Knowledge Reference

Kaiyu Zheng
PhD student at Brown University

May 2019 -

Abstract

This is the ultimate compilation of techinical knowledge from Math, Computer Sciecne, etc., learned from courses or other places, useful for my PhD career and thereafter.

Contents

Part I

Basics

Chapter 1

Discrete Mathematics

Chapter 2

Calculus

2.1 Integration

Theorem 2.1.1 (Integration by Parts)**.**

$$\int udv = uv - \int vdu \tag{2.1}$$

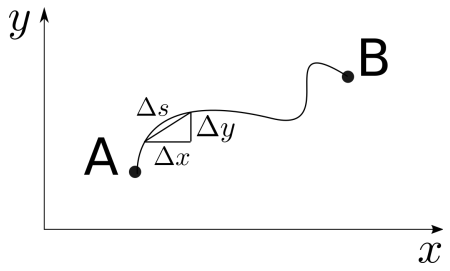
2.2 Partial Differentiation

Theorem 2.2.1 (Chain Rule of Partial Differentiation)**.** *Suppose* $u = f(x)$*,* $v = g(x)$*,* $w = h(u, v) = h(f(x), g(x)) = H(x)$*. Then,*

$$\frac{\partial w}{\partial x} = \frac{\partial w}{\partial f}\frac{\partial f}{\partial x} + \frac{\partial w}{\partial g}\frac{\partial g}{\partial x} \tag{2.2}$$

2.3 Calculus of Variations

Motivating Example Find the shortest path between two points in the plane, $A = (x_1, y_1)$, $B = (x_2, y_2)$. Assume $x_1 \neq x_2$, and let $y(x_1) = y_1$, $y(x_2) = y_2$, which means the second coordinate is a function of the first. So $y(x)$ for $x \in [x_1, x_2]$ is the trajectory.



Since the length of the path depends on the trajectory, let the length be a function $I(y)$. Δs is a small step along the trajectory. Therefore,

$$I(y) = \sum \Delta s \tag{2.3}$$

Note that

$$\Delta s = \sqrt{(\Delta x)^2 + (\Delta y)^2} = \sqrt{1 + \left(\frac{\Delta y}{\Delta x}\right)^2}\Delta x \tag{2.4}$$

As the limit of $\Delta x \to 0$, $\left(\frac{\Delta y}{\Delta x}\right) \to \left(\frac{\partial y}{\partial x}\right)$. Then,

$$I(y) = \int_{x_1}^{x_2} \sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2} dx \tag{2.5}$$

Therefore, the problem of finding the shortest path, is to minimize I.

More Generally *The fundamental problem of calculus of variations:* Given a function $f = f(x, y, y')$, find the functions $y(x)$ corresponding to the extrema points of the integral

$$I = \int_{x_1}^{x_2} f(x, y, y')dx \tag{2.6}$$

subject to the boundary conditions $y(x_1) = y_1$ and $y(x_2) = y_2$.

The Minimization Case Suppose our goal is to minimize I. Assume that y_* is the particular function that does so. Therefore, it must be $y_*(x_1) = y_1$, $y_*(x_2) = y_2$. Now, let's discuss a form, described as

$$Y(\epsilon, x) = y_*(x) + \epsilon g(x) \tag{2.7}$$

where ϵ is a real number, and $g(x)$ is an arbitrary function with $g(x_1) = 0$, $g(x_2) = 0$. This guarantees that $Y(\epsilon, x) = y_1$, $Y(\epsilon, x_2) = y_2$.

The term $\epsilon g(x)$ is called a *variation* of the minimizing function:

$$I(\epsilon) \int_{x_1}^{x_2} f(x, Y, Y')dx \tag{2.8}$$

Previously, $I(y)$ was a function of a function whose extrema is hard to define. Now, $I(\epsilon)$ is a function of a real value. Therefore, it is straightforward that

$$I'(\epsilon) = \frac{d}{d\epsilon}\int_{x_1}^{x_2} f(x, Y, Y')dx \tag{2.9}$$

$$= \int_{x_1}^{x_2} \frac{d}{d\epsilon} f(x, Y, Y')dx \tag{2.10}$$

$$= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial x}\frac{\partial x}{\partial \epsilon} + \frac{\partial f}{\partial y}\frac{\partial Y}{\partial \epsilon} + \frac{\partial f}{\partial y'}\frac{\partial Y'}{\partial \epsilon}\right) [chainrule?] \tag{2.11}$$

$$= \int_{x_1}^{x_2} \left(\frac{\partial f}{\partial y}g + \frac{\partial f}{\partial y'}g'\right) dx \tag{2.12}$$

$$\tag{2.13}$$

Note that, using integration by parts,

$$\int_{x_1}^{x_2} \frac{\partial f}{\partial y'}g'dx = \left[\frac{\partial f}{\partial y'}g\right]_{x_1}^{x_2} - \int_{x_1}^{x_2} g\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) dx \tag{2.14}$$

$$= 0 - \int_{x_1}^{x_2} g\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) dx \tag{2.15}$$

Therefore,

$$I'(\epsilon) = \int_{x_1}^{x_2} \left[\frac{\partial f}{\partial g} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right)\right] gdx \tag{2.16}$$

Since we have assumed that $y_*(x)$ is the extrema function, which is obtained when $\epsilon = 0$. This implies at $\epsilon = 0$, $I'(\epsilon) = I'(0) = 0$. Therefore, to have $I'(\epsilon) = 0$, since $g(x)$ is arbitrary, it must be the case that

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0 \tag{2.17}$$

This is called the *Euler-Lagrange Equation.*

◦ In the special case where f is *explicitly independent* of x, meaning that in f, x always appears as part of y and y' but not explcitly as in $f(x, y', y) = x^2 + y(x)$. In other words,

$$\frac{\partial^2 f}{\partial x \partial y'} = 0 \tag{2.18}$$

In this case, the Euler-Lagrange Equation can be simplified as

$$y'\frac{\partial f}{\partial y'} - f = C \tag{2.19}$$

This is called the *Beltrami identity.*

Back to the Motivating Problem Now, looking at the motivating problem, we have

$$I(y) = \int_{x_1}^{x_2} \sqrt{1 + \left(\frac{\partial y}{\partial x}\right)^2} dx = \int_{x_1}^{x_2} \sqrt{1 + (y')^2} dx \tag{2.20}$$

Here, $f(x, y, y') = \sqrt{1 + (y')^2}$. Thus, $\frac{\partial f}{\partial y}$ in the Euler-Lagrange equation equals to 0. Therefore, we have

$$\frac{d}{dx}\left(\frac{\partial f}{\partial y'}\right) = 0 \tag{2.21}$$

$$\Rightarrow \frac{d}{dx}\left(\frac{y'}{\sqrt{1 + (y')^2}}\right) = \frac{y''}{(1 + (y')^2)^{3/2}} = 0 \tag{2.22}$$

$$\Rightarrow y'' = 0 \tag{2.23}$$

Therefore, the solution is $y(x) = ax + b$ with a and b satisfying the boundary conditions; This is, indeed, a straight line.

2.3.1 Reyleigh-Ritz Method

In general, the *Rayleigh-Ritz Method* can be used to approximate eigen pairs (λ_i, ψ_i) and their Ritz residuals, σ_i. In calculus of variations, this method is used to approximate the solution $y(x)$. This method is illustrated by the following example (homework problem in ENGN 2010):

Example *Use **Rayleigh-Ritz approximate method** to minimize the integral. Use 3 terms in a series:*

$$I = \int_0^1 \Big[(y')^2 + y^2 - 2xy\Big] dx \tag{2.24}$$

where $y(0) = 1$ and $y(1) = 2$. Plot the graphs for y_0, y_1, and y_2 and the exact solution.

Assume that

$$y(x) \approx \phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x) + \cdots + c_N\phi_N(x) \tag{2.25}$$

For $N = 0$, we can choose $y_0 = y(x) \approx \phi_0(x) = 1 + x$, which satisfies the boundary conditions.
For $N = 1$, we can choose $y_1 = y(x) \approx \phi_0(x) + c_1\phi_1(x) = 1 + x + c_1x(x - 1)$. Evaluating the integral,

$$I(y_1) = \frac{11\,{c_1}^2}{30} - \frac{c_1}{3} + \frac{5}{3} \tag{2.26}$$

Solving $\frac{\partial I}{\partial c_1} = 0$ for c_1, we get $c_1 = \frac{5}{11}$. This means

$$y_1 = x + \frac{5\,x\;(x-1)}{11} + 1 \tag{2.27}$$

$$= \frac{5\,x^2}{11} + \frac{6\,x}{11} + 1 \tag{2.28}$$

For $N = 2$, we choose $y_2 = y(x) \approx \phi_0(x) + c_1\phi_1(x) + c_2\phi_2(x) = 1 + x + (c_1 + c_2 x)x(x-1)$. Evaluating the integral,

$$I(y_2) = \frac{11\,{c_1}^2}{30} + \frac{11\,c_1\,c_2}{30} - \frac{c_1}{3} + \frac{{c_2}^2}{7} - \frac{c_2}{6} + \frac{5}{3} \tag{2.29}$$

Extremizing I, we obtain a system of equations:

$$\frac{\partial I}{\partial c_1} = 0 \;\Rightarrow \frac{11\,c_1}{15} + \frac{11\,c_2}{30} - \frac{1}{3} = 0 \tag{2.30}$$

$$\frac{dI}{\partial c_2} = 0 \;\Rightarrow \frac{11\,c_1}{30} + \frac{2\,c_2}{7} - \frac{1}{6} = 0 \tag{2.31}$$

Solving for c_1 and c_2 and we get $c_1 = \frac{11}{15}$ and $c_2 = 0$. This means y_2 happens to be the same as y_1.

For the exact solution, derive the Euler-Lagrange equation, with $f(y, y', x) = (y')^2 + y^2 - 2xy$.

$$\frac{\partial f}{\partial y} - \frac{d}{dx}\Big(\frac{\partial f}{\partial y'}\Big) = 0 \tag{2.32}$$

$$2y - 2x - 2y'' = 0 \tag{2.33}$$

Solving the above differential equation with MATLAB yields

$$y(x) = x + \frac{e^x}{e+1} + \frac{e^{-x+1}}{e+1} \tag{2.34}$$

2.3.2 Finite Difference

This is another approximation method to find $y(x)$. Suppose we want to maximize $I(y) = \int_{x_0}^{x_{n+1}} F(x, y, y')dx$, given boundary conditions $g(x_0) = y_0$, $y(x_{n+1}) = y_{n+1}$. Divide the interval $[x_0, x_{n+1}]$ into $n+1$ parts,

$$\Delta x = \frac{x_{n+1} - x_0}{n+1} \tag{2.35}$$

Let $y_1, y_2, \cdots, y_n$ be the values of y corresponding to $x_1 = x_0 + \Delta x$, $x_2 = x_0 + 2\Delta x$, $\cdots$ Then the integral is approximated by a function of n variables:

$$\phi(y_1, y_2, \cdots y_n) = \sum_{i=0}^{n} F\left(x_i, y_i, \frac{y_{i+1} - y_i}{\Delta x}\right)\Delta x \tag{2.36}$$

The quantities $y_1, y_2, \cdots y_n$ are determined so that

$$\frac{\partial \phi}{\partial y_i} = 0 \qquad i = 1, 2, \cdots, n \tag{2.37}$$

Again, we will use an example to illustrate how to apply this method.

Example Solve the same problem (as in Rayleigh-Ritz) with **finite difference** for $n = 2$, 4, and 6.

The idea of finite difference is to divide the interval $[x_0, x_{n+1}]$ of x-axis into $n + 1$ segments of the same length Δx given by:

$$\Delta x = \frac{x_{n+1} - x_0}{n + 1} \tag{2.38}$$

Then, the integral$I(y) = \int_{x_0}^{x_{n+1}} f(y, y', x)dx$ is approximated by a function of n variables,

$$\phi(y_1, y_2, \cdots, y_n) = \sum_{i=0}^{n} f\Big(x_i, y_i, \frac{y_{i+1} - y_i}{\Delta x}\Big)\Delta x \tag{2.39}$$

where $x_i = x_0 + i\Delta x$. To extremize ϕ, we would want

$$\frac{\partial \phi}{\partial y_i} = 0 \qquad i = 1, 2, \cdots n \tag{2.40}$$

For the integral in Eq (2.24),

$$\phi(y_0, y_1, \cdots, y_{n+1}) = \sum_{i=0}^{n} \Big[\Big(\frac{y_{i+1} - y_i}{\Delta x}\Big)^2 + y_i^2 - 2x_i y_i\Big]\Delta x \tag{2.41}$$

When $n = 2$,

$$x_0 = 0, x_1 = \frac{1}{3}, x_2 = \frac{2}{3}, x_3 = 1 \tag{2.42}$$

$$\phi(y_{0:3}) = \frac{y_0^2}{3} + \frac{y_1^2}{3} - \frac{2y_1}{9} + \frac{y_2^2}{3} - \frac{4y_2}{9} \tag{2.43}$$

$$+ \frac{1}{3}(-3y_0 + 3y_1)^2 + \frac{1}{3}(-3y_1 + 3y_2)^2 + \frac{1}{3}(-3y_2 + 3y_3)^2 \tag{2.44}$$

Plug in the boundary conditions, and solve for the y values,

$$y_0 = 1,\ y_1 = \frac{37}{30},\ y_2 = \frac{47}{30},\ y_3 = 2, \tag{2.45}$$

When $n = 4$,

$$x_0 = 0, x_1 = \frac{1}{5}, x_2 = \frac{2}{5}, x_3 = \frac{3}{5}, x_4 = \frac{4}{5}, x_5 = 1 \tag{2.46}$$

$$\phi(y_{0:5}) = \frac{y_0^2}{5} + \frac{y_1^2}{5} - \frac{2y_1}{25} + \frac{y_2^2}{5} - \frac{4y_2}{25} + \frac{y_3^2}{5} \tag{2.47}$$

$$- \frac{6y_3}{25} + \frac{y_4^2}{5} - \frac{8y_4}{25} + \frac{1}{5}(-5y_0 + 5y_1)^2 \tag{2.48}$$

$$+ \frac{1}{5}(-5y_1 + 5y_2)^2 + \frac{1}{5}(-5y_2 + 5y_3)^2 \tag{2.49}$$

$$+ \frac{1}{5}(-5y_3 + 5y_4)^2 + \frac{1}{5}(-5y_4 + 5y_5)^2 \tag{2.50}$$

Plug in the boundary conditions, and solve for the y values,

$$y_0 = 1,\ y_1 = \frac{3951}{3505},\ y_2 = \frac{4527}{3505},\ y_3 = \frac{5228}{3505},\ y_4 = \frac{6054}{3505},\ y_5 = 2, \tag{2.51}$$

When $n = 6$,

$$x_0 = 0, x_1 = \frac{1}{7}, x_2 = \frac{2}{7}, x_3 = \frac{3}{7}, x_4 = \frac{4}{7}, x_5 = \frac{5}{7}, x_6 = \frac{6}{7}, x_7 = 1 \tag{2.52}$$

$$\phi(y_{0:7}) = \frac{y_0^2}{7} + \frac{y_1^2}{7} - \frac{2y_1}{49} + \frac{y_2^2}{7} - \frac{4y_2}{49} + \frac{y_3^2}{7} - \frac{6y_3}{49} \tag{2.53}$$

$$+ \frac{y_4^2}{7} - \frac{8y_4}{49} + \frac{y_5^2}{7} - \frac{10y_5}{49} + \frac{y_6^2}{7} - \frac{12y_6}{49} + \frac{1}{7}(-7y_0 + 7y_1)^2 \tag{2.54}$$

$$+ \frac{1}{7}(-7y_1 + 7y_2)^2 + \frac{1}{7}(-7y_2 + 7y_3)^2 + \frac{1}{7}(-7y_3 + 7y_4)^2 \tag{2.55}$$

$$+ \frac{1}{7}(-7y_4 + 7y_5)^2 + \frac{1}{7}(-7y_5 + 7y_6)^2 + \frac{1}{7}(-7y_6 + 7y_7)^2 \tag{2.56}$$

Plug in the boundary conditions, and solve for the y values,

$$y_0 = 1,\ y_1 = \frac{1006608}{926107},\ y_2 = \frac{1104952}{926107},\ y_3 = \frac{1220446}{926107},\ y_4 = \frac{1352747}{926107}, \tag{2.57}$$

$$y_5 = \frac{1501855}{926107},\ y_6 = \frac{1668113}{926107},\ y_7 = 2, \tag{2.58}$$

2.4 Complex Analysis

2.4.1 Basics

A complex number can be described as $z = (x, y) = x + iy$. The arithmetic of complex numbers are defined as follows:

- Addition: $(x_1, y_1) \pm (x_2, y_2) = (x_1 \pm x_2, y_1 \pm y_2)$
- Multiplication:

  $(x_1 + iy_1)(x_2 + iy_2) = x_1x_2 + ix_1y_2 + iy_1x_2 + i^2y_1y_2 = (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1)$

- Division: TODO

2.4.2 Differentiation

Definition 2.4.1 (Continuous)**.** A function $f(z)$ is *continuous* at $z = z_0$ if $f(z_0)$ is defined, and

$$\lim_{z \to z_0} f(z) = f(z_0) \tag{2.59}$$

Note that $|z - z_0| < \delta$ is a circle; $|f(z) - f(z_0)| < \epsilon$ is also a circle.

