a Gaussian

Homework P3-1:
Verify the expression for the variance of a Gaussian.
Hint: See Example 4.1-7 in [1]

Optional Exercise:
Find expressions for all moments of a Gaussian.
Hint: See e.g. https://arxiv.org/abs/1209.4340

Basic probability bounds

Often we want to bound the probability of certain events or random variables without having to specify/compute their distribution

c.f. the first several pages of Wainwright's book [2]

Markov inequality

Given a non-negative random variable X with finite mean, we have

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t} \quad \text{for all } t > 0 \tag{28}$$

"X is probably small when its mean is small"

The most basic tail bound.

Basis for several "classical" concentration inequalities.

Chebyshev inequality

Given a random variable X with finite mean μ and variance σ^2, we have

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\sigma^2}{t^2} \quad \text{for all } t > 0 \tag{29}$$

"X is probably close to its mean whenever its variance is small"

The most basic concentration inequality.

Proof: Follows by applying Markov inequality to the non-negative random variable $(X - \mu)^2$.

Moment bound

Given a non-negative random variable X with finite moments up to order k, we have

$$\mathbb{P}[|X - \mu| \geq t] \leq \frac{\mathbb{E}\left[|X - \mu|^k\right]}{t^k} \qquad \text{for all } t > 0 \tag{30}$$

Proof: Follows by applying Markov inequality to the random variable $|X - \mu|^k$

Chernoff bound

Given a non-negative random variable X with a moment generating function in a neighborhood of zero, we have

$$\mathbb{P}[X \geq 0] \leq \inf_{\theta > 0} \mathbb{E}\left[e^{\theta X}\right] \tag{31}$$

Proof: Follows by applying Markov inequality to the random variable $e^{\theta(X-\mu)}$ and optimizing over θ.

The moment bound with an optimal choice of k is never worse than the Chernoff bound.

Nonetheless, the Chernoff bound is most widely used in practice, possibly due to the ease of manipulating moment generating functions.

Homework P3-2:
Compare the Markov inequality bound with the exact tail probability from the exponential cdf with parameter $\lambda = 1$; compute the probability bounds at the level $t = 2$. How bad is the Markov bound compared with the exact tail probability?
Hint: The mean of an exponential random variable is $\mu = 1/\lambda$.

Homework P3-3:
Compare the Chebyshev inequality bound with the exact tail bound from the standard normal cdf; compute the probability bounds at the level $t = 2$. How bad is the Chebyshev bound compared with the exact concentration probability?
Hint: The standard normal cdf does not have a closed-form expression, so either use the `cdf()` method of `scipy.stats.norm` or a table of the standard normal cdf to get the exact value. In case you run into issues, $\Phi(2) = 1 - \Phi(-2) = 0.9772$.

Joint moments summarize different aspects of the shape of a joint distribution

Joint moments

The ij**th (raw) joint moment** of random variables X and Y is

$$m_{ij} = \mathbb{E}[X^i Y^j] \tag{32}$$

The ij**th central joint moment** of random variables X and Y is

$$c_{ij} = \mathbb{E}[(X - \mathbb{E}[X])^i (Y - \mathbb{E}[Y])^j] \tag{33}$$

Some joint moments have special, confusing names

The **correlation** is

$$m_{11} = \mathbb{E}[XY] \tag{34}$$

The **covariance** is

$$c_{11} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \tag{35}$$

The **correlation coefficient** is

$$\rho = \frac{c_{11}}{\sqrt{c_{02}c_{20}}} \tag{36}$$

Homework P3-4:
Prove the relation

$$m_{11} = c_{11} + \mathbb{E}[X]\mathbb{E}[Y]$$

Hint: It is similar to the earlier second moment relation $m_2 = c_2 + m_1^2$

Homework P3-5:
When are the correlation and covariance equal?
Hint: Use the relation $m_{11} = c_{11} + \mathbb{E}[X]\mathbb{E}[Y]$ *you just proved.*

Homework P3-6:
Prove that $\rho \in [-1, 1]$
Hint: See Ch. 4.3 of [1]

Uncorrelated and orthogonal random variables

Uncorrelated random variables

Two random variables are **uncorrelated** if their **covariance** is zero.

Orthogonal random variables

Two random variables are **orthogonal** if their **correlation** is zero.

- Yes I know the terminology is confusing :/

Homework: Uncorrelated random variables

Homework P3-7:
Prove that if X and Y are uncorrelated, then $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y$
i.e. "the variance of the sum is the sum of the variances."
Hint: Use linearity of expectation.

Homework P3-8:
Prove that if X and Y are independent, then they are uncorrelated.
Remark: The converse does not hold unless X and Y are both Gaussian.

Homework P3-9:
Under what condition(s) can a pair of uncorrelated random variables be orthogonal?
Hint: This is a special case of one of the earlier exercises.

Random vectors

Random vector

A **random vector** is a vector of random variables.

The **cdf** of a random vector is defined as

$$F_X(x) = \mathbb{P}[X_1 \leq x_1 \text{ and } X_2 \leq x_2 \text{ and } \ldots X_n \leq x_n] \tag{37}$$

The **pdf** is defined as

$$f_X(x) = \frac{\partial^n F_X(x)}{\partial x_1 \partial x_2 \cdots \partial x_n} \tag{38}$$

Similar definitions for joint, marginal, and conditional distributions

- See Ch. 5.1 of [1]

The **expectation** of a random vector X is the vector μ_X with entries

$$[\mu_X]_i = \mathbb{E}[X]_i = \int_{-\infty}^{\infty} x_i f_{X_i}(x_i) dx_i \tag{39}$$

where $f_{X_i}(x_i)$ is the ith marginal pdf.

Moments are defined similarly as with random variables.

(Auto)-covariance matrix of X

$$K_X = \mathbb{E}[(X - \mu_X)(X - \mu_X)^\intercal] \tag{40}$$

(Cross)-covariance matrix between X and Y

$$C_{XY} = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)^\intercal] \tag{41}$$

We can gather these up into the block covariance matrix

$$D_{XY} = \begin{bmatrix} K_X & C_{XY} \\ C_{XY}^\intercal & K_Y \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix}\begin{bmatrix} X - \mu_X \\ Y - \mu_Y \end{bmatrix}^\intercal\right] \tag{42}$$

(Auto)-correlation matrix of X

$$R_X = \mathbb{E}[XX^\intercal] \succeq 0 \tag{43}$$

(Cross)-correlation matrix between X and Y

$$S_{XY} = \mathbb{E}[XY^\intercal] \tag{44}$$

We can gather these up into the block correlation matrix

$$B_{XY} = \begin{bmatrix} R_X & S_{XY} \\ S_{XY}^\intercal & R_Y \end{bmatrix} = \mathbb{E}\left[\begin{bmatrix} X \\ Y \end{bmatrix}\begin{bmatrix} X \\ Y \end{bmatrix}^\intercal\right] \tag{45}$$

Homework 3-10:
Prove the identity between covariance and correlation matrices

$$R = K + \mu\mu^\intercal \tag{46}$$

Hint: Use linearity of expectation.

Homework 3-11:
Write an expression for D in terms of B, μ_X, μ_Y.
Hint: It follows immediately from $R = K + \mu\mu^\intercal$ by stacking X and Y.

Homework 3-12:
Prove that $R \succeq K \succeq 0$ and $B \succeq D \succeq 0$ where $A \succeq B$ means $A - B$ is symmetric positive semidefinite.
Hint: It follows by the above relations and the property of outer product matrices $AA^\intercal \succeq 0$ for any matrix A, and taking $A = \mu$.

Uncorrelated and orthogonal random vectors

A random vector X is **uncorrelated** with itself if K is diagonal.

A random vector X is **orthogonal** with itself if R is diagonal.

Two random vectors X and Y are **uncorrelated** if $C = 0$.

Two random vectors X and Y are **orthogonal** if $S = 0$.

Optional Exercise:
Think about how these expressions can be summarized in terms of the block matrices C and D.

Optional Exercise:
Under what condition(s) can a pair of uncorrelated random vectors be orthogonal?
Hint: You already solved this in the scalar case.

Whitening transformation

Sometimes we need to get a standardized version of a random variable

In the scalar case we used the standardizing transform

$$Z = \frac{X - \mu}{\sigma} \tag{47}$$

- Subtract out the mean and normalize by the standard deviation, so Z has zero mean and variance one
- Need to assume $\sigma > 0$ for non-degeneracy

The **whitening transformation** is the multivariate generalization of the scalar standardizing transform

- Based on the eigen-decomposition of the covariance matrix

The **whitening transformation** is

$$Z = \Lambda_X^{-1/2} U_X^{\intercal} (X - \mu) \tag{48}$$

- Subtract the mean out and normalize, so Z has zero mean and identity auto-covariance
- Λ_X is a diagonal matrix whose entries are the n eigenvalues of K_X
  - The eigenvalues λ_i are real numbers since K_X is symmetric
  - Need to assume $\lambda_i > 0$ for $i = 1, \ldots, n$ for non-degeneracy
    - Equivalent to assuming K_X full rank
  - $\Lambda_X^{-1/2}$ is diagonal with entries $\lambda_i^{-1/2}$
- U_X is an orthogonal matrix whose columns are n eigenvectors of K_X

Coloring transformation

Sometimes we need to get a a random vector Y with nonzero mean μ_Y and non-identity covariance K_Y from a white random vector

- Inverse operation of the whitening transformation

The **coloring transformation** is

$$Y = U_Y \Lambda_Y^{1/2} X + \mu \tag{49}$$

- Λ_Y is a diagonal matrix whose entries are the n eigenvalues of K_Y
- U_Y is an orthogonal matrix whose columns are n eigenvectors of K_Y