Definition 2.4.2 (Differentiation)**.**

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} \tag{2.60}$$

$f'(z_0)$ is the derivative of f at point z_0.

If we write $\Delta z = z - z_0$, then

$$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0} \tag{2.61}$$

2.4.3 Cauchy-Riemann equations

Definition 2.4.3 (Analytic function)**.** A function $f(z)$ is said to be *analytic in a domain* D if $f(z)$ is defined and differentiable at all points of D. The function $f(z)$ is said to be *analytic at a point* $z = z_0$ in D if $f(z)$ is analytic in a neighborhood of z_0.

Theorem 2.4.1. *Given a complex function $w = f(z) = u(x, y) + iv(x, y)$, this function f is analytic in a domain D if and only if*

$$u_x = v_y \qquad \text{and} \qquad u_y = -v_x \tag{2.62}$$

are satisfied. These two equations are called the Cauchy-Rieman equations.

Theorem 2.4.2. *Suppose f is continuous in some neighborhood of a point $z = x + iy$ and differentiable at z, then, the Cauchy-Riemann equations are satisfied at that point.*

Caucy-Riemann in Polar Form This is a homework problem in ENGN 2010: Prove Cauchy-Riemann equations in polar form.

Proof. Suppose we have complex function $f(z) = u(x, y) + iv(x, y)$. Consider polar coordinates (r, θ) such that $x = r\cos\theta$ and $y = r\sin\theta$. Assume f is differentiable at z.

According to the Chain rule,

$$u_x = \frac{\partial u}{\partial x} = \frac{\partial u}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial u}{\partial \theta}\frac{\partial \theta}{\partial x} = \frac{u_r}{\cos\theta} - \frac{u_\theta}{r\sin\theta} \tag{2.63}$$

$$u_y = \frac{\partial u}{\partial y} = \frac{\partial u}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial u}{\partial \theta}\frac{\partial \theta}{\partial y} = \frac{u_r}{\sin\theta} + \frac{u_\theta}{r\cos\theta} \tag{2.64}$$

$$v_x = \frac{\partial v}{\partial x} = \frac{\partial v}{\partial r}\frac{\partial r}{\partial x} + \frac{\partial v}{\partial \theta}\frac{\partial \theta}{\partial x} = \frac{v_r}{\cos\theta} - \frac{v_\theta}{r\sin\theta} \tag{2.65}$$

$$v_y = \frac{\partial v}{\partial y} = \frac{\partial v}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial v}{\partial \theta}\frac{\partial \theta}{\partial y} = \frac{v_r}{\sin\theta} + \frac{v_\theta}{r\cos\theta} \tag{2.66}$$

Because f is differentiable at z, the Cauchy-Riemann equations are satisfied. This means $u_x = v_y$ and $u_y = -v_x$. Substituting with the above equations, we get

$$\frac{u_r}{\cos\theta} - \frac{u_\theta}{r\sin\theta} = \frac{v_r}{\sin\theta} + \frac{v_\theta}{r\cos\theta} \tag{2.67}$$

$$\frac{u_r}{\sin\theta} + \frac{u_\theta}{r\cos\theta} = \frac{v_\theta}{r\sin\theta} - \frac{v_r}{\cos\theta} \tag{2.68}$$

It is possible to simplify and obtain

$$u_r = \frac{v_\theta}{r} \qquad u_\theta = -\frac{v_r}{r} \tag{2.69}$$

which are the Cauchy-Riemann equations in polar form. □

Chapter 3

Linear Algebra

Notation

We denote vectors using bold lower case letters such as $\boldsymbol{x}$, matrices using bold upper case letters such as $\boldsymbol{X}$, and entries of matrices using normal upper case letters such as X_{ij} or $X_{i,j}$ (The comma is used if the indices are expressed by equations). The vector $\boldsymbol{e_i}$ by default means the ith column vector in an indentity matrix with dimension depending on the context.

3.1 Linear System of Equations

Definition 3.1.1 (Row Echelon Form)**.** Each variable can be the leading variable for at most one equation.

For example,

$$\begin{aligned} x_1 + x_2 + x_3 - x_4 &= 0 \\ -x_2 + 7x_4 - x_5 &= -1 \\ x_4 + x_5 &= 2 \end{aligned} \tag{3.1}$$

Definition 3.1.2. Linear systems are *equivalent* if they are related by a sequence of elementary operations:

(1) Interchange position of rows

(2) Multiply an equal constant

(3) Add a multiple of one equation to another

Definition 3.1.3 (Augmented Matrix)**.** The linear system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m &= b_1, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m &= b_n \end{aligned} \tag{3.2}$$

can be written as an *augmented matrix* as follows:

$$\begin{bmatrix} a_{11} & \dots & a_{1m} & b_1 \\ \vdots & \ddots & \vdots & \vdots \\ a_{n1} & \dots & a_{nm} & b_n \end{bmatrix} \tag{3.3}$$

Definition 3.1.4 (Row Echelon Form)**.** A matrix is in *row echelon form* if

a) Every leading term is in a column to the left of the leading term of the row below it.

b) Any zero rows are at the bottom of the matrix

For example, the left matrix below is not an echelon form, because "0=7" has no leading variable. It is an *inconsistent* matrix. The right matrix is a echelon form.

$$\begin{bmatrix} 1 & 2 & 3 & 0 & 0 \\ 0 & 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 & 7 \end{bmatrix} \quad \begin{bmatrix} 1 & -2 & 5 & 2 & -1 \\ 0 & 3 & 4 & 5 & 6 \\ 0 & 0 & 22 & 14 & 4 \end{bmatrix}$$

The leading variable positions in the matrix are called *pivot positions*. A column in the matrix that contains a pivot position is a *pivot column*. The process of converting a linear system into echelon form is *Gaussian Elimination*.

Definition 3.1.5 (Reduced Row Echelon Form)**.** A matrix is said to be in *reduced row echelon form* if:

a) all pivot positions have 1

b) the only nonzero term in each pivot column is the pivot

c) it is in row echelon form.

Try finding the reduced row echelon form of the following matrix:

$$\begin{bmatrix} 0 & 3 & 4 & 5 & 6 \\ 1 & -2 & 5 & 2 & -1 \\ 3 & 0 & 1 & 2 & 5 \end{bmatrix} \tag{3.4}$$

Definition 3.1.6 (Homogeneity)**.** A *homogeneous linear equation* is

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = 0 \tag{3.5}$$

The equation is said to be in homogeneous form. A linear system where all equations are in homogeneous form is a *homogenous system*.

Every homogenous system is *consistent*, i.e. solvable.

3.2 Vectors

Definition 3.2.1 (Norm)**.** The *norm*, or magnitude of a vector $\boldsymbol{a} \in \mathbb{R}^n$ is defined as the *L2-norm* of the vector.

$$|\boldsymbol{a}| = \sqrt{\sum_{i=1}^{n} a_i^2} \tag{3.6}$$

Definition 3.2.2 (Dot Product)**.** *(Algebraic definition)* Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. Then the dot product (or inner product) between $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined as:

$$\boldsymbol{a} \cdot \boldsymbol{b} = \boldsymbol{a}^T \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i \tag{3.7}$$

(Geometric definition) The dot product of two Euclidean vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined by

$$\boldsymbol{a} \cdot \boldsymbol{b} = |\boldsymbol{a}||\boldsymbol{b}| cos(\theta_{\boldsymbol{a},\boldsymbol{b}}) \tag{3.8}$$

Also, The dot product $\boldsymbol{w} \cdot \boldsymbol{x} = b$ is a hyperplane, where $\boldsymbol{w}$ is normal to it.

Definition 3.2.3 (Projection)**.** Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. The projection of $\boldsymbol{b}$ onto $\boldsymbol{a}$ is defined

$$proj_{\boldsymbol{a}} \boldsymbol{b} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|} \frac{\boldsymbol{a}}{|\boldsymbol{a}|} = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}|^2} \boldsymbol{a} \tag{3.9}$$

Definition 3.2.4 (Outer Product)**.** Let $\boldsymbol{a}$ and $\boldsymbol{b}$ be two vectors in $\mathbb{R}^n$. Then the outer product (or tensor product) between $\boldsymbol{a}$ and $\boldsymbol{b}$ is defined such that $(\boldsymbol{a}\boldsymbol{b}^T)_{ij} = a_i b_j$:

$$\boldsymbol{a}\boldsymbol{b}^T = \begin{bmatrix} a_1 b_1 & a_1 b_2 & \cdots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \cdots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n b_1 & a_n b_2 & \cdots & a_n b_n \end{bmatrix} \tag{3.10}$$

Definition 3.2.5 (Linear Combination)**.** If $\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots \boldsymbol{u}_m$ are vectors and $c_1, c_2 \cdots c_m$ are scalars, then $c_1 \boldsymbol{u}_1 + c_2 \boldsymbol{u}_2 + \cdots + c_m \boldsymbol{u}_m$ is a linear combination of the vectors.

Definition 3.2.6 (Span)**.** Let $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ be a set of m vectors in $\mathbb{R}^n$. The *span* of the set is the set of linear combinations of $\boldsymbol{u}_1 \cdots \boldsymbol{u}_m$.

For example, suppose $\boldsymbol{u}_1 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$ and $\boldsymbol{u}_2 = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$, what is the span of $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$? A vector $\boldsymbol{v} = \begin{bmatrix} a \\ b \\ c \end{bmatrix} \in \text{span}\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ if and only if $\exists s, t . s\boldsymbol{u}_1 + t\boldsymbol{u}_2 = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$. s, t exist if $\begin{bmatrix} 1 & 3 & a \\ 2 & 2 & b \\ 3 & 1 & c \end{bmatrix}$ has

a solution. This matrix is reduced to $\begin{bmatrix} 1 & 3 & a \\ 0 & 4 & 2a-b \\ 0 & 0 & a-2b+c \end{bmatrix}$, therefore it has a solution when $a - ab + c = 0$ holds. So the *span* of $\{\boldsymbol{u}_1, \boldsymbol{u}_2\}$ is the plane $x - 2y + z = 0$.

Definition 3.2.7 (Relation of Span and Augmented Matrix)**.** If a vector $\boldsymbol{v}$ is in the span of vectors $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ then the matrix $[\boldsymbol{u}_1 \ \cdots \ \boldsymbol{u}_m \ \boldsymbol{v}]$ has at least 1 solution.

Theorem 3.2.1 (Relation of Span and Linearly Independence)**.** *If* $\boldsymbol{u} \in \text{span}\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ *then* $\text{span}\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\} = \text{span}\{\boldsymbol{u}, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$

3.2.1 Linear independence

Definition 3.2.8 (Linear Independence)**.** Let $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ be a set of vectors in $\mathbb{R}^n$. If the only solution to the equation $x_1\boldsymbol{u}_1 + \cdots + x_m\boldsymbol{u}_m = 0$ is the trivial solution (i.e. all zeros), then $\boldsymbol{u}_1 \cdots \boldsymbol{u}_m$ are *linearly independent.*

Fact: If any set of vector contains $\mathbf{0}$, this set of vectors are not linearly independent.

Definition 3.2.9 (Orthonormal Vectors)**.** Vectors in a set $\mathcal{U} = \{\boldsymbol{u}_1, \cdots \boldsymbol{u}_m\}$ are *orthonormal* if every vector in $\mathcal{U}$ is a unit vector and every pair $\boldsymbol{u_i}, \boldsymbol{u_j} \in \mathcal{U}$ of vectors are orthogonal, i.e. $\boldsymbol{u_i}^T\boldsymbol{u_j} = 0$.

Theorem 3.2.2. *Every set of orthonormal vectors is linearly independent (i.e. the vectors in the set are linearly independent).*

3.2.2 Linear dependence

Theorem 3.2.3 (Linear Dependence)**.** *Let* $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ *be a set of vectors in* $\mathbb{R}^n$*. If* $n < m$*, the set is* linearly dependent*.*

Corollary 3.2.3.1 (Relation of Span and Linearly Independence)**.** *If there is a set of* m linearly independent vectors *in* $\mathbb{R}^n$ *that spans all of* $\mathbb{R}^n$*, then* $m = n$*.*

Theorem 3.2.4 (Relation of Linear Combination and Linearly Dependence)**.** *Let* $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_m\}$ *be a set of vectors in* $\mathbb{R}^n$*. The vectors in this set are* linearly dependent *if one vector is a linear combination of others.*

3.2.3 Linear transformation

Definition 3.2.10 (Linear Transformation)**.** Function $T : \mathbb{R}^m \to \mathbb{R}^n$ is a *linear transformation* if for all $\boldsymbol{v}, \boldsymbol{u} \in \mathbb{R}^m$ and for all $r \in \mathbb{R}$, $T(\boldsymbol{v} + \boldsymbol{u}) = T\boldsymbol{v} + T\boldsymbol{u}$ and $T(r\boldsymbol{v}) = rT(\boldsymbol{v})$. $\mathbb{R}^m$ is the *domain*, and $\mathbb{R}^n$ is the *co-domain.* For $\boldsymbol{u} \in \mathbb{R}^m$, $T(\boldsymbol{u})$ is the *image* of $\boldsymbol{u}$ under T.