The n-**dimensional multivariate Gaussian** pdf is

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det(K)}} \exp\left[-\frac{1}{2}(x-\mu)^\intercal K^{-1}(x-\mu)\right] \tag{50}$$

- Mean is $\mu \in \mathbb{R}^n$
- Covariance is $K \in \mathbb{R}_+^{n \times n}$

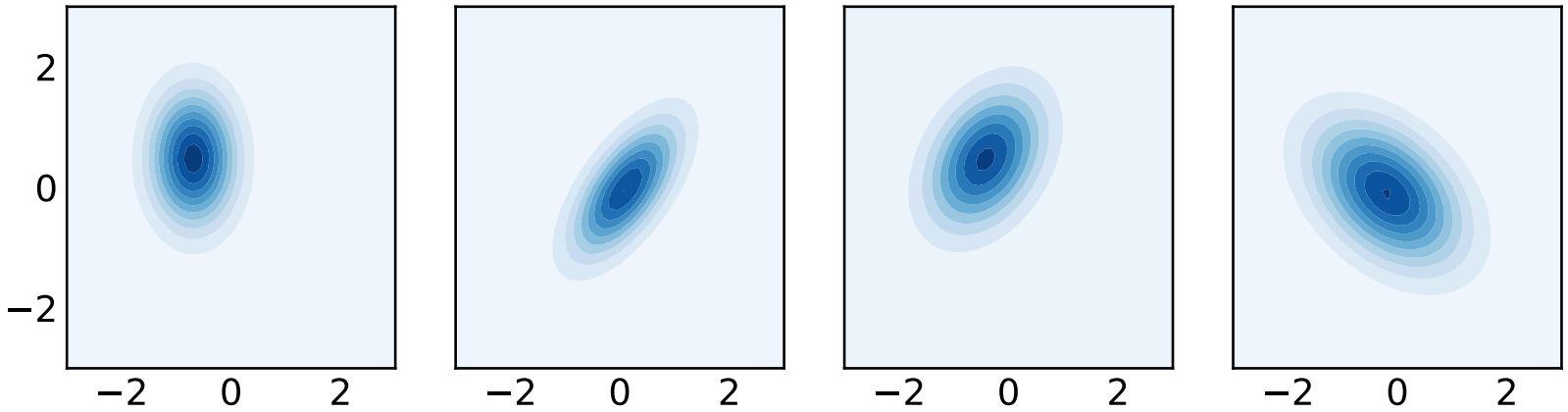


Figure 2: Various multivariate Gaussian pdfs for $n = 2$.
See `multivariate_gaussian.py`

Properties of multivariate Gaussians

Gaussians are extremely special distributions with nice properties

- Marginals of a Gaussian are Gaussian
- Gaussians conditioned on Gaussians are Gaussian
- Any affine transformation of a Gaussian is Gaussian
- All pertinent information about a Gaussian is encoded in the mean and covariance
- Sums of random vectors tend towards a Gaussian (central limit theorem, coming up)

Homework 3-13:
What is the pdf of a white (zero mean and identity covariance) multivariate Gaussian random vector X? Can it be expressed in terms of the marginal densities of each component of X? If so, write the expression. Are the components of X statistically independent?

Bibliography I

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

[2] Martin J Wainwright.
High-dimensional statistics: A non-asymptotic viewpoint, volume 48.
Cambridge University Press, 2019.
https://people.eecs.berkeley.edu/~wainwrig/BibPapers/Wai19.pdf.

Parameter Estimation

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

1. Parameter estimation
2. Laws of large numbers
3. Central limit theorem

Parameter estimation

Parameter estimation

In many applications:

- Distribution of a random variable X is unknown or too complicated to compute
- Only need some parameter θ that characterizes the distribution

Goal: Obtain a good approximation of parameter θ based only on observations of X.

Estimator

An **estimator** $\hat{\Theta}$ is a function of the data $\{X_i\}$ that approximates θ, but is not an explicit function of θ.

Judging parameter estimators

How do we judge the quality of an estimator?

Consistency

An estimator $\hat{\Theta}_n$ computed from n samples is **consistent** if

$$\lim_{n \to \infty} P\Big[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \Big] = 0 \tag{1}$$

for any positive tolerance $\varepsilon > 0$.

Consistency means "we can guarantee arbitrarily accurate estimates if we use an arbitrarily large amount of data"

Judging parameter estimators

What we really want:

Confidence bound

An estimator $\hat{\Theta}_n$ is **ε-accurate with $1-\delta$ confidence** if

$$P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] \leq \delta \tag{2}$$

- This is like soft consistency w/ finite data
- Consistency allows us to take ε and δ as small as we like (so long as we can pay for it with infinite data $n \to \infty$)
- Quantifying n
  - Can be done exactly in certain special cases
    - e.g. estimating the mean of a Gaussian
  - Can be done conservatively using concentration inequalities in more general cases
    - e.g. estimating the mean of any distribution w/ finite variance

Confidence interval

Consider an estimator $\hat{\Theta}_n$. Fix the number of samples n and fix a failure probability δ. The $1-\delta$ **confidence interval** is the smallest accuracy tolerance ε such that

$$P\left[\ |\hat{\Theta}_n - \theta| > \varepsilon\ \right] \leq \delta \tag{3}$$

i.e. the estimator $\hat{\Theta}_n$ is ε-accurate with $1-\delta$ confidence.

Basically the same as the confidence criterion where we fixed ε and sought n, but here we fix n and seek ε

Judging parameter estimators

Many classical results use two proxies for the ε-δ criterion:

- Bias
  - "systematic errors"
  - "location"
- Variance
  - "random errors"
  - "spread"

Bias

The **bias** of an estimator $\hat{\Theta}$ is

$$|\mathbb{E}[\hat{\Theta}] - \theta|. \tag{4}$$

The estimator is **unbiased** if

$$\mathbb{E}[\hat{\Theta}] = \theta. \tag{5}$$

Variance

The **variance** of an estimator $\hat{\Theta}$ is

$$\mathbb{E}[(\hat{\Theta} - \theta)^2]. \tag{6}$$

The estimator is **minimum variance** if

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}}\ \mathbb{E}[(\Theta - \theta)^2]. \tag{7}$$

Judging parameter estimators

Sometimes bias can be eliminated without affecting the variance

- We will see an example of such a correction

Sometimes bias can only be reduced at the expense of higher variance

- In machine learning this is a well-studied phenomenon called the **bias-variance tradeoff**

Sample average estimator of a RV

The **sample average estimator** of a random variable X given N observations $\{X_i\}_{i=1}^N$ is

$$\hat{\mu}_X(n) := \frac{1}{N}\sum_{i=1}^{N} X_i$$

Sample average estimator of a function of a RV

The **sample average estimator** of a function g of a random variable X given N observations $\{X_i\}_{i=1}^N$ is

$$\hat{\mu}_{g(X)}(n) := \frac{1}{N}\sum_{i=1}^{N} g(X_i)$$

It's easy to show that the sample average is **unbiased**:

$$\begin{aligned}
\mathbb{E}\left[\hat{\mu}_X(n)\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] && \text{(def. of } \hat{\mu}_X(n)\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[X_i\right] && \text{(linearity of } \mathbb{E}[\cdot]\text{)} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu_X && \text{(def. of } \mu_X\text{)} \\
&= \frac{1}{n} \cdot n \cdot \mu_X && (8) \\
&= \mu_X && (9)
\end{aligned}$$

Properties of the sample average: variance

The **variance** of the sample average is not much harder to find:

$$\begin{aligned}
\sigma^2_{\hat{\mu}}(n) &:= \mathbb{E}\left[\left(\hat{\mu}_X(n) - \mathbb{E}\left[\hat{\mu}_X(n)\right]\right)^2\right] && \text{(def. of } \sigma^2_{\hat{\mu}}(n)\text{)} \\
&= \mathbb{E}\left[\left(\hat{\mu}_X(n) - \mu_X\right)^2\right] && \text{(since } \hat{\mu} \text{ unbiased)} \\
&= \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \mu_X\right)\right)^2\right] && \text{(def. of } \hat{\mu}\text{)} \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\left(X_i - \mu_X\right)^2\right] + \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\left(X_i - \mu_X\right)\left(X_j - \mu_X\right)\right] \\
& && \text{(expand squared sum)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\left(X_i - \mu_X\right)^2\right] + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n}\mathbb{E}\left[\left(X_i - \mu_X\right)\left(X_j - \mu_X\right)\right] \\
& && \text{(linearity of } \mathbb{E}[\cdot]\text{)} \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2_X + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{i\neq j}^{n}0 && \text{(def. of } \sigma^2_X\text{, uncorrelation of } X_i\text{)} \\
&= \sigma^2_X/n && \text{(10)}
\end{aligned}$$

Properties of the sample average: confidence

We can get a **confidence bound** by using the Chebyshev inequality:

$$P\left[|\hat{\mu}_X(n) - \mu_X| \geq \varepsilon\right] \leq \frac{\sigma^2_{\hat{\mu}}(n)}{\varepsilon^2} = \frac{1}{n} \cdot \frac{\sigma^2_X}{\varepsilon^2} \tag{11}$$

Taking $n \to \infty$ reveals that the **sample average is consistent**:

$$\lim_{n\to\infty} P\left[|\hat{\mu}_X(n) - \mu_X| \geq \varepsilon\right] = \lim_{n\to\infty} \frac{1}{n} \cdot \frac{\sigma^2_X}{\varepsilon^2} = 0 \tag{12}$$

Remark: If we knew the form of the distribution e.g. Gaussian we could get an exact confidence bound using the standard normal CDF.

Remark: This confidence bound involves the true variance σ^2_X, which is typically unknown. If X is Gaussian and σ^2_X is replaced by a sample variance estimate, an exact confidence bound can still be obtained using the **student T-distribution** CDF - see Ch. 6.3 of [1].