Definition 3.2.11 (Subspace)**.** A subset S of $\mathbb{R}^n$ is a *subspace* if S satisfies:

a) S contains $\mathbf{0}$.

b) if $\boldsymbol{u}$ and $\boldsymbol{v}$ are in S then $\boldsymbol{u} + \boldsymbol{v}$ is also in S. (*closure under addition*)

c) If r is a real number, and $\boldsymbol{u} \in S$ then, $r\boldsymbol{u} \in S$. (*closure under multiplication*)

Definition 3.2.12 (One-to-one and On-to)**.** Let $T : \mathbb{R}^m \to \mathbb{R}^n$, $T(\boldsymbol{v}) = \boldsymbol{A}\boldsymbol{v}$ thus T is a linear transformation. T is *one-to-one* (injective) if and only if $T(\boldsymbol{x}) = \mathbf{0}$ has only the trivial solution (i.e. $\boldsymbol{x} = \mathbf{0}$), or equivalently, $T(\boldsymbol{a}) = T(\boldsymbol{b})$ implies $\boldsymbol{a} = \boldsymbol{b}$. This means the columns of $\boldsymbol{A}$ are linearly independent. T is *on-to* (surjective) if and only if columns of $\boldsymbol{A}$ span $\mathbb{R}^n$.

Note, $\boldsymbol{A}$ is a $n \times m$ matrix. If $m > n$, T is *not* one-to-one. If $m < n$, T is *not* on-to.

In more general terms, if a function is one-to-one (**injective**), every element of the co-domain is mapped to by *at most one* element of the domain. If a function is on-to (**surjective**) if every element of the co-domain is mapped to by at least one element of the domain. A function is *one-to-one and on-to* (**bijective**) if every element of the co-domain is mapped to by exactly one element of the domain.

3.3 Matrix Algebra

3.3.1 Addition

If $\boldsymbol{A}, \boldsymbol{B} \in M_{n\times m}(\mathbb{R})$ and $r \in \mathbb{R}$,

$$(\boldsymbol{A} + \boldsymbol{B})_{ij} = (\boldsymbol{A})_{ij} + (\boldsymbol{B})_{ij} \tag{3.11}$$

3.3.2 Scalar Multiplication

$$(r\boldsymbol{A})_{ij} = r(\boldsymbol{A})_{ij} \tag{3.12}$$

3.3.3 Matrix Multiplication

If $T : \mathbb{R}^m \to \mathbb{R}^n$ is represented by $\boldsymbol{A} \in M_{n\times m}(\mathbb{R})$ and $W : \mathbb{R}^n \to \mathbb{R}^l$ is represented by $\boldsymbol{B} \in M_{l\times m}(\mathbb{R})$, then $\boldsymbol{B}\boldsymbol{A}$ should be represented as $W \circ T : \mathbb{R}^m \to \mathbb{R}^l$. So $\boldsymbol{B}\boldsymbol{A} \in M_{l\times m}(\mathbb{R})$.

Matrix multiplication can be thought of as applying a series of linear transformation to vectors in an initial domain. For example, $\boldsymbol{B}\boldsymbol{A}$ is illustrated as

$$\mathbb{R}^m \xrightarrow{T} \mathbb{R}^n \xrightarrow{W} \mathbb{R}^l$$

Notice that although the final transformation is $\mathbb{R}^m \to \mathbb{R}^l$ which reads "a transformation going from $\mathbb{R}^m$ (domain of T) to $\mathbb{R}^l$ (codomain of W)", the formal notation is "reversed", which is $W \circ T$.

Alternative definition: Let $\boldsymbol{A} \in M_{n\times p}(\mathbb{R})$ and $\boldsymbol{B} \in M_{p\times m}(\mathbb{R})$, then $\boldsymbol{AB} \in M_{n\times m}(\mathbb{R})$. We will look at several equivalent algebraic definitions of $\boldsymbol{AB}$ from different perspectives. But first of all, let us look at two interpretations of *matrix-vector multiplication* $\boldsymbol{Ax}$ where $\boldsymbol{x} \in \mathbb{R}^p$.

1) We consider $\boldsymbol{Ax}$ from the perspective of considering *row vectors* of $\boldsymbol{A}$, that is, we view $\boldsymbol{A}$ as

$$\boldsymbol{A} = \begin{bmatrix} \underline{\boldsymbol{a}}_1 \\ \underline{\boldsymbol{a}}_2 \\ \vdots \\ \underline{\boldsymbol{a}}_n \end{bmatrix} \tag{3.13}$$

where each component $\underline{\boldsymbol{a}}_{\boldsymbol{i}}$ is a row vector. Then, $\boldsymbol{Ax}$ can be computed by performing dot product $\underline{\boldsymbol{a}}_{\boldsymbol{i}}^T\boldsymbol{x}$ for $i \in \{1, \cdots, n\}$, therefore $(\boldsymbol{Ax})_i = \underline{\boldsymbol{a}}_{\boldsymbol{i}}^T\boldsymbol{x}$. Specifically,

$$\boldsymbol{Ax} = \begin{bmatrix} \underline{\boldsymbol{a}}_1^T\boldsymbol{x} \\ \underline{\boldsymbol{a}}_2^T\boldsymbol{x} \\ \vdots \\ \underline{\boldsymbol{a}}_n^T\boldsymbol{x} \end{bmatrix} \tag{3.14}$$

2) We can also compute $\boldsymbol{Ax}$ by considering *column vectors* of $\boldsymbol{A}$, such that

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_{|1} & \boldsymbol{a}_{|2} & \cdots & \boldsymbol{a}_{|p} \end{bmatrix} \tag{3.15}$$

where each component $\boldsymbol{a}_{|i}$ is a column vector. Then, the matrix multiplication $\boldsymbol{Ax}$ can be viewed as a linear combination of columns of $\boldsymbol{A}$ with coefficients determined by entries x_i for $i \in \{1, \cdots, k\}$.

$$\begin{aligned} \boldsymbol{Ax} &= x_1\boldsymbol{a}_{|1} + x_2\boldsymbol{a}_{|2} + \cdots + x_n\boldsymbol{a}_{|p} \\ &= \sum_{i=1}^{p} \boldsymbol{a}_{|i}x_i \end{aligned} \tag{3.16}$$

Now, let us look at *matrix-matrix multiplication* also from two perspectives.

1) When we consider row vectors of $\boldsymbol{A}$ and column vectors of $\boldsymbol{B}$, the multiplication $\boldsymbol{AB}$ can be viewed as

$$\boldsymbol{AB} = \begin{bmatrix} \boldsymbol{Ab}_{|1} & \boldsymbol{Ab}_{|2} & \cdots & \boldsymbol{Ab}_{|m} \end{bmatrix} \tag{3.17}$$

where $\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_{|1} & \boldsymbol{b}_{|2} & \cdots & \boldsymbol{b}_{|m} \end{bmatrix}$. From Equation 3.14, we know $(\boldsymbol{Ab}_{|k})_i = \underline{\boldsymbol{a}}_i^T\boldsymbol{b}_{|k}$. Therefore, $(\boldsymbol{AB})_{ij} = \underline{\boldsymbol{a}}_i^T\boldsymbol{b}_{|j}$.

2) When we consider column vectors of $\boldsymbol{A}$ and row vectors of $\boldsymbol{B}$, the multiplication $\boldsymbol{AB}$ can be viewed as

$$\boldsymbol{AB} = \sum_{i=1}^{p} \boldsymbol{a}_{|i} \underline{\boldsymbol{b}}_i^T \tag{3.18}$$

where $\boldsymbol{a}_{|i} \underline{\boldsymbol{b}}_i^T$ is the *outer product* with output dimension of $n \times m$.

Properties of Matrix Multiplication:

1) $\boldsymbol{A}(\boldsymbol{BC}) = (\boldsymbol{AB})\boldsymbol{C}$
2) $\boldsymbol{A}(\boldsymbol{B} + \boldsymbol{C}) = \boldsymbol{AB} + \boldsymbol{AC}$
3) $(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{C} = \boldsymbol{AC} + \boldsymbol{BC}$
4) $s\boldsymbol{AB} = \boldsymbol{A}s\boldsymbol{B}$
5) $\boldsymbol{IA} = \boldsymbol{AI} = \boldsymbol{A}$

Caveats:

1) $\boldsymbol{AB} \neq (\boldsymbol{BA})$ (usually)
2) $\boldsymbol{AC} = \boldsymbol{AB} \not\Rightarrow \boldsymbol{C} = \boldsymbol{B}$

3.3.4 Transpose

If $\boldsymbol{A} \in M_{n\times m}(\mathbb{R})$, then $\boldsymbol{A}^T \in M_{m\times n}(\mathbb{R})$.

Properties of Transpose:

1) $(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$
2) $(s\boldsymbol{A})^T = s(\boldsymbol{A}^T)$
3) $(\boldsymbol{AC})^T = \boldsymbol{C}^T\boldsymbol{A}^T$

Theorem 3.3.1. *A matrix $\boldsymbol{A}$ has the property that for all $\boldsymbol{v}, \boldsymbol{W} \in \mathbb{R}^2$, $\boldsymbol{v} \cdot \boldsymbol{w} = \boldsymbol{Av} \cdot \boldsymbol{Aw}$ if and only if $\boldsymbol{A}$ is* orthogonal, *that is, $\boldsymbol{AA}^T = \boldsymbol{I}_n$, or equivalently, $\boldsymbol{A}^T = \boldsymbol{A}^{-1}$.*

Conjugate Transpose

Definition 3.3.1 (Conjugate Transpose)**.** Given an $n \times n$ matrix $\boldsymbol{A}$ with complex entries (i.e. entries are complex numbers), the *conjugate transpose* (or Hermitian transpose, Hermitian conjugate) of $\boldsymbol{A}$ is given by

$$\boldsymbol{A}^H = \left(\bar{\boldsymbol{A}}\right)^T \tag{3.19}$$

where $\bar{\boldsymbol{A}}$ has the complex conjugate entries of $\boldsymbol{A}$.

Properties of Conjugate Transpose:

1) $(\boldsymbol{A} + \boldsymbol{B})^H = \boldsymbol{A}^H + \boldsymbol{B}^H$

2) $(s\boldsymbol{A})^H = s(\boldsymbol{A}^H)$

3) $(\boldsymbol{AC})^H = \boldsymbol{C}^H\boldsymbol{A}^H$

3.3.5 Inverse

Definition 3.3.2 (Invertibility)**.** A linear map $T : \mathbb{R}^m \to \mathbb{R}^n$ is *invertible* if it is one-to-one and on-to. Two implications follows if $T : \mathbb{R}^m \to \mathbb{R}^n$ is invertible:

1) $m = n$ (required)

2) T^{-1} is also linear.

Theorem 3.3.2 (Invert of Matrix)**.** *An $n \times n$ matrix $\boldsymbol{A}$ is invertible if there exists a matrix $\boldsymbol{B}$ so that $\boldsymbol{BA} = \boldsymbol{I}_n$. If $\boldsymbol{A}$ is invertible, $\boldsymbol{B}$ is* unique *and define $\boldsymbol{A}^{-1} = \boldsymbol{B}$.*

$$\implies \boldsymbol{BA} = \boldsymbol{AB} = \boldsymbol{I}_n$$

To compute $\boldsymbol{A}^{-1}$, form an $n \times 2n$ matrix $\begin{bmatrix}\boldsymbol{A} & \boldsymbol{I}_n\end{bmatrix}$. Then convert it to reduced row echelon form, which results in $\begin{bmatrix}\boldsymbol{I}_n & \boldsymbol{A}^{-1}\end{bmatrix}$.

Theorem 3.3.3 (Invertibility Implies Non-zero Determinant)**.** *An $n \times n$ matrix is invertible if and only if its* determinant *is not zero.*

Theorem 3.3.4 (Invertibility and Positive-Definite)**.**

Any positive-definite *matrix is invertible.*

Properties of Matrix Inverse:
If $\boldsymbol{A}, \boldsymbol{B}$ are invertible $n \times n$ matrix, and $\boldsymbol{C}$, $\boldsymbol{D}$ are $n \times m$ matrix. Then:

a) $\boldsymbol{A}^{-1}$ is invertible. $(\boldsymbol{A}^{-1})^{-1} = \boldsymbol{A}$

b) $\boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$

c) $\boldsymbol{A}\boldsymbol{B}$ is invertible. $(\boldsymbol{A}\boldsymbol{B})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$

d) If $\boldsymbol{A}\boldsymbol{C} = \boldsymbol{A}\boldsymbol{D}$, then $\boldsymbol{C} = \boldsymbol{D}$

e) If $\boldsymbol{A}\boldsymbol{C} = \boldsymbol{0}$, then $\boldsymbol{C} = \boldsymbol{0}$

f) $(\boldsymbol{A}^T)^{-1} = (\boldsymbol{A}^{-1})^T$

Proof. We will prove c) and d).

c) Show that $\boldsymbol{A}\boldsymbol{B}(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) = \boldsymbol{I}_n$:

$$\boldsymbol{A}\boldsymbol{B}(\boldsymbol{B}^{-1}\boldsymbol{A}^{-1}) = \boldsymbol{I}_n = \boldsymbol{A}(\boldsymbol{B}\boldsymbol{B}^{-1})\boldsymbol{A}^{-1} = \boldsymbol{A}\boldsymbol{I}_n\boldsymbol{A}^{-1} = \boldsymbol{A}\boldsymbol{A}^{-1} = \boldsymbol{I}_n \tag{3.20}$$

d) Show that $\boldsymbol{A}\boldsymbol{C} = \boldsymbol{A}\boldsymbol{D} \implies \boldsymbol{C} = \boldsymbol{D}$:

$$\begin{aligned} &\boldsymbol{A}\boldsymbol{C} = \boldsymbol{A}\boldsymbol{D} && (3.21)\\ \implies &\boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{C} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{D} && (3.22)\\ \implies &\boldsymbol{I}_n\boldsymbol{C} = \boldsymbol{I}_n\boldsymbol{D} && (3.23)\\ \implies &\boldsymbol{C} = \boldsymbol{D} && (3.24) \end{aligned}$$

From the above proof, we see that $\boldsymbol{A}$ being invertible is important, because otherwise $\boldsymbol{A}^{-1}$ does not exist.

□

3.3.6 Trace

Definition 3.3.3 (Trace)**.** Let $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$. The trace of $\boldsymbol{A}$ is defined as the sum of entries along the main diagonal:

$$tr(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii} \tag{3.25}$$

Properties of Trace:

a) $tr(\boldsymbol{A} + \boldsymbol{B}) = tr(\boldsymbol{A}) + tr(\boldsymbol{B})$

b) $tr(c\boldsymbol{A}) = c \cdot tr(\boldsymbol{A})$

c) $tr(\boldsymbol{A}\boldsymbol{B}) = tr(\boldsymbol{B}\boldsymbol{A})$

d) $tr(\boldsymbol{A}) = tr(\boldsymbol{A}^T)$

e) $tr(\boldsymbol{X}^T\boldsymbol{Y}) = tr(\boldsymbol{X}\boldsymbol{Y}^T) = tr(\boldsymbol{Y}^T\boldsymbol{X}) = \sum_{ij} X_{ij}Y_{ij}$

f) Similarity-invariant:

$$tr(\boldsymbol{P}^{-1}\boldsymbol{AP}) = tr(\boldsymbol{P}^{-1}(\boldsymbol{AP})) = tr((\boldsymbol{AP})\boldsymbol{P}^{-1}) = tr(\boldsymbol{A}(\boldsymbol{PP}^{-1})) = tr(\boldsymbol{A})$$

g) $d\ tr(\boldsymbol{X}) = tr(d\boldsymbol{X})$

Trace and Eigenvalues: In Section 3.5, we discuss eigenvectors and eigenvalues in more detail. For the sake of proximity, we describe the relation of trace and eigenvalues here.