Sample variance

So far we estimated the mean - what about estimating the variance?

If we **knew the true mean μ** we could create the variance estimator

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \mu)^2 \tag{13}$$

But of course we **don't know the true mean μ!**

Natural idea: just use the sample mean in place of the true mean:

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \hat{\mu})^2 \tag{14}$$

But there is an issue with this...

Homework P4-1

Compute the expectation of the sample variance estimator

$$\hat{\sigma}_X^2(n) = \frac{1}{n}\sum_{i=0}^{n}(X_i - \hat{\mu}_X(n))^2 \tag{15}$$

where

$$\hat{\mu}_X(n) = \frac{1}{n}\sum_{i=0}^{n} X_i \tag{16}$$

1. Is this sample variance estimator $\hat{\sigma}_X^2(n)$ biased?
2. If so, how much is the bias?
3. How does the bias change with the number of samples n?
4. What correction needs to be made to $\hat{\sigma}_X^2(n)$ in order to make the estimator unbiased?

Maximum likelihood estimation provides a principled way to design estimators based on optimization.

Likelihood

The **likelihood** function $L(\theta)$ of the random variables $\{X_i\}_{i=1}^n$ for outcome $\{x_i\}_{i=1}^n$ under parameter θ is the parametric joint pdf

$$L(\theta) = f_{\{X_i\}_{i=1}^n}(\{x_i\}_{i=1}^n; \theta). \tag{17}$$

As a special case, if $\{X_i\}_{i=1}^n$ are i.i.d. random variables then

$$L(\theta) = \prod_{i=1}^{n} f_X(x_i; \theta) \tag{18}$$

Maximum likelihood estimate

The **maximum likelihood estimate** for outcome $\{x_i\}_{i=1}^n$ is the parameter $\theta^*(\{x_i\}_{i=1}^n)$ that maximizes the likelihood, i.e.

$$\theta^*(\{x_i\}_{i=1}^n) = \underset{\theta}{\text{argmax}}\ L(\theta) \tag{19}$$

The **maximum likelihood estimator** is the random variable

$$\hat{\theta} = \theta^*(\{X_i\}_{i=1}^n) \tag{20}$$

MLE: mean of a Gaussian

We start by assuming the *form* of the distribution is Gaussian with variance σ^2. We are estimating the mean, so the parameter is $\theta = \mu$

The likelihood is

$$L(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \tag{21}$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left(\sum_{i=1}^{n} -\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right) \tag{22}$$

Since the log function is monotonic increasing, the argmax of $L(\mu)$ is the same as the argmax of $\log L(\mu)$. The log is easier to work with.

$$\log L(\mu) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 \tag{23}$$

To maximize the log likelihood we find the stationary point

$$0 = \left.\frac{\partial \log L(\mu)}{\partial \mu}\right|_{\mu^*} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu^*) \tag{24}$$

which implies the MLE is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{25}$$

which happens to be the sample mean.

Homework: MLE for variance of a Gaussian

Homework P4-2: Derive the expression for the maximum likelihood estimator of the mean and variance of a Gaussian. Is the MLE variance biased?
Hint: Use the log-likelihood

$$\log L(\mu, \sigma) = -\frac{n}{2}\log(2\pi) - n\log(\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{26}$$

Suppose we wish to estimate a vector parameter which is exposed through the **linear observation model**

$$Y = H\theta + N \tag{27}$$

- Y is an **observation vector**
- H is a known constant **observation matrix**
- θ is an unknown constant **parameter vector**
- N is a **random observation noise vector**

The observation Y is directly measured, but the noise N is not.

Define the **residual**

$$E = Y - H\theta \tag{28}$$

which measures the error between the observation and its expected value.

A natural idea is to choose a parameter estimate that minimizes an objective function $v(\theta)$ which increases with the size of the residual.

$$\hat{\theta} = \underset{\theta}{\text{argmin}}\ v(\theta) \tag{29}$$

In particular, choose $v(\theta)$ as the squared norm of the residual:

$$v(\theta) = \|E\|^2 = (Y - H\theta)^\intercal (Y - H\theta) \tag{30}$$

Next we need some basic facts from optimization and matrix calculus.

Fact 1: The minimum of a continuous function $f(\theta)$ can only occur at a **stationary point** where the gradient vanishes

$$0 = \frac{\partial f(\theta)}{\partial \theta} \tag{31}$$

Fact 2: The derivative of an affine form is

$$\frac{d}{dx} a^\intercal x = a \tag{32}$$

and the derivative of a quadratic form is

$$\frac{d}{dx} x^\intercal Q x = 2Qx \tag{33}$$

Since $v(\theta)$ is a quadratic form, we can compute the minimizer in closed-form by finding the **stationary point** where the gradient of the objective vanishes:

$$0 = \frac{\partial v(\theta)}{\partial \theta}\bigg|_{\hat{\theta}} = 2(H^\intercal H)\hat{\theta} - 2H^\intercal Y \tag{34}$$

Rearranging yields the so-called **normal equation**

$$(H^\intercal H)\hat{\theta} = H^\intercal Y \tag{35}$$

If $H^\intercal H$ is invertible, we obtain the **least-squares estimate (LSE)**

$$\hat{\theta} = (H^\intercal H)^{-1} H^\intercal Y \tag{36}$$

Remark: If N is a white Gaussian noise, i.e. $N \sim \mathcal{N}(0, I)$, then it can be shown that the LSE is an unbiased, minimum variance, and maximum likelihood estimator.

Homework P4-3: We are given the following data:

$$\begin{bmatrix} 6.2 \\ 7.8 \\ 2.2 \end{bmatrix} = \begin{bmatrix} 3 \\ 4 \\ 1 \end{bmatrix} \theta + \begin{bmatrix} n_1 \\ n_2 \\ n_3 \end{bmatrix} \tag{37}$$

where n_i are random variables. Find a least-squares estimate for θ.

Asymptotics

Asymptotics

In this section we see major results from classical statistics

Claims are **asymptotic**; they only hold as the amount of data $\to \infty$

Claims are all about **convergence** of some kind

Contrast with finite-sample results c.f. [2]

Weak law of large numbers

Let X_i be an infinite sequence of i.i.d. random variables with a finite, common true mean μ and variance σ^2. Consider the sample mean

$$\hat{\mu}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{38}$$

Then for any fixed positive tolerance $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathbb{P}\left[|\hat{\mu}(n) - \mu| < \varepsilon\right] = 1 \tag{39}$$

i.e. the sample mean **converges in probability** to the true mean.

Proof: We already proved that the sample mean is consistent, which is the same thing as the WLLN.

Strong law of large numbers (SLLN)

Strong law of large numbers

Let X_i be an infinite sequence of i.i.d. random variables with a finite, common true mean μ and variance σ^2. Consider the sample mean

$$\hat{\mu}(n) = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{40}$$

Then we have

$$\mathbb{P}\left[\lim_{n\to\infty} \hat{\mu}(n) = \mu\right] = 1 \tag{41}$$

i.e. the sample mean **converges almost surely** to the true mean.

Proof: More involved than the WLLN. Also SLLN implies WLLN.

Notice the difference between weak and strong laws:

1. WLLN: Sequence of success probabilities approaches one
2. SLLN: Sequence of sample means approaches the true mean

Central limit theorem

Let X_i be an infinite sequence of independent random variables with cdf's F_{X_i}, finite means μ_i and finite variances σ_i^2.

Define the variance sum s_n^2 and normalized random variable Z_n

$$s_n^2 = \sum_{i=1}^{n} \sigma_i^2, \quad Z_n = \sum_{i=1}^{n} (X_i - \mu_i)/s_n \tag{42}$$

Suppose there exists $\varepsilon > 0$ and for all n sufficiently large that

$$\sigma_i < \varepsilon s_n, \quad i = 1, \ldots, n \tag{43}$$

Then

$$\lim_{n \to \infty} F_{Z_n}(z) = \Phi(z) \tag{44}$$

i.e. Z_n **converges in distribution** to a standard normal.

Homework P4-4: Let $\{X_i\}_{i=1}^n$ be a sequence of n i.i.d. random variables. Compute the approximate probability

$$\mathbb{P}[a \leq S \leq b] \tag{45}$$

of the sum

$$S(n) = \sum_{i=1}^{n} X_i \tag{46}$$

using the central limit theorem.
For concreteness, assume the X_i are uniform random variables on the unit interval $[0, 1]$, $n = 100$, $a = 45$, and $b = 52.5$.

Bibliography I

[1] John Woods and Henry Stark.
Probability, Statistics, and Random Processes for Engineers.
Pearson Higher Ed, 4 edition, 2011.

[2] Martin J Wainwright.
High-dimensional statistics: A non-asymptotic viewpoint, volume 48.
Cambridge University Press, 2019.
https://people.eecs.berkeley.edu/~wainwrig/BibPapers/Wai19.pdf.

Information Theory

Ben Gravell

benjamin.gravell@utdallas.edu
The Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
800 W. Campbell Rd.
Richardson, TX 75080

Outline

1. What is information theory?
2. Entropy
3. Wasserstein metric

Information theory

Information theory concerns quantifying the amount of information present in signals

- Originally developed for sending and receiving messages over communication channels
- Deals primarily with discrete random variables

Applications

- Machine learning e.g. classify images
- Reinforcement learning e.g. teach robots how to balance

c.f. Ch. 1-3 of Mackay's "Information Theory, Inference, and Learning Algorithms" [1]

c.f. Ch. 3 of Goodfellow's "Deep Learning" [2]

Intuitively, we want a quantity that measures

- The amount of information communicated by an outcome
- How surprising an outcome is