Theorem 3.3.5. *If $\boldsymbol{A}$ is an $n \times n$ matrix with real or complex entries and if $\lambda_1, \cdots, \lambda_n$ are eigenvalues of $\boldsymbol{A}$, then*

$$tr(\boldsymbol{A}) = \sum_i \lambda_i \tag{3.26}$$

$$tr(\boldsymbol{A}^k) = \sum_i \lambda_i^k \tag{3.27}$$

3.3.7 Power

Definition 3.3.4 (Integral Power)**.** $\boldsymbol{A}^n$ is raising matrix $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$ to the power of n. It is defined as the multiplication of n the same matrix $\boldsymbol{A}$:

$$\boldsymbol{A}^n = \boldsymbol{AA}\cdots\boldsymbol{A} \tag{3.28}$$

The matrix to the 0th power is defined to be the identity matrix, i.e. $\boldsymbol{A}^0 = \boldsymbol{I}$. The exponentiation of a non-square matrix is not well-defined; One reason is that the 0th power is undefined. Note that $\boldsymbol{A}^{-1} \neq 1/\boldsymbol{A}$, as it is the matrix inverse.

Definition 3.3.5 (Square Root)**.** Matrix $\boldsymbol{B} = \boldsymbol{A}^{1/2}$ if and only if $\boldsymbol{BB} = \boldsymbol{A}$.

To compute the square root of an arbitrary square matrix, a method that involves Jordan Normal Form (Section **??**) can be used. We discuss the case when the matrix $\boldsymbol{A}$ is diagonalizable (Section 3.7.4), meaning there exist matrix $\boldsymbol{V}$ and diagonal matrix $\boldsymbol{D}$ such that $\boldsymbol{A} = \boldsymbol{VDV}^{-1}$. The square root of $\boldsymbol{A}$ is $\boldsymbol{R}$ such that:

$$\boldsymbol{R} = \boldsymbol{VSV}^{-1} \tag{3.29}$$

where $\boldsymbol{S}$ is *any* square root of $\boldsymbol{D}$. To verify,

$$\boldsymbol{RR} = \boldsymbol{VS}(\boldsymbol{V}^{-1}\boldsymbol{V})\boldsymbol{SV}^{-1} = \boldsymbol{VSSV}^{-1} = \boldsymbol{VDV}^{-1} = \boldsymbol{A} \tag{3.30}$$

The square root of $\boldsymbol{D}$ is simply obtained by taking the square root of all entries along the diagonal. To raise a matrix $\boldsymbol{A}$ to an arbitrary real value p, we can follow

$$\boldsymbol{A}^p = \exp(p\ln(\boldsymbol{A})) \tag{3.31}$$

where $\ln(\boldsymbol{A})$ is defined in Section 3.3.8 below.

3.3.8 Exponential and Logarithm

Definition 3.3.6 (Exponential of Matrix)**.** The exponential of matrix $\boldsymbol{A}$ is defined as

$$e^{\boldsymbol{A}} = \sum_{n=0}^{\infty} \frac{\boldsymbol{A}}{n!} \tag{3.32}$$

This is a generalization of ordinary exponential function e^x which is

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!} \tag{3.33}$$

Definition 3.3.7 (Logarithm of Matrix)**.** Matrix $\boldsymbol{B}$ is the logarithm of matrix $\boldsymbol{A}$ if

$$\ln(\boldsymbol{A}) = \boldsymbol{B} \tag{3.34}$$

which is equivalent as $e^{\boldsymbol{B}} = \boldsymbol{A}$.

The logarithm of $\boldsymbol{A}$ does not always exist; At least, $\boldsymbol{A}$ needs to be invertible, but this is not enough. For more, please refer to Wikipedia.

3.3.9 Conversion Between Matrix Notation and Summation

Outer products Suppose $\boldsymbol{x_i} \in \mathbb{R}^d$, and $\boldsymbol{X} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}]^T$. Then,

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T = \boldsymbol{X}^T \boldsymbol{X} \tag{3.35}$$

To understand this intuitively, note that the vertical vectors $\boldsymbol{x_i}$ are rows of $\boldsymbol{X}$. Then, recall from Equation 3.18, matrix multiplication $\boldsymbol{AB}$ can be viewed as the sum of outer products between column vectors of $\boldsymbol{A}$ and row vectors of $\boldsymbol{B}$. Therefore, we need to transpose $\boldsymbol{X}$ and multiply it by itself, yielding $\boldsymbol{X}^T\boldsymbol{X}$.

Similarly, if $\boldsymbol{y} \in \mathbb{R}^s$, and $\boldsymbol{Y} = [\boldsymbol{y}_1, \cdots, \boldsymbol{y}_n]^T$, we have:

$$\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{y}_i^T = \boldsymbol{X}^T \boldsymbol{Y} \tag{3.36}$$

Examples:

- Conversion from primal objective to dual objective for *kernel ridge regression*. In ridge regression, with $\boldsymbol{X} \in \mathbb{R}^{N \times d}$, $\boldsymbol{y} \in \mathbb{R}^N$, $\boldsymbol{x}_i \in \mathbb{R}^d$ features each we can formulate the objective as:

$$\min_{\boldsymbol{w}} \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i \right)^2 + \lambda \boldsymbol{w}^T \boldsymbol{w} \tag{3.37}$$

According to the Representer Theorem, $\boldsymbol{w}^* = \sum_{i=1}^N \alpha_i \boldsymbol{x}_i$ is the optimal weights. Thus, with $\boldsymbol{\alpha} \in \mathbb{R}^N$, the above can be transformed into the following (kernel ridge regression objective), where $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_j)$ is the kernel function:

$$\min_{\boldsymbol{\alpha}} \frac{1}{N} \sum_{i=1}^N \Big(y_i - \sum_{j=1}^N \alpha_j \boldsymbol{x}_j^T \boldsymbol{x}_i \Big)^2 + \lambda \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j \boldsymbol{x}_i^T \boldsymbol{x}_j \tag{3.38}$$

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} \sum_{i=1}^N \Big(y_i - \sum_{j=1}^N \alpha_j k(\boldsymbol{x}_j, \boldsymbol{x}_i) \Big)^2 + \lambda \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \tag{3.39}$$

To transform Equation 3.39 into matrix notation, first let $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ be the kernel matrix where $\boldsymbol{K}_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$. Then, we have:

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha})^T (\boldsymbol{y} - \boldsymbol{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{3.40}$$

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} (\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{K} \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{y} - \boldsymbol{y}^T \boldsymbol{K} \boldsymbol{\alpha} + \boldsymbol{y}^T \boldsymbol{y}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{3.41}$$

Because $\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{y}$ and $\boldsymbol{y}^T \boldsymbol{K} \boldsymbol{\alpha}$ are just scalars, we can just write:

$$\Leftrightarrow \min_{\boldsymbol{\alpha}} \frac{1}{N} (\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{K} \boldsymbol{\alpha} - 2\boldsymbol{\alpha}^T \boldsymbol{K}^T \boldsymbol{y} + \boldsymbol{y}^T \boldsymbol{y}) + \lambda \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{3.42}$$

The matrix notation conversion of the ridge regularization term is important.

Singular value decomposition Given matrix $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, how do you write its singular value decomposition $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^T$ using summation notation?

3.4 Vector Spaces

Definition 3.4.1 (Vector Space)**.** A *vector space* $\mathcal{V}$ over a field, such as real numbers $\mathbb{R}$, is a set $\mathcal{V}$ with two functions:

$$\text{addition } + : V \times V \to V \qquad (\text{e.g. } \boldsymbol{v} + \boldsymbol{w}) \tag{3.43}$$

$$\text{scalar multiplication } \cdot : \mathbb{R} \times V \to V \qquad (\text{e.g. } a\boldsymbol{v},\ a \in \mathbb{R}) \tag{3.44}$$

and satisfy these properties (*axioms* for all $\boldsymbol{v}, \boldsymbol{w}, \boldsymbol{u} \in \mathcal{V}$ and $s, t \in \mathbb{R}$:

1) $\boldsymbol{u} + (\boldsymbol{v} + \boldsymbol{w}) = (\boldsymbol{u} + \boldsymbol{v}) + \boldsymbol{w}$ (Associativity of addition)

2) $\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{v} + \boldsymbol{u}$ (Commutativity of addition)

3) There exists an element $\boldsymbol{0} \in \mathcal{V}$, called the zero vector, such that $\boldsymbol{v} + \boldsymbol{0} = \boldsymbol{v} for all \boldsymbol{v} \in \mathcal{V}$. (Identity element of addition)

4) $\cdots$ For more, refer to the Wikipedia's article on vector space.

Definition 3.4.2 (Subspace)**.** A *linear subspace* is a subset of $\mathbb{R}^n$ that is a vector space with the induced multiplication and addition from $\mathbb{R}^n$.

For example, $\mathcal{S} \in \mathbb{R}^n$ is a vector subspace if for all $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{S}$, $\boldsymbol{v} + \boldsymbol{w} \in \mathcal{S}$, and for all $r \in \mathbb{R}$, $\boldsymbol{v} \in \mathcal{S}$, $r\boldsymbol{v} \in \mathcal{S}$. The latter implies $\mathbf{0} \in \mathcal{S}$.

$\left\{ \begin{bmatrix} a \\ b \\ 1 \end{bmatrix} \in \mathbb{R}^3, a, b \in \mathbb{R} \right\}$ is *not* a subspace.

Definition 3.4.3 (Null space)**.** If $\boldsymbol{A}$ is an $n \times n$ matrix, the set of solutions to the system $\boldsymbol{Ax} = \mathbf{0}$ is a subspace of $\mathbb{R}^n$, called the *null space* of $\boldsymbol{A}$ or *null($\boldsymbol{A}$)*.

Proof. Suppose $\boldsymbol{v}$, $\boldsymbol{w}$ are vectors in $\mathbb{R}^n$ that satisfy $\boldsymbol{Av} = \boldsymbol{Aw} = \mathbf{0}$. Then $\boldsymbol{A}(\boldsymbol{v} + \boldsymbol{w}) = \boldsymbol{Av} + \boldsymbol{Aw} = \mathbf{0}$. And $\boldsymbol{A}(r\boldsymbol{v}) = r\boldsymbol{Av} = \mathbf{0}$. Therefore, the set of solutions to $\boldsymbol{Ax} = \mathbf{0}$ is closed both under addition and multiplication. □

3.4.1 Determinant

Before we formally define determinants, let us use $det(\boldsymbol{A})$ to refer to the determinant of matrix $\boldsymbol{A}$, which is a real value.

Definition 3.4.4. (Determinant and Minor) If $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, define $\boldsymbol{M}_{ij}$ as the $n-1 \times n-1$ matrix formed by deleting the i-th row and j-th column. $\boldsymbol{A}$. $det(\boldsymbol{M}_{ij})$ is called the *minor* of entry a_{ij} in $\boldsymbol{A}$.

Definition 3.4.5 (Cofactor)**.** If $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, the *cofactor* of a_{ij}, or $C_{ij} = (-1)^{i+j} det(\boldsymbol{M}_{ij})$.

Definition 3.4.6 (Singularity)**.** A square matrix $\boldsymbol{A}$ that is invertible is called *nonsingular*. Otherwise, it is called *singular* or *degenerate*.

Theorem 3.4.1 (Singularity and Determinant)**.** *A square matrix is* singular *if and only if its determinant is 0.*

Now, we formally introduce determinant of a matrix.

Definition 3.4.7 (Determinant)**.** The determinant of $\boldsymbol{A}$ is an $n \times n$ matrix

$$\begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}$$

The determinant of $\boldsymbol{A}$ is recursively defined as:

$$det(\boldsymbol{A}) = |\boldsymbol{A}| = a_{11}C_{11} + a_{12}C_{12} + \cdots + a_{1n}C_{1n} \tag{3.45}$$

And when $n = 1$, $det(a_{11}) = a_{11}$ (base case).

The above definition is recursive because the definition of cofactor contains determinant.

Geometric Meaning of Determinants First, we focus on 2D. Suppose

$$\boldsymbol{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \boldsymbol{x}_1 = \begin{bmatrix} a \\ c \end{bmatrix}, \boldsymbol{x}_2 = \begin{bmatrix} b \\ d \end{bmatrix}$$

We have $det(\boldsymbol{A}) = ad - bc$. This is the *signed* area of the parallelogram formed by vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. In the 3D case, the determinant represents the signed volume of the hexahedron formed by the three column vectors in the matrix.

Theorem 3.4.2 (Invertibility and Determinant)**.** *For $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, it is invertible if and only if $det(\boldsymbol{A} \neq 0$.*

In other words, the determinant of an n by n matrix $\boldsymbol{A}$ is 0 if and only if the rows are linearly dependent (and not zero if and only if they are linearly independent).

Properties of Determinants:

a) The determinant equals to the product of eigenvalues λ_i:

$$det(\boldsymbol{A}) = \prod_i \lambda_i$$

b) $det(c\boldsymbol{A}) = c^n \cdot det(\boldsymbol{A})$

c) $det(\boldsymbol{AB}) = det(\boldsymbol{A})det(\boldsymbol{B})$

d) $det(\boldsymbol{A}^{-1}) = \dfrac{1}{det(\boldsymbol{A})}$

e) $det(\boldsymbol{A}^T) = det(\boldsymbol{A})$

f) $det(\boldsymbol{A}^n) = det(\boldsymbol{A})^n$

Cool Facts about Determinants[1]:

1) Interchanging any two rows of an n by n matrix $\boldsymbol{A}$ reverses the sign of its determinant.

2) If two rows of a matrix are equal, its determinant is 0. (Because $det(\boldsymbol{A}) = -det(\boldsymbol{A})$ implies $det(\boldsymbol{A}) = 0$.

3) If $\boldsymbol{A}$ is an n by n matrix, adding a multiple of one row to a different row does not affect its determinant.

4) An n by n matrix with a row of zeros has determinant zero.

[1]Source: http://www.math.lsa.umich.edu/~hochster/419/det.html

3.4.2 Kernel

Definition 3.4.8 (Kernel)**.** Suppose $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear transformation. The *kernel* of T is the set of vectors $\boldsymbol{x}$ such that $T(\boldsymbol{x}) = \mathbf{0}$, denoted by $ker(T)$. In other words,

$$ker(T) = \{\boldsymbol{x} \in \mathbb{R}^m | T(\boldsymbol{x} = \mathbf{0})\} \tag{3.46}$$

Theorem 3.4.3 (Kernel and Injectivity)**.** *Suppose $T : \mathbb{R}^m \to \mathbb{R}^n$ is a linear transformation. Then T is one-to-one if and only if $ker(T) = \{\mathbf{0}\}$.*

This is rather intuitive. T being one-to-one means $T(\boldsymbol{x}) = \mathbf{0}$ has only the trivial solution which is $\boldsymbol{x} = \mathbf{0}$. By definition of kernel, $ker(T) = \{\mathbf{0}\}$.

3.4.3 Basis

Definition 3.4.9 (Basis)**.** A set $\mathcal{B} = \{\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}\}$ is a basis for a subspace $\mathcal{S}$ if

a) $\mathcal{B}$ spans $\mathcal{S}$.

b) $\mathcal{B}$ is linearly independent.

For example, $\mathcal{S} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$ is the standard basis in the $\mathbb{R}^2$.
To find basis for $\mathcal{S} = span\{\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}\}$,

1. Use $\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}$ to form the rows of a matrix $\boldsymbol{A}$.
2. Transform $\boldsymbol{A}$ into row echelon form $\boldsymbol{B}$.
3. The nonzero rows give a basis for $\mathcal{S}$.

3.4.4 Change of Basis

Definition 3.4.10 (Change of Basis)**.** Suppose subspaces $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{R}^n$ each have a basis $\mathcal{B}_1 = \{\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}\}$ and $\mathcal{B}_2 = \{\boldsymbol{v_1}, \cdots, \boldsymbol{v_m}\}$, respectively. Let $\boldsymbol{A} = \begin{bmatrix} [\boldsymbol{a_1}]_{\mathcal{B}_1} \cdots [\boldsymbol{a_n}]_{\mathcal{B}_1} \end{bmatrix}$ be a matrix with column vectors relative to the basis $\mathcal{B}_1$[2]. Then, to represent column vectors in $\boldsymbol{A}$ with $\mathcal{B}_2$, we apply a *change-of-basis* matrix $\boldsymbol{P}_{\mathcal{B}_1 \to \mathcal{B}_2}$ from $\mathcal{B}_1$ to $\mathcal{B}_2$, such that

$$[\boldsymbol{a_i}]_{\mathcal{B}_2} = \boldsymbol{P}_{\mathcal{B}_1 \to \mathcal{B}_2} [\boldsymbol{a_i}]_{\mathcal{B}_1} \tag{3.47}$$

To find the change-of-basis matrix from $\mathcal{B}_1$ to $\mathcal{B}_2$, notice first that the definition of the (ordered) bases $\mathcal{B}_1 = \{\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}\}$ and $\mathcal{B}_2 = \{\boldsymbol{v_1}, \cdots, \boldsymbol{v_m}\}$ involve vectors relative to the standard basis. For example, if $\mathcal{B}_1 = \left\{ \begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 2 \end{bmatrix} \right\}$, the coordinates of the basis vectors

[2]Usually we omit the subscript when denoting vectors since by default the basis is the standard basis $\mathcal{S}$.

are relative to the the $\mathbb{R}^2$ space, even though they are the "unit basis vectors" relative to the subspace spanned by $\mathcal{B}_1$. That is, $[\boldsymbol{u_1}]_{\mathcal{S}} = \begin{bmatrix}1\\0\end{bmatrix}_{\mathcal{B}_1}$. Therefore, we can obtain the change-of-basis matrix from $\mathcal{B}_1$ to $\mathcal{S}$ effortlessly, given by

$$\boldsymbol{P}_{\mathcal{B}_1\to\mathcal{S}} = \begin{bmatrix}\boldsymbol{u_1} & \cdots & \boldsymbol{u_n}\end{bmatrix} \tag{3.48}$$

Because $\boldsymbol{u_i} = \boldsymbol{P}_{\mathcal{B}_1\to\mathcal{S}}[\boldsymbol{e_i}]_{\mathcal{B}_1}$. The same goes for $\mathcal{B}_2$. Therefore, we can easily obtain $\boldsymbol{P}_{\mathcal{B}_1\to\mathcal{S}}$ and $\boldsymbol{P}_{\mathcal{B}_2\to\mathcal{S}}$. Thus, to change the basis from $\mathcal{B}_1$ to $\mathcal{B}_2$, we can first change to the standard basis, then change to $\mathcal{B}_2$, summarized by:

$$\boldsymbol{P}_{\mathcal{B}_1\to\mathcal{B}_2} = \boldsymbol{P}_{\mathcal{S}\to\mathcal{B}_2}\boldsymbol{P}_{\mathcal{B}_1\to\mathcal{S}} \tag{3.49}$$

$$= \boldsymbol{P}^{-1}_{\mathcal{B}_2\to\mathcal{S}}\boldsymbol{P}_{\mathcal{B}_1\to\mathcal{S}} \tag{3.50}$$

For an entire matrix $\boldsymbol{A}$ representing the transformation $T : \mathbb{R}^n \to \mathbb{R}^n$, we can construct a matrix to represent the same linear transformation within a different subspace $\mathcal{B} \subset \mathbb{R}^n$, say $W : \mathcal{B} \to \mathcal{B}$, by leveraging the change-of-basis matrix $\boldsymbol{P}_{\mathcal{B}\to\mathcal{S}}$:

$$[\boldsymbol{A}]_{\mathcal{B}} = \boldsymbol{P}^{-1}_{\mathcal{B}\to\mathcal{S}}\boldsymbol{A}\boldsymbol{P}_{\mathcal{B}\to\mathcal{S}} \tag{3.51}$$

3.4.5 Dimension, Row & Column Space, and Rank

Definition 3.4.11 (Dimension)**.** Let $\mathcal{S}$ be a subspace of $\mathbb{R}^n$. Then the dimension of $\mathcal{S}$, denoted as $dim(\mathcal{S})$, is the number of vectors in any basis of $\mathcal{S}$.

Definition 3.4.12 (Row Space, Column Space)**.** Suppose $\boldsymbol{A} \in M_{n\times m}(\mathbb{R})$. Then:

- $row(\boldsymbol{A})$ = span of rows of $\boldsymbol{A}$ (row space)
- $col(\boldsymbol{A})$ = span of columns of $\boldsymbol{A}$ (column space)

$row(\boldsymbol{A}) \subseteq \mathbb{R}^m$, $col(\boldsymbol{A}) \subseteq \mathbb{R}^n$.

Theorem 3.4.4 (Basis for Row and Column Spaces)**.** *Let $\boldsymbol{A}$ be a matrix, and $\boldsymbol{B}$ be a row-echelon form of that matrix. Then*

a) The nonzero rows of $\boldsymbol{B}$ form a basis for $row(\boldsymbol{A})$.

b) The columns of $\boldsymbol{A}$ corresponding to pivot columns of $\boldsymbol{B}$ form a basis for $col(\boldsymbol{A})$.

Theorem 3.4.5 (Dimension of Row and Column Spaces Are Equal)**.** *The following is always true for matrix $\boldsymbol{A}$:*

$$dim(col(\boldsymbol{A})) = dim(row(\boldsymbol{A})) \tag{3.52}$$

Definition 3.4.13 (Rank)**.** The rank of a matrix $\boldsymbol{A}$ is defined by:

$$rank(\boldsymbol{A}) = dim(col(\boldsymbol{A})) = dim(row(\boldsymbol{A})) \tag{3.53}$$

Definition 3.4.14 (Nullity)**.** The *nullity* of $\boldsymbol{A}$ is $dim(null(\boldsymbol{A}))$.

Theorem 3.4.6 (Rank-Nullity Theorem)**.** *Let $\boldsymbol{A}$ be an $n \times m$ matrix. Then*

$$rank(\boldsymbol{A}) + nullity(\boldsymbol{A}) = m \tag{3.54}$$

3.5 Eigen

Definition 3.5.1 (Eigenvector and Eigenvalue)**.** Let $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, then a nonzero vector $\boldsymbol{u}$ is an *eigenvector* of $\boldsymbol{A}$ if there exists a scalar λ such that $\boldsymbol{Au} = \lambda\boldsymbol{u}$. The scalar λ is called the *eigenvalue*

$\mathbf{0}$ is never an eigenvector.

Theorem 3.5.1 (Scaled Eigenvectors)**.** *Suppose $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, and $\boldsymbol{u}$ is an eigenvector with eigenvalue λ. Then for any $r \neq 0$, $r \in \mathbb{R}$, $r\boldsymbol{u}$ is another eigenvector with eigenvalue λ.*

It is important to note that the theorem above does not imply that all eigenvectors with eigenvalue λ should be related by the scalar λ. With this in mind, it is more intuitive to accept the following theorem.

Theorem 3.5.2 (Eigenspace)**.** *If $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$, then the set of eigenvectors with eigenvalue λ, together with $\mathbf{0}$ is a subspace of $\mathbb{R}^n$, called the* eigenspace*.*

Theorem 3.5.3 (Condition for an Eigenvalue)**.** *Let $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$. Then λ is an eigenvalue of $\boldsymbol{A}$ if and only if*

$$det(\boldsymbol{A} - \lambda\boldsymbol{I_n}) = 0. \tag{3.55}$$

We refer to $det(\boldsymbol{A} - \lambda\boldsymbol{I_n}) = 0$ as the *characteristic polynomial*.

Definition 3.5.2 (Characteristic Polynomial)**.** The *characteristic polynomial* of an $n \times n$ matrix $\boldsymbol{A}$, $char_{\boldsymbol{A}}(\lambda)$, is the degree n polynomial $det(\boldsymbol{A} - \lambda\boldsymbol{I_n}) = 0$.

Caveat: Some linear maps do not have eigenvalues or eigenvectors, such as below:

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

The intuition of eigenvectors is to think of them as the axis of the corresponding linear transformation. The eigenvalue λ helps to know if $\boldsymbol{x}$ is stretched or shrunk, when multiplied by a matrix $\boldsymbol{A}$ (i.e. $\boldsymbol{Ax}$).

3.5.1 Multiplicity of Eigenvalues

Definition 3.5.3 (Algebraic Multiplicity)**.** The algebraic multiplicity of an eigenvalue α of $\boldsymbol{A}$ is found by k in $char_{\boldsymbol{A}} = (\alpha - \lambda)^k Q(\lambda)$ where $Q(\lambda)$ is a polynomial with $Q(\lambda) \neq 0$.

For example, for $char_{\boldsymbol{A}} = -\lambda(\lambda-2)^2 = -(\lambda-0)(\lambda-2)^2$. Therefore, $\lambda = 0$ has algebraic multiplicity of 1, and $lambda = 2$ has algebraic multiplicity of 2.

Definition 3.5.4 (Geometric Multiplicity)**.** The geometric multiplicity of an eigenvalue λ is the dimension of the eigenspace associated with λ, i.e. number of linearly independent eigenvectors of that eigenvalue.

- 0 is eigenvalue if $\boldsymbol{A} \in M_{n\times n}(\mathbb{R})$ is *singular* (See definition 3.4.6).
- Geometric multiplicity $\leq$ algebraic multiplicity (of an eigenvalue).

3.5.2 Eigendecomposition

Definition 3.5.5 (Eigendecomposition of a Matrix)**.** Let $\boldsymbol{A}$ be an $n \times n$ matrix, with n linearly independent eigenvectors $\boldsymbol{u_i}$ for $i \in \{1, \cdots, n\}$. Then we can perform an eigendecomposition of $\boldsymbol{A}$ as follows

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1} \tag{3.56}$$

where $\boldsymbol{U}$ is an $n \times n$ matrix whose ith column is the eigenvector $\boldsymbol{u_i}$ of $\boldsymbol{A}$, and $\boldsymbol{\Lambda}$ is the diagonal matrix whose diagonal entries are the corresponding eigenvalues (i.e. $\Lambda_{ii} = \lambda_i$).

This definition implies that $\boldsymbol{A}$ must be *diagonalizable* (Section 3.7.4). It is usually convenient to have $\boldsymbol{U}$ be a orthonormal matrix.

3.6 The Big Theorem

Theorem 3.6.1 (The Big Theorem)**.** *Let $\mathcal{A} = \{\boldsymbol{a_1}, \cdots, \boldsymbol{a_n}\}$ be a set of vectors in $\boldsymbol{R}^n$. Let $\boldsymbol{A} = [\boldsymbol{a_1} \quad \cdots \quad \boldsymbol{a_n}]$ be an $n \times n$ matrix, and let $T : \mathbb{R}^n \to \mathbb{R}^n$ be given by $T(\boldsymbol{X}) = \boldsymbol{A}\boldsymbol{x}$. Then the following statements are equivalent:*

a) $\mathcal{A}$ spans $\mathbb{R}^n$

b) $\mathcal{A}$ is linearly independent (i.e. $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{0}$ has only the trivial solution)

c) $\mathcal{A}$ is a basis for $\mathbb{R}^n$

d) $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ has a unique solution for all $\boldsymbol{b} \in \mathbb{R}^n$

e) T is onto (surjective)

f) T is one-to-one (injective)

g) $\boldsymbol{A}$ *is an invertible matrix*

h) $ker(T) = \{\mathbf{0}\}$

i) $col(\boldsymbol{A}) = \mathbb{R}^n$

j) $row(\boldsymbol{A}) = \mathbb{R}^n$

k) $rank(\boldsymbol{A}) = n$

l) $det(\boldsymbol{A}) \neq 0$

m) $\lambda = 0$ *is not an eigenvalue of* $\boldsymbol{A}$

3.7 Special Matrices

3.7.1 Block Matrix

Definition 3.7.1 (Block Matrix)**.** A block matrix $\boldsymbol{M}$ is defined as

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}$$

where $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are matrices (or block matrices) themselves.

Block matrices share many useful properties as normal matrices, by treating block entries as normal matrix entries. For example:

$$\boldsymbol{M}^2 = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} \tag{3.57}$$

$$= \begin{bmatrix} \boldsymbol{A}^2 + \boldsymbol{BC} & \boldsymbol{AB} + \boldsymbol{BD} \\ \boldsymbol{CA} + \boldsymbol{DC} & \boldsymbol{CB} + \boldsymbol{D}^2 \end{bmatrix} \tag{3.58}$$

3.7.2 Orthogonal

Definition 3.7.2 (Orthogonal Matrix)**.** An *orthogonal matrix* $\boldsymbol{Q}$ is a square matrix with real entries whose columns and rows are orthogonal unit vectors (i.e., *orthonormal* vectors), i.e.

$$\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^T = \boldsymbol{I} \tag{3.59}$$

Therefore, we have $\boldsymbol{Q}^T = \boldsymbol{Q}^{-1}$. To fully understand why Equation 3.59 holds, we need to know that for two orthogonal vectors $\boldsymbol{u_1}$ and $\boldsymbol{u_2}$, $\boldsymbol{u_1}^T\boldsymbol{u_2} = 0$. And $\boldsymbol{u_1}^T\boldsymbol{u_1} = |\boldsymbol{u_1}|^2 = 1$. Therefore, in the resulting matrix, all entries are 0 except for ones along the diagonal.

3.7.3 Diagonal

Definition 3.7.3 (Diagonal Matrix)**.** A square matrix $\boldsymbol{D}$ is a diagonal matrix if all entries except for ones along the main diagonal are 0.

Simple fact: for two diagonal matrices $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$, their multiplication $\boldsymbol{D_1 D_2} = \boldsymbol{D_3}$ is also a diagonal matrix with each entry $\boldsymbol{D_3}[i]$ along[3] the main diagonal equals to $\boldsymbol{D_1}[i]\boldsymbol{D_2}[i]$.

Therefore, every diagonal matrix is invertible. The inverse $\boldsymbol{D}^{-1}$ of diagonal matrix $\boldsymbol{D}$ has entries $\boldsymbol{D}^{-1}[i] = 1/\boldsymbol{D}[i]$.

Another fact: The determinant of a diagonal matrix is the product of the diagonal entries.