Our definition of "information" or "surprise" should satisfy three axioms:

1. Certain events yield zero information
   - They always happen, so they are not surprising
2. Less probable events yield more information
   - They happen less, so they are more surprising
3. The total information of independent events is the sum of the information of each individual event
   - Their chances of happening are unrelated, so knowing one outcome has no effect on how surprising the other outcome is

Information

The **(Shannon) information** of measuring random variable X with pmf P_X as outcome x is the quantity

$$I_X(x) = -\log_b(P_X(x)) \tag{1}$$

The log base b is an arbitrary choice which has the effect of fixing the units of information. Common choices:

- $b = 2$, "bits"
- $b = e$, "nats"
- $b = 10$, "dits"

Information is a **description of a distribution** like the pmf or cdf.

Sometimes the random variable $I(X) = I_X(X)$ is also called the information.

Entropy

The **entropy** of random variable X is the expected information of X

$$H(X) = \mathbb{E}_X[I(X)] \tag{2}$$

$$= \sum_i P_X(x_i) I_X(x_i) \tag{3}$$

$$= -\sum_i P(x_i) \log(P_X(x_i)) \tag{4}$$

Entropy measures the amount of randomness in X.

Entropy is a **summary statistic** like the mean or variance.

Example: Entropy of a coin flip

Let X be a Bernoulli random variable with success probability p

Let's compute the entropy of X as a function of the probability p

$$H(X) = -\sum_i P(x_i) \log(P_X(x_i)) \tag{5}$$

$$= -p \log(p) - (1-p) \log(1-p) \tag{6}$$

Example: Entropy of a coin flip

Exercise: Compute p which maximize and minimize entropy.

Solution:

- Max entropy when $p = 1/2$
  - Most random, heads and tails equally likely
- Min entropy when $p = 0$ or $p = 1$
  - Least random, heads or tails is certain

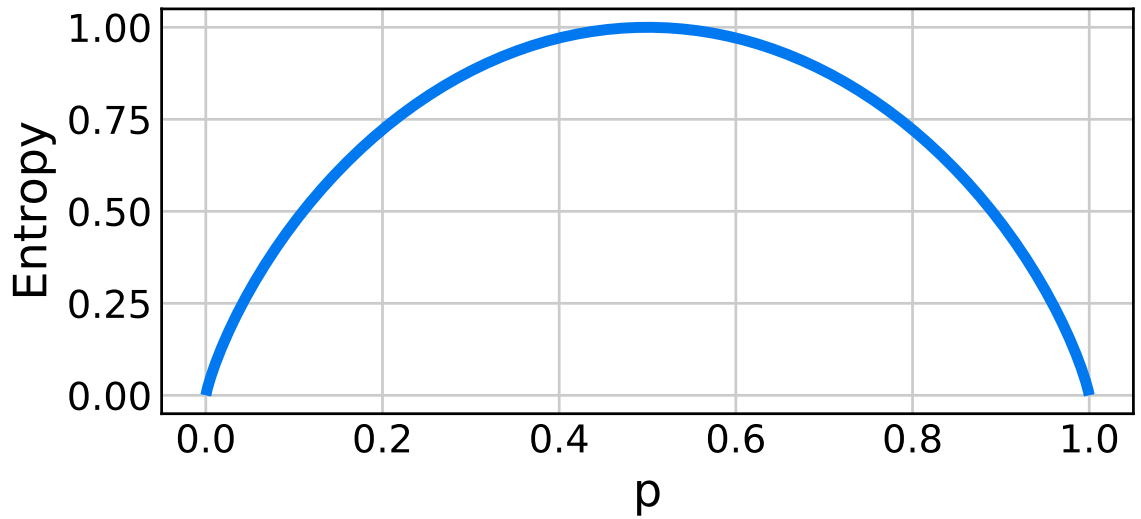


Figure 1: Entropy vs. parameter p for a Bernoulli random variable.
See `entropy_bernoulli.py`

Joint entropy

The **joint entropy** between two random variables X and Y with joint pmf P_{XY} is

$$H(X,Y) = -\sum_i \sum_j P_{XY}(x_i, y_i) \log(P_{XY}(x_i, y_i)) \tag{7}$$

Joint entropy measures the amount of randomness in X and Y.

Special case:
X and Y independent if and only if the joint entropy is additive

$$H(X,Y) = H(X) + H(Y) \tag{8}$$

Mutual information

The **mutual information** between two random variables X and Y is

$$I(X,Y) = H(X) + H(Y) - H(X,Y) \tag{9}$$

$$= \sum_i \sum_j P_{XY}(x_i, y_i) \log \left(\frac{P_{XY}(x_i, y_i)}{P_X(x_i) P_Y(y_i)} \right) \tag{10}$$

Mutual information measures the average reduction in uncertainty about X that results from learning the value of Y.