Yet another fact: The column vectors of a diagonal matrix $\boldsymbol{D}$ are the eigenvectors of $\boldsymbol{D}$, and each diagonal entry is the eigenvalue for the eigenvector at the corresponding column, that is

$$\boldsymbol{D} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \tag{3.60}$$

This can be verified simply by solving the characteristic polynomial $det(\boldsymbol{D} - \lambda \boldsymbol{I}) = 0$.

3.7.4 Diagonalizable

Definition 3.7.4 (Diagonalizable Matrix)**.** An $n \times n$ matrix $\boldsymbol{A}$ is *diagonalizable* if there exists an $n \times n$ matrix $\boldsymbol{P}$ such that

$$\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} \tag{3.61}$$

where $\boldsymbol{D}$ is a diagonal matrix.

Note that $\boldsymbol{D} = \boldsymbol{P}^{-1}\boldsymbol{A}\boldsymbol{P} \Longrightarrow \boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$

Theorem 3.7.1 (The Diagonalization Theorem)**.**

a) An $n \times n$ matrix $\boldsymbol{A}$ is diagonalizable if and only if $\boldsymbol{A}$ has n linearly independent eigenvectors.

b) $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$ where $\boldsymbol{D}$ is a diagonal matrix if and only if *all n columns of $\boldsymbol{P}$ are linearly independent eigenvectors of $\boldsymbol{A}$* and *the diagonal entries of $\boldsymbol{D}$ are their corresponding eigenvalues.*

If we can find n linearly independent eigenvectors for an $n \times n$ matrix $\boldsymbol{A}$, then we know the matrix is diagonalizable. Furthermore, we can use those eigenvectors and their corresponding eigenvalues to find the invertible matrix $\boldsymbol{P}$ and diagonal matrix $\boldsymbol{D}$ necessary to show that $\boldsymbol{A}$ is diagonalizable.

[3] The $[i]$ just means the ith entry along the main diagonal.

Theorem 3.7.2 (Power of Diagonalizable Matrix)**.** *If* $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{D}\boldsymbol{P}^{-1}$*, then* $\boldsymbol{A}^k = \boldsymbol{P}\boldsymbol{D}^k\boldsymbol{P}^{-1}$

3.7.5 Symmetric

Definition 3.7.5 (Symmetric Matrix)**.** A square matrix $\boldsymbol{A}$ is symmetric if and only if

$$\boldsymbol{A} = \boldsymbol{A}^T \tag{3.62}$$

For any $n \times m$ matrix $\boldsymbol{B}$, the matrix $\boldsymbol{B}^T\boldsymbol{B} \in \mathbb{R}^{n \times n}$ is symmetric. Also, every square diagonal matrix is symmetric.

Facts about symmetric matrix

- Any symmtric matrix:
  - has only real eigenvalues;
  - is always *diagonalizable*;
  - has orthogonal eigenvectors;
- The symmetric matrix $\boldsymbol{A}$ is
  - positive definite if all its eigenvalues are positive.
  - positive semidefinite if all its eigenvalues are non negative..

3.7.6 Positive-Definite

We omit the discussion of complex matrices for now.

Definition 3.7.6 (Positive-Definite)**.** A *symmetric* $n \times n$ real matrix $\boldsymbol{A}$ is *positive definite* if for all $\boldsymbol{x} \in \mathbb{R}^n$
$\{\boldsymbol{0}\}$,

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0 \tag{3.63}$$

The negative definite, positive semi-definite, and negative semi-definite matrices are defined analogously. For "$*$ semi-$*$", zero is allowed (e.g. $\boldsymbol{A}$ is positive semi-definite implies $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0$).

Theorem 3.7.3. *Covariance matrix is positive semi-definite.*

Given data $\boldsymbol{X} \in \ltimes \times$, its covariance matrix $\boldsymbol{\Sigma}$ is computed by the following:

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T] \tag{3.64}$$

For nonzero $\boldsymbol{y} \in \mathbb{R}^d$

$$\boldsymbol{y}^T \boldsymbol{\Sigma} \boldsymbol{y} = \boldsymbol{y}^T \mathbb{E}[(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T] \boldsymbol{y} \tag{3.65}$$

$$= \mathbb{E}[\boldsymbol{y}^T (\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])(\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T \boldsymbol{y}] \tag{3.66}$$

$$= \mathbb{E}[\boldsymbol{Q}^T \boldsymbol{Q}] \tag{3.67}$$

For $\boldsymbol{Q} = (\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}])^T \boldsymbol{y}$. Therefore, $\boldsymbol{y}^T \boldsymbol{\Sigma} \boldsymbol{y} \geq 0$, which means $\boldsymbol{\Sigma}$ is *positive semi-definite*.

3.7.7 Singular Value Decomposition

Theorem 3.7.4. *For any given real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$, there exists a unique set of matrices $\boldsymbol{U}, \boldsymbol{S}, \boldsymbol{V}$ such that*

$$\boldsymbol{A} = \boldsymbol{U} \boldsymbol{S} \boldsymbol{V}^T \tag{3.68}$$

where $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{S} \in \mathbb{R}^{n \times p}$ and $\boldsymbol{V} \in \mathbb{R}^{p \times p}$ $\boldsymbol{U}^T \boldsymbol{U} = \boldsymbol{I}$ and $\boldsymbol{V}^T \boldsymbol{V} = \boldsymbol{I}$. This is called the singular value decomposition *of $\boldsymbol{A}$.*

$\boldsymbol{U}$ and $\boldsymbol{V}$ are orthonormal matrices. $\boldsymbol{S}$ is a diagonal matrix[4]. The elements in $\boldsymbol{S}$ are called *singular values* of $\boldsymbol{A}$. The eigenvectors of $\boldsymbol{A}^T \boldsymbol{A}$ are columns of $\boldsymbol{V}$, and the eigenvectors of $\boldsymbol{A} \boldsymbol{A}^T$ are columns of $\boldsymbol{U}$. The entries in $\boldsymbol{S}$ are positive, and sorted in decreasing order ($S_{11} \geq S_{22} \geq \cdots$).

3.7.8 Similar

Definition 3.7.7 (Similar)**.** An $n \times n$ matrix $\boldsymbol{A}$ is *similar* to an $n \times n$ matrix $\boldsymbol{B}$ (denoted as $\boldsymbol{A} \sim \boldsymbol{B}$) if there exists an invertible matrix $\boldsymbol{P}$ such that $\boldsymbol{P}^{-1} \boldsymbol{A} \boldsymbol{P} = \boldsymbol{B}$.

The intuition behind matrix similarity is that the two matrices represent the same linear operator with respect to (possibly) different bases. The matrix $\boldsymbol{P}$ can be viewed as a change-of-basis matrix.

Theorem 3.7.5. *All similar matrices have the same eigenvalues.*

Proof. Suppose $\boldsymbol{A} \sim \boldsymbol{B}$ with $\boldsymbol{B} = \boldsymbol{P}^{-1} \boldsymbol{A} \boldsymbol{P}$, and $\boldsymbol{B} \boldsymbol{x} = \lambda \boldsymbol{x}$. Then $\boldsymbol{P}^{-1} \boldsymbol{A} \boldsymbol{P} \boldsymbol{x} = \lambda \boldsymbol{x}$. Then $\boldsymbol{A} \boldsymbol{P} \boldsymbol{x} = \lambda \boldsymbol{P} \boldsymbol{x}$, which means $\boldsymbol{A}$ also has eigenvalue λ (but with eigenvector $\boldsymbol{P} \boldsymbol{x}$). □

Note that some matrices have the same eigenvalues, but they are not similar.

Theorem 3.7.6 (Determine Matrix Similarity[5])**.** *Two square matrices are similar if and only if they have the same* Jordan normal form*.*

There are multiple corollaries from this theorem in the reference.

[4] More precisely, it is a rectangular diagonal matrix because n may not equal to p. Still, $S_{ij} = 0$ if $i \neq j$.

[5] Reference: http://kom.aau.dk/~jakob/selPubl/papers1995/ijmest_1995.pdf

3.7.9 Jordan Normal Form

Definition 3.7.8 (Jordan Normal Form[6])**.** The *Jordan normal form* of a linear transformation $T : \mathcal{V} \to \mathcal{V}$ is a special type of block matrix in which each block consists of Jordan blocks with possibly differing constants λ_i. In particular, it is a block matrix of the form:

$$\boldsymbol{J} = \begin{bmatrix} \boldsymbol{J_1} & 0 & \cdots & 0 \\ 0 & \boldsymbol{J_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \boldsymbol{J_p} \end{bmatrix} \tag{3.69}$$

where every $\boldsymbol{J_k}$ is a block matrix (Jordan block), defined as:

$$\boldsymbol{J_k} = \begin{bmatrix} \lambda_k & 1 & 0 & \cdots & 0 \\ 0 & \lambda_k & 1 & \ddots & 0 \\ 0 & 0 & \lambda_k & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 1 \\ 0 & 0 & 0 & \cdots & \lambda_k \end{bmatrix} \tag{3.70}$$

3.7.10 Hermitian

Definition 3.7.9. A *Hermitian matrix* (or self-adjoint matrix) $\boldsymbol{A}$ is a complex square matrix that is equal to its own conjugate transpose, namely,

$$\boldsymbol{A} = \boldsymbol{A}^H \tag{3.71}$$

For properties, refer to Wikipedia.

3.7.11 Discrete Fourier Transform

The $N \times N$ DFT matrx is given by

$$\boldsymbol{F}_N = [e_{nk}]$$

where $n = 0, 1, \cdots, 5$, $k = 0, 1, \cdots, 5$, and $e_{nk} = e^{-inx_k} = e^{-i2\pi nk/N}$. Define

$$w = e^{-i2\pi/N} = e^{-i\pi/4} = \frac{(1-i)}{\sqrt{2}} \tag{3.72}$$

$$w^{nk} = \left(\frac{(1-i)}{\sqrt{2}}\right)^{nk} \tag{3.73}$$

[6]Reference: http://mathworld.wolfram.com/JordanCanonicalForm.html. Jordan normal form is named after French mathematician Camille Jordan.

The inverse inverse $N \times N$ DFT.

$$\boldsymbol{F}_6^{-1} = \overline{\boldsymbol{F}_N} \tag{3.74}$$

where

$$\overline{\boldsymbol{F}_N} = \frac{1}{N}[\overline{w}^{nk}] \tag{3.75}$$

More details will be on my review of Fourier Series and Fourier Transform.

3.8 Matrix Calculus

This section references [link 1] and [link 2]. Suppose $\boldsymbol{x} \in \mathbb{R}^n$, and $\boldsymbol{y} \in \mathbb{R}^m$, and that $\boldsymbol{y}$ and $\boldsymbol{x}$ are related through a function $\boldsymbol{\psi}$,

$$\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{x}) = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} \tag{3.76}$$

where each $y_i \in \mathbb{R}$, a *scalar*, is produced by a function of $\boldsymbol{x}$ (i.e. a function of a vector).

3.8.1 Differentiation

Vector to Vector The derivative of vector $\boldsymbol{y}$ with respect to vector $\boldsymbol{x}$ is an $n \times m$ matrix:

$$\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \equiv \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \tag{3.77}$$

Scalar to Vector The derivative of scalar y with respect to vector $\boldsymbol{x}$ is a column vector:

$$\frac{\partial y}{\partial \boldsymbol{x}} \equiv \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} \tag{3.78}$$

Vector to Scalar The derivative of vector $\boldsymbol{y}$ with respect to vector x is a row vector:

$$\frac{\partial \boldsymbol{y}}{\partial x} \equiv \begin{bmatrix} \frac{\partial y_1}{\partial x} & \frac{\partial y_2}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix} \tag{3.79}$$

3.8.2 Jacobian

Definition 3.8.1. For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, the determinant of the square matrix $\dfrac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}}$

$$\boldsymbol{J} = \left| \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{y}} \right| \tag{3.80}$$

is called the *Jacobian* of the transform determined by $\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{x})$.

Example D.2 of [link 1] is a good example of how to calculate the Jacobian.

3.8.3 The Chain Rule

Definition 3.8.2 (Chain Rule)**.** Suppose we have,

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{bmatrix} \qquad \boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_m \end{bmatrix} \tag{3.81}$$

where $\boldsymbol{z}$ is a function of $\boldsymbol{y}$, which is a function of $\boldsymbol{x}$. Then, we have

$$\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} = \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}} \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{x}} \tag{3.82}$$

3.9 Algorithms

3.9.1 Gauss-Seidel Method

Gauss-Seidel Method Below is my Python implementation of the Gauss-Seidel method, also known as the Liebmann method or the method of successive displacement, which is an iterative method used to solve a linear system of equations $\boldsymbol{Ax} = \boldsymbol{b}$.

```
def gauss_seidel(A, b, x_0, err, N):
    """Approximates solution for Ax=b"""
    def sigma(aj, x, start, end):
        return sum(aj[k] * x[k] for k in range(start, end))

    n = A.shape[0]
```

```
    x_m = x_0
    for m in range(N):
        x_mp1 = np.zeros(n)
        for j in range(n):
            x_mp1[j] = 1 / A[j,j] * (b[j] - sigma(A[j], x_mp1, 0, j)
                                     - sigma(A[j], x_m, j+1, n))
        j = np.argmax(np.abs(x_mp1 - x_m))
        if np.max(np.abs(x_mp1 - x_m)) < err * x_mp1[j]:
            return x_mp1
        x_m = x_mp1
        print(x_m)
    print("No solution satisfying tolerance condition after %d iterations." % N)
    return None
```

Chapter 4

Probability

Notation

Unless otherwise specified, capital letters such as X, Y, Z are random variables, and lower case letters such as x, y, z are values. $\Pr(X = x)$, or just $\Pr(x)$, denotes the probability of the event "$X = x$". $\mathbb{E}[X]$ denotes the expectation of variable X. We use $X \sim \Pr$ to denote that X is distributed according to probability distribution Pr. There are some messy notations regarding the subscript of $\mathbb{E}$ or Pr, which will be explained in the context. Ω denotes the probability space.

4.1 Probability Basics

4.1.1 Probability Space

Definition 4.1.1 (Sample Space)**.** A *sample space* Ω is the set of all possible outcomes (of all possible events).

Definition 4.1.2 (Event space)**.** An *event space* $\mathcal{S}$ is the set of measurable *events* α such that $\alpha \in \mathcal{S}$ and $\alpha \subseteq \Omega$ to which we are willing to assign probabilities.

For example, if we roll a dice, then $\Omega = \{1, 2, 3, 4, 5, 6\}$. A possible event could be $\{1\}$ (we rolled one), $\{1, 3, 5\}$ (we rolled odd), etc. We say an event α *happened* if we observed an outcome $r \in \alpha$. The event space $\mathcal{S}$ is closed under union ($\alpha \in \mathcal{S} \wedge \beta \in \mathcal{S} \to \alpha \cup \beta \in \mathcal{S}$) and complementation ($\alpha \in \mathcal{S} \to \Omega - \alpha \subseteq \mathcal{S}$).

Definition 4.1.3 (Probability distribution)**.** Given $(\Omega, \mathcal{S})$, a *probability distribution* $\Pr : \mathcal{S} \to \mathbb{R}$ is a mapping from events to real values, such that (1) for all $\alpha \in \mathcal{S}$, $\Pr(\alpha) \geq 0$, (2) $\Pr(\Omega) = 1$, and (3) for $\beta \in \mathcal{S}$, $\Pr(\alpha \cup \beta) = \Pr(\alpha) + \Pr(\beta) - \Pr(\alpha \cap \beta)$.

There is a term *probability measure* that means basically the same thing, a function that assign probabilities to measurable events.

Definition 4.1.4 (Probability Space)**.** A *probability space* is a triple $(\Omega, \mathcal{S}, \Pr)$.

4.1.2 Random Variables

Definition 4.1.5 (Random variable)**.** A random varialbe $X : \Omega \to \mathbb{R}$ associates each outcome in Ω with a value. The set of possible values that X can take is denoted as $val(X)$.

Random variables can be discrete or continuous. We primarily consider discrete ones. For simplicity, if x, y are arbitrary values for random variables X and Y, then we write $\Pr(X = x, Y = y)$ as $\Pr(X, Y)$. For a specific value x, we write $\Pr(X = x)$ as $\Pr(x)$.

Definition 4.1.6 (Marginal distribution)**.** The marginal distribution over random variable X is $\Pr(X)$.

Definition 4.1.7 (Joint distribution)**.** The joint distribution over random variables $X_1, \cdots X_n$ is $\Pr(X_1, \cdots, X_n)$ satisfying $\Pr(X_1) = \sum_{x_i, \cdots x_n} \Pr(X_1, x_2, \cdots, x_n)$. Note that 1 is arbitrarily chosen.

4.1.3 Conditional Probability

Definition 4.1.8 (Conditional probability)**.** Given two events $A, B \subseteq \Omega$, the conditional probability $P(A|B)$ is defined to be equal to

$$P(A|B) = \frac{P(A, B)}{P(B)} \tag{4.1}$$

Intuitively, we can consider "condition" as a form of additional information. Concretely, $P(A|B)$ is $P(A)$ in the universe where B happens. Therefore, $P(A|B)$ is essentially the joint probability of A, B normalized by $P(B)$ to confine it within the "new" universe.

Convenient facts:

For random variables $X_1, \cdots, X_n$, we have

$$\Pr(X_1, \cdots, X_n) = \Pr(X_1)(X_2|X_1) \Pr(X_3|X_1, X_2) \cdots \Pr(X_n|X_1, \cdots, X_{n-1}) \tag{4.2}$$

For events A, B, C, we have

$$\Pr(A, B|C) = \Pr(A|B, C) \Pr(B|C) \tag{4.3}$$

Proof.

$$\Pr(A, B|C) = \frac{\Pr(A, B, C)}{\Pr(C)} \tag{4.4}$$

$$= \frac{\Pr(A|B, C)\Pr(B, C)}{\Pr(C)} \tag{4.5}$$

$$= \Pr(A|B, C)\Pr(B, C|C) \tag{4.6}$$

$$= \Pr(A|B, C)\Pr(B|C) \tag{4.7}$$

Note for dummies (like me): $\Pr(B, C|C) = \frac{\Pr(B,C)}{\Pr(C)} = \Pr(B|C)$. □

4.1.4 Independence

Definition 4.1.9 (Independence)**.** X and Y are *independent* if $\Pr(X, Y) = \Pr(X)\Pr(Y)$.

Theorem 4.1.1 (Central Limit Theorem)**.** *Let $X_1, \cdots, X_n$ be independent identically distributed random variables with common mean μ and variance σ^2. Then*

$$\lim_{n\to\infty} \Pr\left(\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n} \leq z}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt \tag{4.8}$$

4.1.5 Conditional Independence

Definition 4.1.10 (Conditional Independence)**.** X and Y are *conditionally independent* given Z if $\Pr(X|Y, Z) = \Pr(X|Z)$ or $\Pr(X, Y|Z) = P(X|Z)P(Y|Z)$.

4.1.6 Context-Specific Independence

The term "context" refers to a specific setting of the assignment of some random variables.

Definition 4.1.11 (Context)**.** Let $\boldsymbol{Y}$ be a set of random variables. Then a specific assignment $\boldsymbol{Y} = \boldsymbol{y}$ is called a *context*, abbreviated as context $\boldsymbol{y}$.

Definition 4.1.12 (Context-Specific Independence)**.** Let $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{C}$ be disjoint sets of random variables. Then, given the context $\boldsymbol{C} = \boldsymbol{c}$, if $\boldsymbol{Y}$ is not empty, and $\Pr(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{c}) = \Pr(\boldsymbol{X}|\boldsymbol{Y}, \boldsymbol{c})$ is true, then $\boldsymbol{X}$ and $\boldsymbol{Z}$ are conditionally independent given $\boldsymbol{Y}$ *under contenxt* $\boldsymbol{C} = \boldsymbol{c}$. If $\boldsymbol{Y}$ is empty, then simply, $\boldsymbol{X}$ and $\boldsymbol{Z}$ are independent *under contenxt* $\boldsymbol{C} = \boldsymbol{c}$.

4.1.7 Bayes's Theorem

Theorem 4.1.2 (Bayes's Theorem)**.** *For random variables X, Y,*

$$\Pr(X|Y) = \frac{\Pr(X)\Pr(Y|X)}{\Pr(Y)} \tag{4.9}$$

where $\Pr(X)$ is called the prior, and $\Pr(X|Y)$ is called the posterior.

Proof. Because $\Pr(X,Y) = \Pr(Y)\Pr(X|Y) = \Pr(X)P(Y|X)$. □

4.2 Expectation

Definition 4.2.1. For discrete random variable X, the *expectation* of X under distribution Pr is defined as

$$\mathbb{E}[X] = \mathbb{E}_{X\sim\Pr}[X] = \sum_x x \Pr(x) \tag{4.10}$$

If X is continuous,

$$\mathbb{E}[X] = \mathbb{E}_{X\sim\Pr}[X] = \int_x x \Pr(x) dx \tag{4.11}$$

4.2.1 Expectation as an operator

Definition 4.2.2. Given a function $f(X)$ over $X \sim \Pr$,

$$\mathbb{E}[f(X)] = \mathbb{E}_{X\sim\Pr}[f(X)] = \sum_x f(x) \Pr(x) \tag{4.12}$$

Quote from Caltech lecture note[1]: "Expectation is a population average, i.e., you average values of the random variable $g(X)$ weighting by the population density $\Pr(x)$."

Multivariate case There is no such thing as $\mathbb{E}[X,Y]$. However, a function $g : val(X) \times val(Y) \to \mathbb{R}$ is also a random variable. In this case, we consider the joint probability distribution $\Pr(X,Y)$ such that $(X,Y) \sim \Pr$.

$$\mathbb{E}[g(X,Y)] = \mathbb{E}_{X,Y\sim\Pr}[g(X,Y)] = \sum_{x,y} g(x,y) \Pr(x,y) \tag{4.13}$$

Properties of expectation:

- $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$
- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
- If X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$
- For a constant value c, $\mathbb{E}[c] = c$.

[1] http://www.its.caltech.edu/~mshum/stats/lect2.pdf

4.2.2 Conditional Expectation

Definition 4.2.3 (Conditional Expectation)**.** For random variables Y, Z,

$$\mathbb{E}[Y|Z=z] = \sum_y y \Pr(Y=y|Z=z) \tag{4.14}$$

This statement says that $\mathbb{E}[Y|Z=z]$ is a weighted sum of the values that Y assumes. However, this is confined, again, to the universe where $Z=z$ has happened (additional information). Of course, $Z=z$ can be replaced by any event.

Note The notation $\mathbb{E}[Y|Z]$ is in fact a random variable - a function of Z, whereas $\mathbb{E}[Y|Z=z]$ is a fixed value.

4.2.3 Variance

Definition 4.2.4. The variance of variable X is

$$\mathrm{Var}[X] = \mathbb{E}\Big[(X-\mathbb{E}[X])^2\Big] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \tag{4.15}$$

Properties of variance:

- $\mathrm{Var}[aX+b] = a^2\mathrm{Var}[X]$
- If X and Y are independent, then $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$

Covariance matrix Suppose $X = [X_1, \cdots, X_n]$. Then we define the covariance matrix Σ of X as

$$\Sigma = \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T] \tag{4.16}$$

if $\mu_i = \mathbb{E}[X_i]$, then each entry $\Sigma_{ij} = \mathrm{cov}(X_i, X_j) = \mathbb{E}[(X_i-\mu_i)(X_j-\mu_j)] = \mathbb{E}[X_iX_j] - \mu_i\mu_j$.

4.2.4 Moment

Definition 4.2.5 (Moment)**.** The kth *moment* of a random variable X is $\mathbb{E}[X^k]$.

Definition 4.2.6 (Central Moment)**.** The kth *central moment* is given by $\mathbb{E}[(X-\mathbb{E}[X])^k]$.

4.3 Inequalities

4.3.1 Markov Inequality

Theorem 4.3.1 (Markov Inequality)**.** *Let X be a random variable such that $\forall x \in val(X), x \geq 0$. Then, for all $a > 0$,*

$$\Pr(X \geq a) \leq \frac{\mathbb{E}[X]}{a} \tag{4.17}$$

Proof. For $a > 0$, let I be an indicator variable such that $val(I) \in \{0, 1\}$, and $I = 1$ if $x \geq a$. Since $X \geq 0$, we have $I \cdot X \geq a$ if $X \geq a$, and $I \cdot X = 0 < a$, if otherwise.

Equivalently, $Ia \leq X$ if $X \geq a$, and $Ia \leq X$ if otherwise. Therefore,

$$I \leq \frac{X}{a} \tag{4.18}$$

Since I is a 0-1 random variable,

$$\mathbb{E}[I] = \Pr(I = 1) = \Pr(X \geq a) \tag{4.19}$$

Therefore,

$$\Pr(X \geq a) = \mathbb{E}[I] \leq \mathbb{E}\left[\frac{X}{a}\right] = \frac{\mathbb{E}[X]}{a} \tag{4.20}$$

□

4.3.2 Chebyshev's Inequality

Theorem 4.3.2 (Chebyshev's Inequality)**.** *For any random variable X, and for all $a > 0$,*

$$\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{Var[X]}{a^2} \tag{4.21}$$

Proof. This follows from Markov Inequality.

$$\Pr(|X - \mathbb{E}[X]| \geq a) = \Pr((X - \mathbb{E}[X])^2 \geq a^2) \tag{4.22}$$

Since $(X - \mathbb{E}[X])^2$ is a nonnegative random variable, by Markov's Inequality,

$$\Pr((X - \mathbb{E}[X])^2 \geq a^2) \leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{a^2} \leq \frac{\mathrm{Var}[X]}{a^2} \tag{4.23}$$

□

4.3.3 Chernoff Bound

Theorem 4.3.3 (Generic Cheroff Bound)**.** *Suppose X is a real-valued random variable. For any $t > 0$,*

$$\Pr(X \geq a) = \Pr(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}[e^{tX}]}{e^{ta}} \tag{4.24}$$

Below is the Chernoff Bound for a specific type of random variables. They are independent 0-1 random variables, known as Poisson trials (not Poisson variables!)[2]

[2]Slides in Brown CS 2540: http://cs.brown.edu/courses/csci1550/slides/2019/Chapter-4.pdf

Theorem 4.3.4 (Chernoff Bound for Poisson Trials)**.** *Let $X_1, \cdots, X_n$ be a sequence of independent Poisson trials, with $\Pr(X_i = 1) = p_i$. Let $X = \sum_{i=1}^{n} X_i$ and let $\mu = \mathbb{E}[X] = \mathbb{E}[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} p_i$. Then for any $\delta \in [0, 1]$,*

$$\Pr(\frac{1}{n}\sum_{i=1}^{n} X_i \geq (1+\delta)\mu \leq e^{-\mu n \sigma^2/3}) \tag{4.25}$$

and

$$\Pr(\frac{1}{n}\sum_{i=1}^{n} X_i \leq (1-\delta)\mu \leq e^{-\mu n \sigma^2/2}) \tag{4.26}$$

4.3.4 Hoeffding's Bound

Theorem 4.3.5. *Let $X_1, \cdots, X_n$ be a sequence of independent random variables such that for all $1 \leq i \leq n$, $\mathbb{E}[X_i] = \mu$ and $\Pr(a \leq X_i \leq b) = 1$. Then,*

$$\Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2/(b-a)^2} \tag{4.27}$$

4.4 Bayesian Optimal Classifier

Suppose $\mathcal{D}$ is some distribution of samples (x, y), where $x \in \mathbb{R}^d$ and $y \in val(Y)$ and Y is a discrete random variable. This means $\mathcal{D}(x, y)$ outputs the probability of (x, y) to exist in the world. A classifier $f(x)$ outputs the category (or class) y given input x. The **Bayesian optimal classifier** is one defined as

$$f^*(x) = \operatorname*{argmax}_y \mathcal{D}(x, y) \tag{4.28}$$

Theorem 4.4.1. *Bayesian optimal classifier achieve minimal error among all classifiers.*

Proof. Assume there exists another deterministic classifier f' that produces lower error than f^*. Then, for some input x, we have $f'(x) \neq f(x)$. Suppose in the real world, input x can map to classes $\mathcal{Y} = \{y_1, \cdots, y_n\}$. Suppose $f'(x) = y_p \in \mathcal{Y}$. We have:

- Probability of x to occur is $\mathcal{D}(x) = \sum_{y \in \mathcal{Y}} \mathcal{D}(x, y)$.
- Probability of (x, y_p) to be observed is $\mathcal{D}(x, y_p) = \mathcal{D}(x, f'(x))$.
- Probability of (x, y_q) where $y_q \in \mathcal{Y} \setminus \{y_p\}$ to be observed is $\mathcal{D}(x) - \mathcal{D}(x, f'(x))$. This is the probability that f' made a mistake.

Similarly, the probability that f^* made a mistake is $\mathcal{D}(x) - \mathcal{D}(x, f^*(x))$. By definition of Bayesian optimal classifier,

$$\mathcal{D}(x, f^*(x)) = \max_y \mathcal{D}(x, y) \geq \mathcal{D}(x, f'(x)) \tag{4.29}$$

Therefore,

$$\mathcal{D}(x) - \mathcal{D}(x, f^*(x)) \leq \mathcal{D}(x) - \mathcal{D}(x, f'(x)) \tag{4.30}$$

Thus, f^* makes fewer mistakes. When f' and f^* only disagree on x, it is not possible for $f'(x)$ being correct while $f^*(x)$ being wrong. Therefore, f^* is optimal. □

Probability theory is such a genious human creation that rationalizes the uncertain world. The content of this chapter comes from my undergraduate study as well as the course CSCI 2540: Probablistic Methods, taken at Brown while studying PhD.

Chapter 5

Fourier Series

Review of Fourier Series, Fourier Integral, Fourier Transform, and related knowledge. Material is based on the textbook *Advanced Engineering Mathematics* by Erwin Kreyszig.

5.1 Series Solution to ODEs

This will be captured by several examples from the homework problems in ENGN 2010 that I have taken.