Special case: $I(X,X) = H(X)$, so entropy can be thought of as "self mutual information"

Cross-entropy

The **cross-entropy** from random variable Y to X is the expected information of Y with respect to X

$$H(X||Y) = \mathbb{E}_X[I(Y)] \tag{11}$$

$$= \sum_i P_X(x_i) I_Y(x_i) \tag{12}$$

$$= -\sum_i P_X(x_i) \log(P_Y(x_i)) \tag{13}$$

Cross-entropy measures the amount of randomness in Y, under the fictitious assumption that Y has the distribution of X for the purpose of computing expectation.

Special case: $H(X||X) = H(X)$, so entropy can be thought of as "self cross-entropy"

Relative entropy / Kullback-Leibler divergence

The **relative entropy** or **Kullback–Leibler (KL) divergence** from random variable Y to X is

$$\mathcal{D}_{KL}(X||Y) = H(X||Y) - H(X) \tag{14}$$

$$= \sum_i P_X(x_i) \log \left(\frac{P_X(x_i)}{P_Y(x_i)} \right) \tag{15}$$

KL divergence measures the **difference between two distributions**.

KL divergence is **not a distance metric** because

1. It is not symmetric
2. The triangle inequality fails

See `kl_divergence.py`

Wasserstein metric ("analytic" definition)

The p**th Wasserstein metric** between two pdfs f_X and f_Y is

$$W_p(f_X, f_Y) = \inf_{\pi \in \Pi(f_X, f_Y)} \left(\int_{\mathbb{R}^n \times \mathbb{R}^n} \|x - y\|^p d\Pi(x, y) \right)^{1/p} \tag{16}$$

where $\Pi(f_X, f_Y)$ is the space of joint pdfs with marginals f_X and f_Y.

- There are ∞ different joint pdfs with marginals f_X and f_Y!
- The joint pdf π defines a **transport map** between f_X and f_Y.
  - π is a plan for moving the mass from f_X to f_Y (and vice versa)
  - Finding the infimal π is a special case of the general **optimal transport problem** c.f. [3]
  - In many cases, this ∞-dim infimization problem can be solved analytically or by reformulating as a finite-dim optimization program

Wasserstein metric ("probabilistic" definition) [4]

The p**th Wasserstein metric** can be expressed as

$$W_p(f_X, f_Y) = \inf_{X \sim f_X,\ Y \sim f_Y} \left(\mathbb{E}_{XY}[\|X - Y\|^p]\right)^{1/p} \tag{17}$$

More facts:

- The two pdfs f_X and f_Y need not both be continuous or discrete
- $p = 1$ and $p = 2$ are the most common choices

Comparison with KL divergence:

- Like the KL divergence, the Wasserstein metric measures the **difference between two distributions**
- Unlike the KL divergence, the Wasserstein metric **is a valid distance metric**
  - Formal analysis using generic results for distance metrics is easier

Special case: pth Wasserstein metric of two Dirac deltas $f_X(x) = \delta(x - a)$ and $f_Y(y) = \delta(y - b)$

$$W_p(f_X, f_Y) = \|a - b\| \tag{18}$$

Special case: 2nd Wasserstein metric of two Gaussians $f_X = \mathcal{N}(\mu_X, \Sigma_X)$ and $f_Y = \mathcal{N}(\mu_Y, \Sigma_Y)$

$$W_2(f_X, f_Y) = \sqrt{\|\mu_X - \mu_Y\|^2 + \mathbf{Tr}\left[\Sigma_X + \Sigma_Y - 2\left(\Sigma_Y^{\frac{1}{2}}\Sigma_X\Sigma_Y^{\frac{1}{2}}\right)^{\frac{1}{2}}\right]} \tag{19}$$

Wasserstein metric

For the interested reader:

1. *"Statistical aspects of Wasserstein distances"* [4]
   - `https://arxiv.org/abs/1806.05500`
   - Contains a nice introduction on the Wasserstein metric.
2. *"Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations"* [5]
   - `https://arxiv.org/abs/1505.05116`
   - Quickly becoming a classic.
   - Details how to use the Wasserstein metric to solve optimization problems involving random problem data with unknown distribution while being robust to the worst-case distribution.

Bibliography I

[1] David JC MacKay and David JC Mac Kay.
Information theory, inference and learning algorithms.
Cambridge university press, 2003.
https://www.inference.org.uk/itila/.

[2] Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.
http://www.deeplearningbook.org.

[3] Cédric Villani.
Optimal transport: old and new, volume 338.
Springer, 2009.
https://cedricvillani.org/sites/dev/files/old_images/2012/08/preprint-1.pdf.

Bibliography II

[4] Victor M Panaretos and Yoav Zemel.
Statistical aspects of wasserstein distances.
Annual review of statistics and its application, 6:405–431, 2019.
https://arxiv.org/pdf/1806.05500.pdf.

[5] Peyman Mohajerin Esfahani and Daniel Kuhn.
Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations.
Mathematical Programming, 171(1):115–166, 2018.
https://arxiv.org/pdf/1505.05116.pdf.