Find a **power series solution** in powers of x

1. $y'' - y' + xy = 0$

   Plug in the series form of y, y', and y'':

$$\sum_{m=2}^{\infty} m(m-1)a_m x^{m-2} - \sum_{m=1}^{\infty} m a_m x^{m-1} + x\sum_{m=0}^{\infty} a_m x^m = 0 \tag{5.1}$$

   Change all exponents of x to $m-1$

$$\sum_{m=1}^{\infty}(m+1)(m)a_{m+1}x^{m-1} - \sum_{m=1}^{\infty} m a_m x^{m-1} + \sum_{m=2}^{\infty} a_{m-2}x^{m-1} = 0 \tag{5.2}$$

$$2a_2 - a_1 + \sum_{m=2}^{\infty}\Big((m+1)ma_{m+1} - ma_m + a_{m-2}\Big) = 0 \tag{5.3}$$

   This means

$$2a_2 - a_1 = 0 \tag{5.4}$$

$$(m+1)ma_{m+1} - ma_m + a_{m-2} = 0 \tag{5.5}$$

The second equation can be rewritten as:

$$a_{m+1} = \frac{ma_m - a_{m-2}}{(m+1)m} \tag{5.6}$$

Equations (14) and (16) suggest that a_0, a_1 are arbitrary. For a_2 and above,

$$a_2 = \frac{1}{2}a_1 \tag{5.7}$$

$$a_3 = \frac{2a_2 - a_0}{(3)(2)} = \frac{a_1 - a_0}{6} \tag{5.8}$$

$$a_4 = \frac{3a_3 - a_1}{(4)(3)} = \frac{-a_1 - a_0}{24} \tag{5.9}$$

$$a_5 = \frac{4a_3 - a_2}{(5)(4)} = \frac{-4a_1 - a_0}{120} \tag{5.10}$$

$$\dots \tag{5.11}$$

Therefore,

$$y = a_0\Big(1 - \frac{1}{6}x^3 - \frac{1}{24}x^4 - \frac{1}{120}x^5 + \cdots\Big) + a_1\Big(x + \frac{1}{2}x^2 + \frac{1}{6}x^3 - \frac{1}{24}x^4 - \frac{1}{30}x^5 + \cdots\Big) \tag{5.12}$$

2. $(1 - x^2)y'' - 2xy' + 2y = 0$

Plug in the series form of y, y', and y'':

$$(1 - x^2)\sum_{m=2}^{\infty} m(m-1)a_m x^{m-2} - 2x\sum_{m=1}^{\infty} ma_m x^{m-1} + 2\sum_{m=0}^{\infty} a_m x^m = 0 \tag{5.13}$$

$$\sum_{m=2}^{\infty} m(m-1)a_m x^{m-2} - \sum_{m=2}^{\infty} m(m-1)a_m x^m - 2\sum_{m=1}^{\infty} ma_m x^m + 2\sum_{m=0}^{\infty} a_m x^m = 0 \tag{5.14}$$

Change all exponents of x to m

$$\sum_{m=0}^{\infty} (m+2)(m+1)a_{m+2} x^m - \sum_{m=2}^{\infty} m(m-1)a_m x^m - 2\sum_{m=1}^{\infty} ma_m x^m + 2\sum_{m=0}^{\infty} a_m x^m = 0 \tag{5.15}$$

The second evaluates to 0 when $m = 1$ or $m = 0$. Thus, we can modify the range of the sum to start from $m = 0$. The third term evaluates to 0 when $m = 0$. Thus, we can modify the range of the sum to start from $m = 0$ as well.

$$\sum_{m=0}^{\infty} (m+2)(m+1)a_{m+2} x^m - \sum_{m=0}^{\infty} m(m-1)a_m x^m - 2\sum_{m=0}^{\infty} ma_m x^m + 2\sum_{m=0}^{\infty} a_m x^m = 0 \tag{5.16}$$

$$\sum_{m=0}^{\infty}\Big((m+2)(m+1)a_{m+2}-m(m-1)a_m-2ma_m+2a_m\Big)x^m=0 \tag{5.17}$$

This means

$$(m+2)(m+1)a_{m+2}-\Big(m(m-1)+2m-2\Big)a_m=0 \tag{5.18}$$

$$a_{m+2}=\frac{m^2+m-2}{(m+2)(m+1)}a_m \tag{5.19}$$

The above suggests a_0 and a_1 are arbitrary. For a_2 and above,

$$a_2=\frac{-2}{2}a_0=-a_0 \tag{5.20}$$

$$a_3=\frac{2-2}{(3)(2)}a_1=0 \tag{5.21}$$

$$a_4=\frac{4}{12}a_2=-\frac{1}{3}a_0 \tag{5.22}$$

$$a_5=0 \tag{5.23}$$

$$a_6=-\frac{18}{30}\frac{1}{3}a_0=-\frac{1}{5}a_0 \tag{5.24}$$

$$\cdots \tag{5.25}$$

Therefore,

$$y=a_0\Big(1-x^2-\frac{1}{3}x^4-\frac{1}{5}x^6+\cdots\Big)+a_1x \tag{5.26}$$

5.2 Fourier Series

Given a periodic function $f(x)$ with period $P=2\pi$, the function can be represented by a *Fourier Series* as follows:

$$f(x)=a_0+\sum_{n=1}^{\infty}(a_n\cos(mx)+b_n\sin(nx)) \tag{5.27}$$

where a_0, a_n, b_n are *Fourier Coefficients*, given by

$$a_0=\frac{1}{2\pi}\int_{-\pi}^{\pi}f(x)dx \tag{5.28}$$

$$a_n=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\cos(nx)dx \tag{5.29}$$

$$b_n=\frac{1}{\pi}\int_{-\pi}^{\pi}f(x)\sin(nx)dx \tag{5.30}$$

Examples Find Fourier transform for the following functions without using Tables.

1. $f(x) = \begin{cases} e^{kx} & x < 0 \\ 0 & x > 0 \end{cases}$

$$\mathscr{F}[f](w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-iwx} dx \tag{5.31}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{(k-iw)x} dx \tag{5.32}$$

$$= \frac{1}{\sqrt{2\pi}} \left[\frac{e^{(k-iw)x}}{k-iw} \right]_{-\infty}^{0} \tag{5.33}$$

$$= \frac{1}{\sqrt{2\pi}} \frac{1}{k-iw} \tag{5.34}$$

2. $f(x) = \begin{cases} |x| & -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

$$\mathscr{F}[f](w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-iwx} dx \tag{5.35}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} |x| e^{-iwx} dx \tag{5.36}$$

$$= \frac{1}{\sqrt{2\pi}} \Big(\int_{-1}^{0} -x e^{-iwx} dx + \int_{0}^{1} x e^{-iwx} dx \Big) \tag{5.37}$$

$$= \frac{1}{\sqrt{2\pi}} \left(\frac{-2 + e^{iw}(1-iw) + e^{-iw}(1+iw)}{w^2} \right) \tag{5.38}$$

3. $f(x) = e^{-|x|} \qquad -\infty < x < \infty$

$$\mathscr{F}[f](w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-iwx} dx \tag{5.39}$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-|x|-iwx} dx \tag{5.40}$$

$$= \frac{1}{\sqrt{2\pi}} \left(\int_{-\infty}^{0} e^{x-iwx} dx + \int_{0}^{\infty} e^{-x-iwx} dx \right) \tag{5.41}$$

$$= \frac{1}{\sqrt{2\pi}} \Big(\frac{i}{w+i} + \frac{-i}{w-1} \Big) \tag{5.42}$$

$$= \frac{1}{\sqrt{2\pi}} \Big(\frac{2}{w^2+1} \Big) \tag{5.43}$$

5.3 Fourier Transform

The **Fourier Cosine Transform** is

$$\hat{f}_c(w) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \cos(wx) dx \tag{5.44}$$

When expressed in Fourier Cosine Integral,

$$f(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty \hat{f}_c(w) \cos(wx) dw \tag{5.45}$$

1. $f(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & x > 1 \end{cases}$

$$\hat{f}_c(w) = \sqrt{\frac{2}{\pi}} \int_0^1 \cos(wx) dx = \sqrt{\frac{2}{\pi}} \frac{\sin(w)}{w} \tag{5.46}$$

Therefore,

$$f(x) = \frac{2}{\pi} \int_0^\infty \frac{\sin(w)}{w} \cos(wx) dw \tag{5.47}$$

2. $f(x) = \begin{cases} x^2 & 0 < x < 1 \\ 0 & x > 1 \end{cases}$

$$\hat{f}_c(w) = \sqrt{\frac{2}{\pi}} \int_0^1 x^2 \cos(wx) dx \tag{5.48}$$

$$= \sqrt{\frac{2}{\pi}} \Big(\frac{2\cos(w)}{w^2} + \frac{(w^2 - 2)\sin(w)}{w^3} \Big) \tag{5.49}$$

Therefore,

$$f(x) = \frac{2}{\pi} \int_0^\infty \Big(\frac{2\cos(w)}{w^2} + \frac{(w^2 - 2)\sin(w)}{w^3} \Big) \cos(wx) dw \tag{5.50}$$

The **Fourier Sine Transform** is

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \sin(wx) dx \tag{5.51}$$

When expressed in Fourier Sine Integral,

$$f(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty \hat{f}_s(w) \sin(wx) dw \tag{5.52}$$

1. $f(x) = \begin{cases} x & 0 < x < a \\ 0 & x > a \end{cases}$

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}} \int_0^a x \sin(wx) dx = \sqrt{\frac{2}{\pi}} \Big(\frac{\sin(aw)}{w^2} - \frac{a \cos(aw)}{w} \Big) \tag{5.53}$$

Therefore,

$$f(x) = \frac{2}{\pi} \int_0^\infty \Big(\frac{\sin(aw)}{w^2} - \frac{a \cos(aw)}{w} \Big) \sin(wx) dw \tag{5.54}$$

2. $f(x) = \begin{cases} e^{-x} & 0 < x < 1 \\ 0 & x > 1 \end{cases}$

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}} \int_0^1 e^{-x} \sin(wx) dx = \sqrt{\frac{2}{\pi}} \Big(\frac{ew - \sin(w) - w \cos(w)}{ew^2 + e} \Big) \tag{5.55}$$

Therefore,

$$f(x) = \frac{2}{\pi} \int_0^\infty \Big(\frac{ew - \sin(w) - w \cos(w)}{ew^2 + e} \Big) \sin(wx) dw \tag{5.56}$$

Part II

Artificial Intelligence

Chapter 6

Probabilistic Graphical Models

Chapter 7

Markov Decision Process

7.1 Definition

In general, an agent interacts with its environment by taking *actions* and receiving *sensory observations*. There is typically uncertainty in actions and observations, due to imperfect actuators, external disturbances, and noisy or malfunctioning sensors. A *Markov Decision Process* is a framework (paradigm) that incorporates uncertainty in action but not in sensing. It is defined by a tuple (S, A, T, R, γ). S and A are the state and action spaces of the agent, respectively. T specifies the uncertainty in the transition between states once an action is taken. R specifies the reward signal, discounted by $\gamma \in \mathbb{R}$. In this case, the planner must consider different possibilities which increase, in the worst case, exponentially as a function of the number of actions and the planning horizon. The job of the planner is to produce a policy that maps from states to actions, denoted as $\pi : S \to A$. The fact that an action can be determined sufficiently by the current state makes this framework *Markov*.

7.2 Derivation of Markov Decision Process

In the birth of the universe, no one was told how MDPs should be defined as. The motivating setup is the following. At time t, an environment has a *state* $x_t \in S$ where S is the space of all possible states. The agent is allowed to take an action $u_t \in A$ where A is the space of all possible actions. The action will have an impact on the state of the environment, causing it to change into a potentially different state $x_{t+1} \in S$. In this process, the agent receives a reward $r_t \in \mathfrak{R}$ for taking the action that caused the change.

As a simple first step, it is assumed that the agent knows entirely the value of x. If that is not the case, the environment is partially observable, which will be covered in the next chapter. The goal of the agent is, intuitively, to maximize the discounted future reward

over a *planning horizon* of T steps:

$$R_T(x_t) = \mathbb{E}[\sum_{s=1}^{T} \gamma^s r_{t+s} | x_s] \tag{7.1}$$

We can perform the following derivation on $R_T(x_t)$ to establish its connection to u_t and x_{t+1}. (TODO) Finally we obtain

$$R_T(x) = \sum_u \Pr(u|x) \sum_{x'} \sum_r \Pr(x', r|u, x)(r + \gamma R_T(x')) \tag{7.2}$$

This is called the *Bellman equation* and $R_T(x)$ is called the value of state x.

To evaluate the value of a state under a specific policy π, which is assumed to map x to an action u deterministically (that is, $\Pr(u|x) = 1$ only if $\pi(x) = u$),

$$\begin{aligned}
R_T^\pi(x) &= \sum_{x'} \sum_r \Pr(x', r|\pi(x), x)(r + \gamma R_T^\pi(x')) && (7.3)\\
&= \sum_{x'} \sum_r \Pr(x', r|\pi(x), x) r + \gamma \sum_{x'} \sum_r \Pr(x', r|\pi(x), x) R_T^\pi(x') && (7.4)\\
&= \sum_r \Pr(r|\pi(x), x) r + \gamma \sum_{x'} \Pr(x'|\pi(x), x) R_T^\pi(x') && (7.5)\\
&= \mathbb{E}[r|\pi(x), x] + \gamma \sum_{x'} \Pr(x'|\pi(x), x) R_T^\pi(x') && (7.6)
\end{aligned}$$

Now it makes sense to define the reward function $R(u, x) = \mathbb{E}[r|u, x]$, and the transition function $T(x, u, x') = \Pr(x'|u, x)$. Thus,

$$R_T^\pi(x) = R(\pi(x), x) + \gamma \sum_{x'} T(x, \pi(x), x') R_T^\pi(x') \tag{7.7}$$

This implies, the optimal value $R_T^*(x)$ and the optimal policy π^* are

$$R_T^*(x) = \max_u \left\{ R(u, x) + \gamma \sum_{x'} T(x, \pi(x), x') R_T^*(x') \right\} \tag{7.8}$$

$$\pi^*(x) = \operatorname*{argmax}_u \left\{ R(u, x) + \gamma \sum_{x'} T(x, u, x') R_T^*(x') \right\} \tag{7.9}$$

An interative method to estimate R_T^* and $\pi^*(x)$ is therefore natural because $R_T^*(x')$ and $R_T^*(x)$ cannot be both stable — we need to use $R_{T-1}^*(x')$ instead.

Chapter 8

Partially Observable Markov Decision Process

HYE

Chapter 9

Neural Networks

Chapter 10

Non-Parametric Methods

Chapter 11

Supervised Learning

Chapter 12

Unsupervised Learning

Chapter 13

Reinforcement Learning

Part III

Systems

Part IV

Programming

Part V

Physics

Part VI

Finance

Part VII

Law

Chapter 14

Jargon

Information-theoretic Relating to "information theory", which concerns quantization, storage and communication of information, originally proposed by Claude Shannon. Entropy is a core concept of information theory. Inference, graphical models, information gain, are considered information-theoretic concepts.

Decision-theoretic Relating to "decision theory", which concerns the choice of action of an agent, such as the planning problem. Reinforcement learning is essentially a framework to learn decision making for an agent, thus it is a decision-theoretic view.

Information-theoretic vs Decision-theoretic There could be information-theoretic tasks, such as maximum-likelihood estimation, and there could be decision-theoretic tasks, such as planning to reach a goal. The two can also help each other. For example, an agent may need to gather information and reason about them before making a decision. The experience generated from one or many agents running in the world can be collected as information and analyzed together in probabilistically to reach some conclusion about the information.